

Figure 6–6: MSH

Fast-feedback allocation subheader (FFSH)

The FFSH (Figure 6–7) is used by the BS to request feedback from an AAS SS. The support for AAS functionality is PHY-specific and described in Chapter 13. The allocation offset is defined in units of slots. This is the offset from the beginning of the fast-feedback UL bandwidth allocation of the slot in which the SS must send a fast-feedback message. The allocation applies to the UL subframe of exactly two frames after the frame including the FFSH. The Feedback Type field identifies the type of the feedback, as shown in Table 6–7. Chapter 13 describes these feedback types in detail.

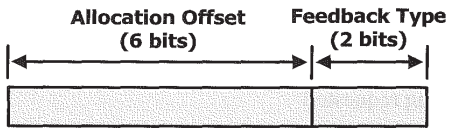


Figure 6–7: FFSH

Table 6–7: Fast-feedback allocation feedback type encodings

Feedback type	Description
00	Fast DL measurement
01	Fast MIMO feedback, antenna #0
10	Fast MIMO feedback, antenna #1
11	MIMO mode and permutation mode feedback

ARQ feedback

ARQ feedback refers to the ARQ control information, such as acknowledgments from a receiver to the transmitter. Since this control information flows in the direction opposite to that of the actual ARQ data flow, efficient mechanisms are required to transport the feedback as a piggybacked payload or as a separate management message. In order to send a separate ARQ feedback management message, especially in the UL, the BS has to make specific allocations for the feedback, or the SS has to steal bandwidth from another allocation. An alternate approach is to piggyback on another connection, which may effectively steal some of its bandwidth.

IEEE Std 802.16 allows such piggybacking of ARQ feedback on any MPDU, subject to system policies. This is similar to the way in which subheaders are transported with MPDUs. However, ARQ feedback is a special payload, not a MAC subheader. The format of the ARQ special payload and rules for piggybacking are described in later sections.

DATA AND MANAGEMENT PDU CONSTRUCTION

Simple MPDU

A simple MPDU consists of the generic MAC header and optionally the payload and CRC. Although CRC is optional, it is a good practice to include it in real-world implementations, unless perhaps in the case of an LOS system with a very low bit error rate (BER). The IEEE 802.16 standard mandates the implementation of CRC if the PHY type is WirelessMAN-SCa, WirelessMAN-OFDM, or WirelessMAN-OFDMA. Figure 6–8 shows the simple MPDU structure.

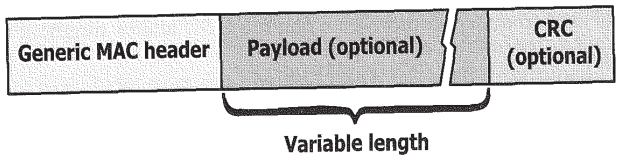


Figure 6–8: MPDU structure

The MPDU format is flexible enough to carry both small and large payloads with minimal overhead. With the appropriate choice of packing and fragmentation and an intelligent scheduler, the overall overhead can be low. When subheaders and special payloads are present, all per-PDU subheaders are placed just after the generic MAC header. The optional subheaders, if present, are considered part of the payload and hence protected by the same CRC that protects the rest of the payload. The format of the MPDU, with subheaders and ARQ special payloads, is shown in Figure 6–9. As shown in the figure, all per-PDU subheaders precede per-SDU subheaders.

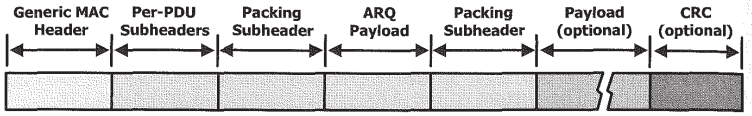


Figure 6–9: MPDU with subheaders

Subheader ordering

As discussed in earlier sections, the generic MAC header indicates the presence or absence of subheaders. However, in order to properly decode MPDUs and subheaders, an ordering of the subheaders and special payloads is defined. The following rules define the use of subheaders in UL and DL:

- ***PSH and FSH:*** The PSH and FSH are mutually exclusive. The FSH is used only to transport a single fragment. The PSH contains the necessary information to support both packing and fragmentation. Therefore, the PSH is used to transport more than one full SDU and/or fragment in the same payload.
- ***ARQ feedback payload:*** The ARQ feedback special payload, if present, is always inserted as the first payload (i.e., after the subheaders, if any) when packing is enabled. This is illustrated in Figure 6–9. Since ARQ feedback is a payload, a PSH must precede it, if packing is required.

- ***MSH:*** The MSH, if present, precedes all other subheaders.
- ***FFSH:*** The FFSH will always be the last per-PDU subheader. Therefore, only the PSH can follow this subheader.
- ***GMSH and FSH:*** The GMSH precedes the FSH.

Based on these rules, the ordering of the subheaders and the special payload is defined as shown in Table 6–8.

Table 6–8: Subheader and special payload ordering

Order	Subheader
1	MSH (mesh subheader)
2	GMSH (grant management subheader)
3	FSH (fragmentation subheader)
4	FFSH (fast-feedback allocation subheader)
5	PSH (packing subheader)
6	ARQ special payload

Figure 6–10 illustrates the ordering of the subheaders in UL and DL. Note that some subheaders are relevant in only one of these.

ARQ blocks

It is important to explain the concept of ARQ blocks before we discuss the fragmentation and packing in detail. The ARQ block is the basic unit of transmission or retransmission for an ARQ-enabled connection. An MSDU is logically partitioned into blocks of a fixed length, negotiated during connection setup. When the length of the SDU is not an integer multiple of the negotiated block size, the final block of the SDU is formed using the remaining SDU bytes after the final full block has been determined.

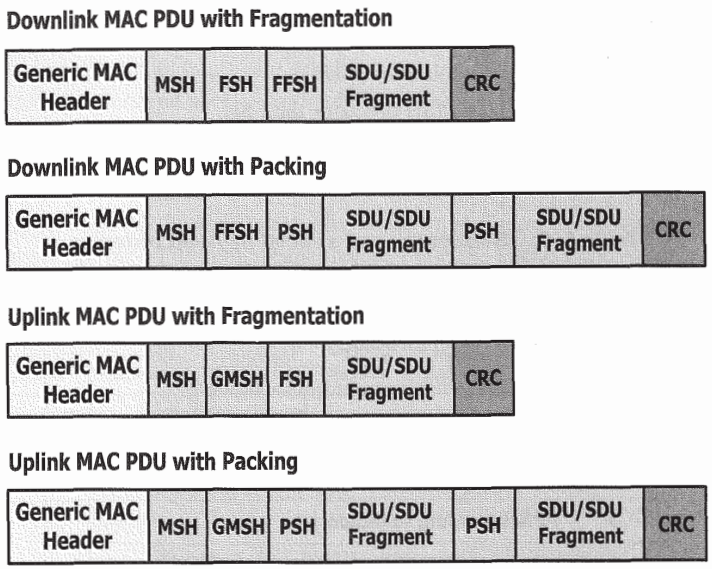


Figure 6–10: Ordering of MAC subheaders

The size of the ARQ block can be of any value from 1 to 2040 bytes. The appropriate choice of the ARQ block size will depend on the ToS carried over the connection, the underlying PHY properties, and the expected feedback delay of the ARQ protocol, among other factors.

Each ARQ block of an SDU is assigned a sequence number called the *block sequence number* (BSN). Fragmentation is allowed to occur only on ARQ block boundaries. Therefore, an ARQ fragment is formed with an integral number of contiguous ARQ blocks. The BSN value carried in the FSH is the BSN for the first ARQ block appearing in the segment.

Figure 6–11 illustrates the concept of ARQ blocks with two consecutive SDUs of the same connection. SDU #1 consists of seven blocks starting from BSN 5. The seventh block, with BSN 11, is of a size smaller than the negotiated ARQ block size. Similarly, SDU #2 consists of five ARQ blocks,

from BSN 12 through BSN 16, where the last ARQ block is smaller than the rest. Both the SDUs are divided into two fragments each: <Frag /SDU#1> (3 blocks, BSN = 5), <Frag 1/SDU#1> (4 blocks, BSN = 8), <Frag 0/SDU#2> (3 blocks, BSN = 12) and <Frag 1/SDU#2> (2 blocks, BSN = 15).

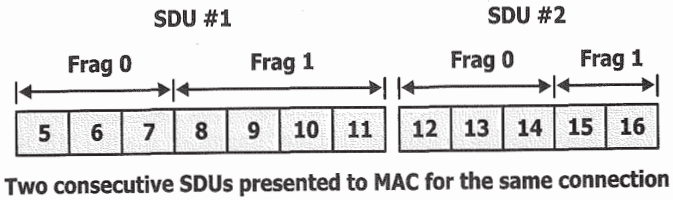


Figure 6–11: ARQ blocks

Fragmentation

The fragmentation function is negotiated on a per-connection basis during connection setup. There are two choices of FSH for non-ARQ connections. The 11-bit FSN is used by all management connections that support fragmentation. The support for the 3-bit FSN is optional in IEEE Std 802.16. If a specific connection has fragmentation enabled, all SDU fragments on that connection are preceded by a FSH. The FSN is incremented by 1 (modulo 8 or modulo 2048) for each SDU fragment and unfragmented SDU.

The fragment size may be chosen dynamically, based on the real-time conditions and other system requirements. This is often called *dynamic fragmentation*. If the regular FSH is used, up to eight fragments are supported per SDU to avoid ambiguity. However, with the extended FSH or when ARQ is enabled, an SDU may be fragmented into any number of fragments to support QoS and other system requirements. There is no restriction of the size of the fragment.

As described in the previous subsection on ARQ blocks, fragmentation works differently for an ARQ connection. For both ARQ and non-ARQ connections, the FSH is of the same size, but the sequence number fields are interpreted

differently. On connections with no ARQ, fragmentation can be done on any byte boundary, and the FSN counter is incremented for every fragment. For ARQ-enabled connections, fragmentation occurs on ARQ block boundaries, and the BSN counter is incremented for every ARQ block.

Fragmentation without ARQ

As soon as a fragment is created, the next valid sequence number is assigned to it. This FSN is carried in the FSH, along with the FC field indicating whether a fragment is the first, an intermediate, or the last fragment from an MSDU. It is also possible to indicate that the "fragment" is a full MSDU (i.e., FC = 00), mostly for convenience of implementation. Although there is no requirement to include a FSH in an MPDU containing a full MSDU, the FSN counter must be incremented even if the FSH is omitted. Figure 6–12 shows the structure of an MPDU with fragmentation.



Figure 6–12: MPDU with fragmentation

The reassembly process for non-ARQ connections in the receiver is fairly straightforward, since fragments are always transmitted in order and can be transmitted only once. If the receiver discovers that a fragment has been lost, it discards any fragments waiting to be reassembled and all fragments until it encounters (a) a fragment indicating it is the first or (b) an MPDU containing a full MSDU. Therefore, the receiver has to buffer at most one full MPDU per connection.

Fragmentation with ARQ

Each MPDU sent on an ARQ-enabled connection must include either the ARQ FSH or the ARQ PSH, even if fragmentation is not enabled, as the BSN must be carried to the receiver. If only fragmentation is used, the MPDU

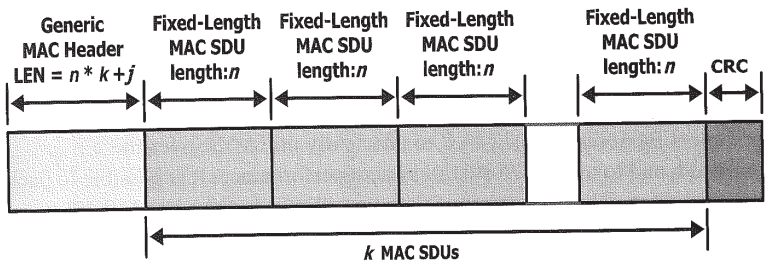
structure is similar to the one shown in Figure 6–12. The only difference is the type of FSH used for ARQ connections.

Packing

There are three types of packing supported in IEEE 802.16: packing of fixed-length SDUs, packing of variable-size SDUs, and packing of ARQ connections.

Packing of fixed-length SDUs

Fixed-size SDUs, such as ATM cells, can be packed and transported on a connection without using PSHs. During connection setup, the SDU parameters can be negotiated for fixed-size packing with no fragmentation. This allows the fixed-size SDUs to be packed efficiently without the PSH overhead. The number of SDUs per MPDU can still be varied dynamically based on QoS and other requirements. The Length field in the generic MAC header implicitly defines the number of packed SDUs. Figure 6–13 shows an MPDU packed with fixed-size SDUs.



Note: j is the length of the MAC Header plus the CRC, in bytes.

Figure 6–13: Packing of fixed-length SDUs

Packing of variable-length SDUs

Variable-length SDUs can be packed using one of the PSHs. Although all PSHs have all the necessary fields to simultaneously support packing and fragmentation, packing can be used without fragmentation. The regular PSH is sufficient to support packing without fragmentation. A packed MPDU, with each SDU preceded by a PSH, is shown in Figure 6–14.

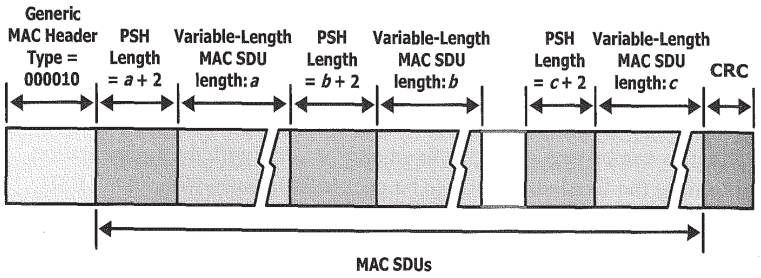


Figure 6–14: Packing of variable-length SDUs

Packing with fragmentation

As discussed earlier, the PSH is a superset of the FSH. Both can be simultaneously supported on the same connection. This combination provides the utmost flexibility to the MAC to efficiently utilize the over-the-air bandwidth and to maximize overall system performance. It is possible to pack multiple consecutive fragments of the same SDU into the same MPDU. However, such fragmentation and packing can only increase the overhead unnecessarily, with no real benefit. Therefore, typically an MPDU with packing and fragmentation will have at most two fragments of two different SDUs and zero or more unfragmented SDUs.

Figure 6–15 shows an example of a packed MPDU with two fragments and some unfragmented SDUs. The first packed payload is the last fragment of an SDU and the last payload is the first fragment of another SDU. Another example of a packed MPDU with one fragment and some unfragmented SDUs is shown in Figure 6–16.

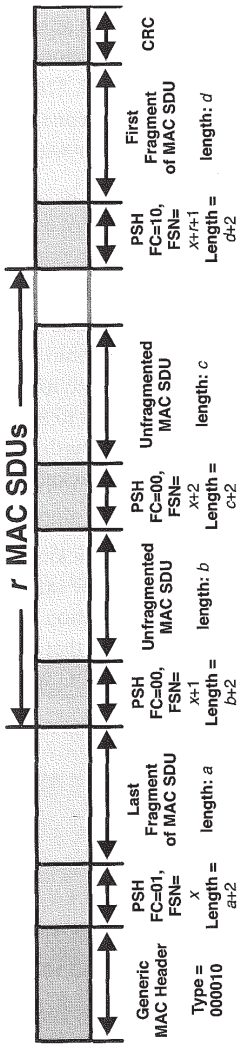


Figure 6–15: Packed MPDU with two fragments

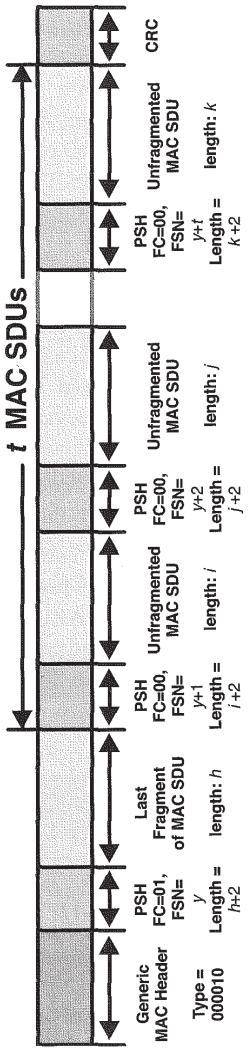


Figure 6–16: Packed MPDU with one fragment

Packing and ARQ

An ARQ connection must use either a FSH or a PSH, as the ARQ algorithm requires the BSN to keep track of ARQ blocks. If ARQ is used with no fragmentation or packing, the ARQ FSH is used to carry the BSN. The same subheader can be used to add fragmentation function to an ARQ connection without packing. If the ARQ connection requires packing (with or without fragmentation), the ARQ PSH must be used.

Packing of ARQ payload

The ARQ feedback payload consists of one or more ARQ Feedback information elements (IEs), the format and contents of which will be covered elsewhere in this book. When the ARQ feedback is piggybacked onto another MPDU with packing enabled, it shall always be the first packed payload. The FC and FSN fields are irrelevant to the ARQ feedback and are set to zero. Since the length of the ARQ feedback payload is variable, the Length field in the PSH is used to correctly decode the ARQ payload from the packed MPDU.

Although only the Length field of the PSH is of significance for packing ARQ feedback payload, all PSHs used within an MPDU should be of the same type (i.e., regular, extended, or ARQ), as negotiated during the connection setup. The ARQ feedback payload can be piggybacked on a connection that may or may not have ARQ enabled. Figure 6–17 shows an example MPDU with an ARQ feedback payload.



Figure 6–17: ARQ feedback and packing

Concatenation

Further efficiency can be achieved by reducing the per-PDU PHY overhead through concatenation. This is achieved by transmitting multiple MPDUs in a single PHY burst. Although concatenation is not a MAC feature, it is covered in this section because it is related to other efficiency functions, such as packing and fragmentation. Concatenation is allowed in both the UL and DL.

For some of the IEEE 802.16 PHY specifications, an intermediate sublayer, called the *transmission CS*, is specified to efficiently support concatenation. This sublayer inserts a pointer byte at the beginning of a payload to help the receiver separate the MPDUs in a burst. This can be used to correctly receive some of the concatenated MPDUs in a PHY burst even if part of the PHY burst was received incorrectly (e.g., if a part of the burst was corrupted). The transmission CS is mandatory for the WirelessMAN-SC PHY and optional for the WirelessMAN-OFDM PHY.

Figure 6–18 shows the concept of concatenation for an UL burst transmission. As shown in the figure, any type of MPDU or message may be concatenated into a single PHY burst. The figure shows two UL bursts, one that concatenates a BW request and a user data MPDU and another that concatenates a MAC management message and user data MPDUs.

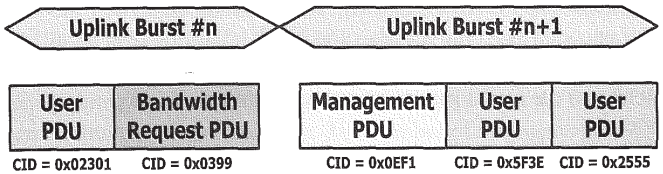


Figure 6–18: Concatenation of MPDUs

MPDU encryption and CRC

The details of the MPDU encryption are covered in Chapter 8. A summary of MPDU encryption and CRC rules are listed here:

- CRC is optional in the IEEE 802.16 MAC, although it is a good practice to include it.
- CRC is mandatory for ARQ connections, as it is used to detect errors in MPDU transmissions.
- If CRC is included, it covers both the MAC header and payload. The generic MAC header is also protected by HCS.
- If encryption is used, the generic MAC header is not encrypted.
- If both CRC and encryption are enabled, CRC is calculated after encryption.

MAC management

The format of the MAC management messages is shown in Figure 6–19. The MAC management messages are carried as (a part of) MPDU payload. All MAC management messages begin with an 8-bit Type field, followed by the message payload. An example management message payload definition is shown in Table 6–9.

The first part of the payload follows no particular set of encoding rules. In the example, it a string of 3 bytes, the two first of which are to be interpreted as the transaction ID and the third one as the confirmation code. The remainder of the message is defined as *TLV-encoded information*, which follows an encoding that closely resembles ASN.1 Basic Encoding Rules. For each parameter encoded as a type-length-value (TLV), the first byte identifies the parameter type, the following byte/bytes[2] gives the length in bytes of the value field, and the actual parameter value follows.



Figure 6–19: Format of the MAC management message

[2] If the length of the Value field is less than 128 bytes, a single byte suffices. For longer parameters, more than 1 byte is used to express the length.

Table 6–9: DSA-RSP message format

Syntax	Size	Notes
DSA-RSP_Message_Format() {		
Management Message Type = 12	8 bits	
Transaction ID	16 bits	
Confirmation Code	8 bits	
TLV Encoded Information	*variable*	TLV-specific
}		

ARQ

The following subsections describe the basic IEEE 802.16 ARQ concepts of block-based retransmissions and bitmap-based ARQ feedback. The actual SR ARQ algorithm, with its operational parameters and state machines, are described in Chapter 7.

ARQ block-based retransmissions

The concept of ARQ blocks and block-based retransmissions is illustrated in Figure 6–20. The figure shows the transmission, loss, and subsequent retransmission of two consecutive SDUs, shown in Figure 6–11, on an ARQ-enabled connection with packing. The number of blocks per SDU is based on the prenegotiated block size. During initial transmission of the SDUs, both are fragmented into two fragments each. The number of fragments per SDU and hence the number of blocks per fragment are dynamically decided based on a variety of factors. For example, the scheduler may limit the fragment size or airtime allocated to a specific connection. Even if the allocated airtime is fixed, the adaptive modulation may change the number of blocks that can be transmitted within the allocated time.

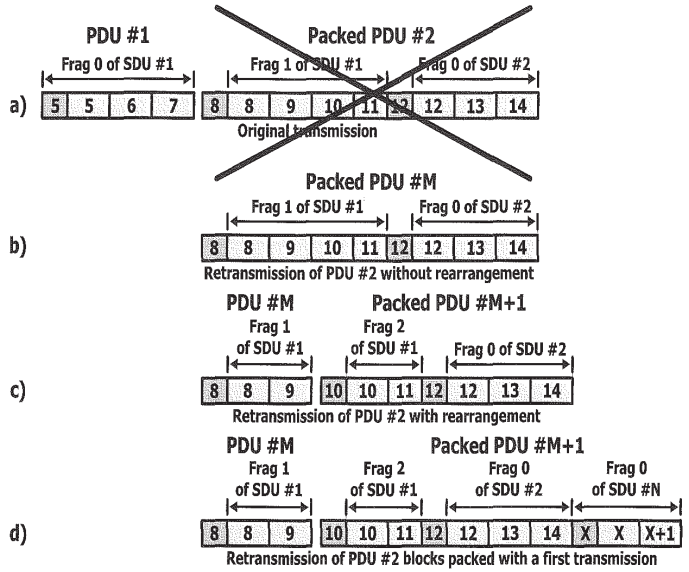


Figure 6–20: ARQ blocks and retransmissions

Figure 6–20 a) shows the construction and transmission of two MPDUs, where PDU #1 consists of fragment #0 of SDU #1, and PDU #2 consists of fragment #1 of SDU #1 and fragment #0 of SDU #2. The transmission of fragment #2 of SDU #2 is not shown in the figure. Assume it is determined (by mechanisms discussed in the next subsection) that PDU #2 did not reach the receiver. The block-based approach offers many possibilities for retransmission of the lost PDU #2. We assumed that, by the time the loss of PDU #2 is detected, fragment #1 of SDU #2 (consisting of blocks 15 and 16) has already been transmitted successfully and that SDU #*N* of the same connection, where *N* is greater than 2, is ready for first transmission.

Out of the many possibilities for the retransmission of PDU #2, three examples are shown in Figure 6–20. The first example, in Figure 6–20 b),

shows the retransmission of PDU#2 with no change or rearrangement as a new PDU #*M* (*M* > 2). The second example, in Figure 6–20 c), shows the retransmission with a rearrangement where the original PDU #2 is split into two MPDUs, *M* and *M*+1. The third example, in Figure 6–20 d), is similar to the second one, except the PDU #*M*+1 consists of the blocks to be retransmitted and the first or original transmission of fragment #0 (with BSN *X*) of SDU #*N*. As shown in Figure 6–20 d), an MPDU on an ARQ connection may contain blocks that are transmitted for the first time as well as those being retransmitted.

It is up to the transmitter's policy to decide whether a set of blocks once transmitted as a single fragment should be retransmitted as a single fragment. Similarly, it is up to the transmitter to decide whether a set of fragments once transmitted in a single MPDU should be retransmitted as a single MPDU. The advantage of keeping the number of blocks per fragment and fragments per MPDU the same between retransmissions is simplicity. On the other hand, the flexibility of refragmentation and rearrangement can significantly improve efficiency and QoS in certain traffic conditions. If fragmentation is not enabled for this connection, the ARQ fragmentation or PSH would still be used with the same block size, as if the fragmentation or packing were enabled. In this case, each "fragment" would contain all the blocks of a single SDU.

ARQ Feedback information element (IE)

Apart from the flexible retransmission, IEEE Std 802.16 defines compact ARQ feedback that is used by the receiver to provide feedback on successfully received and missing blocks. The ARQ Feedback IE contains an ACK Type field and an optional variable-size bitmap indicate positive and/or negative acknowledgments. Four types of acknowledgments are defined to support flexible feedback. Figure 6–21 shows the format of the ARQ Feedback IE.

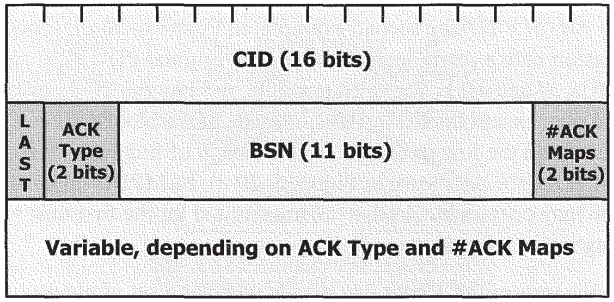


Figure 6–21: ARQ Feedback IE

The fields of the ARQ Feedback IE are described below:

- ***CID:*** The CID of the connection for which this feedback is being sent.
- ***Last:*** The Last bit is used to pack multiple ARQ Feedback IEs to create an ARQ feedback payload. The Last bit is set to zero for all ARQ Feedback IEs, except the last IE in the payload.
- ***ACK Type:*** This field indicates the type of acknowledgment being transported in this feedback IE. Table 6–10 shows the acknowledgment types and their descriptions.
- ***BSN:*** The BSN is interpreted differently based on the acknowledgment type as described in Table 6–10.
- ***# ACK MAPs:*** The value of this field plus one represents the total number of 16-bit acknowledgment bitmaps included in this feedback IE. This field is not valid for acknowledgment type 0b01 and, therefore, is set to zero.
- ***ACK MAPs:*** The number of acknowledgment MAPs is variable, and the acknowledgment type 0b01 does not have an associated acknowledgment MAP. The acknowledgment MAPs are treated as a bit stream, where the MSB is the MSB of the first acknowledgment MAP and the least significant bit (LSB) is the LSB of the last acknowledgment MAP.

Table 6–10: Acknowledgment types

Value	Type	Description
00	Selective	The BSN carried within this IE corresponds to the MSB of the first 16-bit acknowledgment bitmap. Each subsequent bit corresponds to the BSN corresponding to the previous bit plus one. No assumptions should be made about the status of any of the BSNs prior to the first BSN or the last BSN represented by the acknowledgment MAPs. If the bit is set, it indicates that the receiver has successfully received the corresponding ARQ block.
01	Cumulative	The ARQ block indicated by the BSN and all ARQ blocks prior to this within the transmission window have been successfully received. No ARQ MAP is included with this acknowledgment type.
10	Cumulative with selective	This is the combination of the previous two acknowledgment types (0b00 and 0b01). Combines the functionality of types 0x0 and 0x1. The first bit of the first acknowledgment MAP, which corresponds to the included BSN, is always set to 1, indicating cumulative acknowledgment for the BSN.
11	Cumulative with block sequence	This acknowledgment type combines the functionality of the cumulative acknowledgment type 0b10 with the ability to acknowledge a sequence of ARQ blocks in a compact format. If a set of ARQ blocks with consecutive BSN values all have the same reception status (i.e., either positive or negative), a single bit indicating the status, together with a sequence length, is used to indicate the status of the whole sequence of BSNs.

Since all the bits in a selective acknowledgment MAP must refer to an ARQ block that was successfully received or an ARQ block that was presumed lost or received incorrectly, it is important to choose the right BSN for the Selective ACK Feedback IE. The receiver must assign a BSN so that all bits in a MAP either indicate a positive acknowledgment (ACK) or a negative acknowledgment (NACK). One way to achieve this is to choose a lower BSN so that all bits are valid. For example, if only 12 ARQ status bits are available (either ACK or NACK) starting from BSN *N*, the receiver may send a selective acknowledgment with a single 16-bit MAP and a BSN of *N - 4*, where all the first 4 bits will be set to one.

The format of the block sequence acknowledgment MAP is shown in Figure 6–22. There are two block sequence formats. The first (and most significant) bit of the block sequence acknowledgment MAP indicates the format. The first format includes a 2-bit sequence acknowledgment MAP, where the bits are interpreted similarly to the selective acknowledgment type, except that each bit corresponds to a BSN sequence instead of a single BSN. Since this is also a cumulative acknowledgment, the BSN carried within this ACK Feedback IE indicates that the ARQ block with this BSN and all ARQ blocks with lesser BSN values within the transmission window have been successfully received. However, the first bit of the sequence acknowledgment MAP corresponds to the first BSN of a sequence that is being selectively acknowledged, which is the BSN carried in the feedback IE plus one. Therefore, the first bit may be a negative acknowledgment, unlike the acknowledgment type 0b10, where the first bit must always be set to one, as it corresponds to the cumulative acknowledgment BSN. The length of the each BSN sequence is indicated in the two 6-bit sequence length fields that follow. The first format can be used to indicate the selective acknowledgment status of up to 126 ARQ blocks.

The second format shown in Figure 6–22 is very similar, except that it has a 3-bit sequence acknowledgment MAP, where the length of each sequence is indicated by the 4-bit sequence length fields. Therefore, a total of up to 45 ARQ blocks can be selectively acknowledged using this 16-bit compact format.

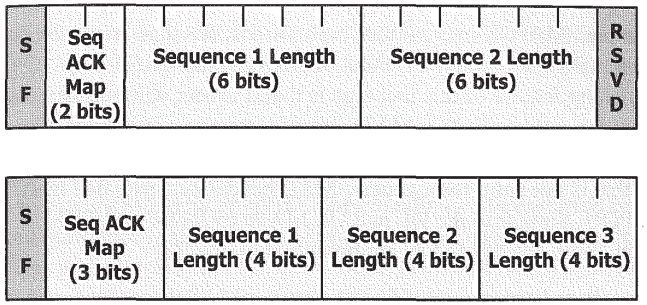


Figure 6–22: Block sequence acknowledgment MAP formats

ARQ feedback payload

The ARQ feedback information can be either sent as a stand-alone MAC management message on the appropriate basic management connection or piggybacked as a special payload on an existing connection. ARQ feedback sent in a basic management connection cannot be fragmented. When transported in a management connection, more than one ARQ Feedback IE may be sent, and the Last bit is used to properly decode the contents.

HYBRID AUTOMATIC REPEAT REQUEST (HARQ)

The HARQ is an optional part of the MAC. HARQ is supported only for the OFDMA PHY. HARQ, along with associated parameters, is negotiated during initialization of the SS. HARQ defines four basic concepts to efficiently support retransmissions:

- Compact MAP IEs that define access information for both UL and DL
- HARQ packet construction
- Reduced connection identifiers (RCIDs)
- HARQ UL/DL acknowledgments

Compact MAP IE

The Compact MAP IE may include both DL/UL IEs that define the access information for the DL and UL burst of HARQ–enabled SS. One distinguishing feature of this message is that it is sent without the generic MAC header, unlike all other MAPs, in order to save a few bytes of overhead. This is accomplished by setting the first 3 bits of the HARQ MAP message to a value that is not valid for a generic MAC header encapsulated management or data PDU. Similar to the DL and UL MAPs for TDM and TDMA operation, these HARQ MAPs are broadcast by the BS.

A BS may broadcast multiple HARQ compact MAP messages using multiple burst profiles (i.e., modulation and coding rate). The HARQ compact MAP always follows the DCD and UCD messages, if present. Otherwise, HARQ compact MAPs are sent after other DL/UL MAP messages. When multiple HARQ compact MAP bursts are present, they are ordered based on the modulation and coding used, starting from the lowest rate to the highest. A HARQ MAP Pointer IE is used by the BS to indicate to the SS the number of slots allocated to the OFDMA burst containing HARQ MAP messages.

HARQ Control IE

HARQ Control IEs that are part of the Compact MAP IEs carry the necessary encoding/decoding information of the HARQ–enabled DL/UL bursts. This information is used to properly decode the HARQ subpackets. The information contained in the HARQ Control IE includes the following:

- A 1-bit HARQ sequence number, toggled between 1 and 0
- A 2-bit SPID that identifies the subpacket as described in the following subsection
- A 4-bit HARQ channel ID that identifies the HARQ channel

Since HARQ is a stop-and-wait ARQ, a single bit sequence number is sufficient for proper operation. However, more than one HARQ channels may be defined per terminal, where each connection may have an encoded packet pending. The HARQ Control IE may also be used by the BS to request a retransmission of a subpacket.

Construction of HARQ packets

The construction of an HARQ packet involves the concatenation of one or more MPDUs, as shown in Figure 6–23. A CRC is added to this concatenated packet to form an HARQ PHY burst.

After the construction, an HARQ packet is encoded according to the PHY specification, and four subpackets are generated from the encoded result. The subpackets are identified by an SPID. The encoding and generation of four subpackets introduce redundancy among the subpackets. Therefore, the subpackets can be transmitted one at a time, in sequence, and the receiver can correctly decode the original encoder packet without having to receive all four subpackets.

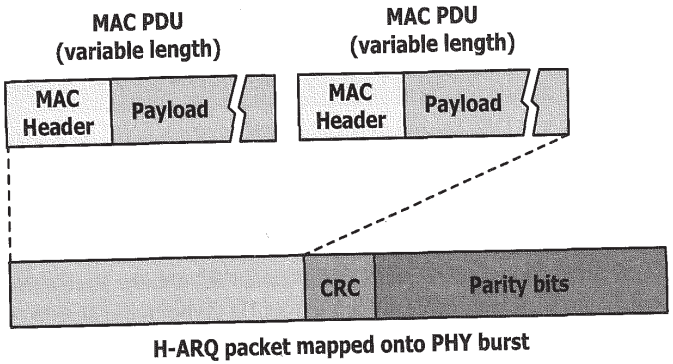


Figure 6–23: Construction of HARQ packet

Reduced connection identifier (RCID)

Figure 6–24 shows the format of a RCID. An OFDMA BS may use RCID instead of a Basic CID or multicast CID to reduce the size of MAP messages, including HARQ MAP messages. There are three types of RCIDs defined in IEEE Std 802.16: RCID-11, RCID-7, and RCID-3. The type of RCID used in a specific system is determined by BS considering the range of Basic CIDs of the SSs connected with the BS. The RCID type to be used is specified in the

RCID_Type field of the Format Configuration IE that is used in the OFDMA PHY to select format configuration.

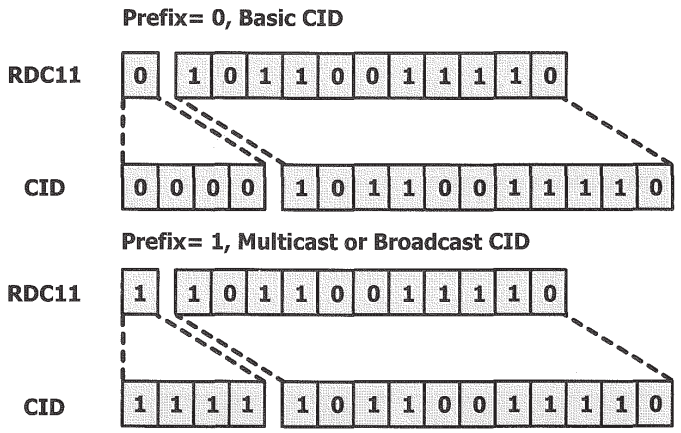


Figure 6–24: RCID decoding

The RCID is carried in a RCID_IE, whose fields are interpreted based on the RCID_Type field (from the Format Configuration IE) and the Prefix field in the RCID_IE itself, if applicable. When the RCID_Type field value is nonzero, the RCID_IE is composed of 1 bit of prefix and *n*-bits of LSB of the SS CID. The prefix is set to one for the Broadcast CID or Multicast Polling CID and set to zero for the Basic CID. If the RCID_Type field value is zero, then the RCID_IE consists of a regular 16-bit CID. Therefore, an RCID_IE can be either 16 bits, 12 bits (RCID-11), 8 bits (RCID-7), or just 4 bits (RCID-3). Table 6–11 shows the interpretation of the RCID_IE fields, where the [R]CID length and RCID_IE length are specified in bits.

Table 6–11: RCID field interpretation

RCID_Type	Prefix	[R]CID length (bit)	RCID_IE length (bit)	Comments
0	NA	16	16	This is the same as normal CID
NA	1	11	12	11 LSB of Multicast, AAS, or Broadcast CIDs
1	0	11	12	11 LSB of Basic CID
2	0	7	8	7 LSB of Basic CID
3	0	3	4	3 LSB of Basic CID

The RCID cannot be used in place of the Transport CID, Primary Management CID, or Secondary Management CID.[3] For the purpose of per-SS HARQ, an SS is always identified by its Basic CID (or a corresponding RCID for the Basic CID).

HARQ acknowledgments

The HARQ acknowledgments are handled differently in the UL and DL. The BS uses a Compact MAP IE to send HARQ acknowledgment bitmap information for the HARQ UL bursts from the SS. The BS specifies a "frame offset" that indicates the offset between the UL HARQ burst and the corresponding HARQ bitmap. This frame offset *f* is communicated in the HARQ_ACK_ Delay_for_UL_Burst field of the DCD messages. The SS uses the frame offset to correlate acknowledgment bitmaps to the HARQ bursts. For an HARQ UL burst in the i^{th} frame corresponding to the j^{th} HARQ burst in the UL MAP, the BS uses the j^{th} bit (LSB) of the HARQ bitmap sent in the $(i+f)^{th}$ frame to indicate the status. The DL Compact MAP IE (type = 5) that

[3] For other uses of RCID in OFDMA, an exception to this rule is made for the Multicast Polling RCID in the DL. If a DL CID with prefix 1 and RCID 11 is in the range of the Multicast Polling CID (0xFF00–0xFFFD), then the DL CID is interpreted as a DL Transport CID by subtracting 0xFF (0xFE01–0xFEFE).

carries the HARQ bitmap consists of a Length field that indicates the length of the bitmap and the actual bitmap.

The DL HARQ bursts are acknowledged in the HARQ region allocated by the BS in a UL Compact MAP IE. The OFDMA subchannels in the HARQ region are divided into two half-subchannels, where the $(2n)^{th}$ half-subchannel is the first half-subchannel and the $(2n+1)^{th}$ half-subchannel is the second half-subchannel of the n^{th} subchannel. An HARQ DL burst at the i^{th} frame is acknowledged by an SS through the half-subchannel in the HARQ region at the $(i+f)^{th}$ frame, where the frame offset f is defined by the HARQ_ACK_Delay_for_DL_Burst field in the UCD message. The order of DL HARQ burst defined in the HARQ MAP determines the half-subchannel offset in the HARQ region. For an HARQ–enabled burst at the i^{th} frame that is the n^{th} HARQ–enabled burst in the HARQ MAP, the SS transmits the HARQ acknowledgment at the n^{th} half-subchannel in the HARQ region that is allocated by the BS at the $(i+f)^{th}$ frame.

Chapter 7 MAC operation

Radio control, QoS, and ARQ

NETWORK ENTRY AND INITIALIZATION

Once the SS has powered up, it begins the network entry and initialization process. After completing the steps of the process, the SS has all the addresses and parameters it needs to communicate with the rest of the network.

Scanning and synchronization to the DL

Before the SS can do anything else, it has to first find a BS transmitting a signal that it is capable of decoding and understanding. To accomplish this, the SS consults the frequency lists with which it shipped, chooses one of the channels mentioned in the list, and starts to search for the periodically occurring frame preamble. If the SS finds the preamble, it tries to determine the DL transmission parameters by looking at the frame control header (FCH) and the DCD message, which is transmitted periodically but not in every frame. Also, the SS learns the specified UL parameters from the UCD messages. Afterwards, it has sufficient knowledge to transmit on the UL, and initial ranging may begin.

Initial ranging

Initial ranging accomplishes the following:

- The timing advance of SS transmissions is adjusted to make the SS appear collocated with the BS.
- The transmission power of the SS is adjusted for optimal reception at the BS.
- The SS is allocated its Basic and Primary Management CIDs.
- In some systems, the frequency of SS UL transmissions is adjusted.

Initial ranging occurs in a part of the frame that is set aside for this purpose and partitioned into initial ranging slots. The slots need to be large enough to accommodate the ranging transmission and twice the propagation delay to the edge of the cell, since the distance between SS and BS is not known at this time. The exact details of the initial ranging procedure depend on the underlying PHY, but in short the process goes as follows: First, the BS signals the available initial ranging slots in the UL MAP. After determining the location in time of the initial ranging slots, the SS randomly selects a slot and sends an RNG-REQ (ranging request) message. If this is received correctly, the BS replies with a message specifying the power, timing, and possibly frequency adjustments the SS needs to make prior to its next transmission. This message can also instruct the SS to repeat the initial ranging process. However, the BS would normally want to instead fine-tune the SS transmission parameters using one of the maintenance ranging procedures, as they are less resource-consuming than initial ranging.

Now, there is always the possibility that two terminals may send their RNG-REQs in the same slot, in which case the BS may be unable to decode either transmission. Also, it might be that the SS transmission fails because its power is too low (noting that the SS will initially use low power so it does not disrupt transmission in neighboring cells). In any case, if the SS does not hear from the BS within a certain interval, it will increase the power by a notch and try again after waiting for a randomly chosen number of initial ranging slots. In addition, for some PHYs, if the BS can detect energy but fails to decode the initial ranging message, it can signal for the SS to retransmit and make adjustments, identifying the SS in the RNG-RSP (ranging response) message by the transmit slot.

SS basic capability negotiation

Once the SS timing and transmit power have been adjusted, the SS needs to tell the BS which optional functionality it supports. Also, the BS needs to tell the SS which of those options the SS is allowed to use. The protocol by which the support and use of options are agreed upon is called *SS basic capability negotiation*. As this name suggests, this basic negotiation covers not all options, but only those having to do with fundamental MAC and PHY features, such as maximum transmit power, modulation schemes, FEC codes,

and support of MAC bandwidth allocation schemes. The idea with exchanging these parameters at an early stage is to make the rest of the network entry process more resource-efficient.

The protocol itself is a straightforward request-grant exchange, as shown in Figure 7–1. The timer T18 has a default value of 50 ms and should be significantly less than the registration timeout (i.e., the timeout at the BS for the maximum time between a RNG-RSP to an SS upon successful ranging and the receipt of an SBC-REQ (SS basic capability request) message from the same SS). In the request message (SBC-REQ), the SS will list all the "basic capabilities" it supports. The BS inspects the list of capabilities and, in the SBC-RSP (SS basic capability response) message, informs the SS of those it does not support or does not want to use.

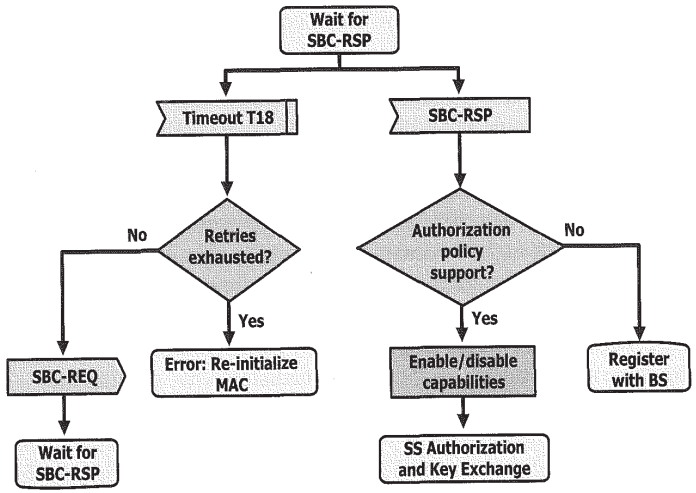


Figure 7–1: SS basic capability negotiation

Note that some of the most fundamental parameters of operation are not subject to negotiation with this protocol. For example, with different PHY system profiles, establishing communication would be challenging or impossible.

Authorization, security association (SA) establishment, and key exchange

At this stage of the network entry process, the SS should be capable of communicating efficiently with the BS. The BS, however, does not yet know with certainty the identity of the SS with which it is dealing. The SS has up to this stage provided its IEEE MAC address in the RNG-REQ, but has offered no credentials verifying its identity. The authorization phase establishes the identity of the SS, the authorization key (AK), and the list of SAs that the SS is authorized to use. To achieve this, well-known public key cryptographic methods are used. For a treatise on these, consult [B45].

The authorization protocol starts with the SS sending the authentication information privacy key management (PKM) message. All PKM messages are encapsulated in either the PKM-REQ (SS to BS) or the PKM-RSP (BS to SS) management message. The authentication information message is strictly informational, but is included to provide a mechanism by which a BS can learn the X.509 certificate of the SS manufacturer in a standard way. However, the IEEE 802.16 standard says nothing about how the BS can authenticate the certificate. It is envisaged that the industry will set up the needed infrastructure for this authentication if this method is to be used in practice. Meanwhile, the assumption is that the manufacturer certificate ends up, by some secure means, in the BS. Immediately after sending the authentication information message, the SS sends the Authorization Request message encapsulated in a PKM-REQ management message containing the X.509 certificate of the SS together with a list of the SS security capabilities. The BS proceeds to authenticate the certificate by checking the signature of the manufacturer. Assuming the certificate is in order, the BS generates the AK and encrypts it using the public key of the SS contained in the X.509 certificate. Additionally, the BS determines which SAs the SS is entitled to use. The AK, the key lifetime, and the list of SAs along with their parameters are sent by the BS to SS in the Authorization Reply message. Also, the BS has

the option to send any PKM protocol parameters that are different from the default values given in the IEEE 802.16 standard.

After receiving the Authorization Reply message and decrypting the AK, the SS becomes authorized to operate normally in the network. However, it still lacks the capability to transmit anything besides MAC management messages. Before becoming fully operational, the SS still needs to (a) acquire a key for each SA given by the BS using the protocol given in the SA parameter perform registration, (b) possibly acquire an IP address, and (c) establish the MAC connections. Each SA's keys, called *traffic encryption keys* (TEKs), are established by the SS sending a digitally signed Key Request PKM message and the BS returning the encrypted key in a Key Reply message, also signed with a key derived from the AK. The algorithm used in the encryption of the TEK is triple data encryption standard (3DES), AES, or RSA, depending on algorithm support of the SS and the parameters of the SA.

Registration

During the registration phase, the SS and BS negotiate additional operational parameters of the MAC. Also, the SS informs the BS whether it will be part of the managed network. If the SS is to be managed, the bidirectional secondary management connection is established between the SS and BS. The registration message exchange also offers the possibility to negotiate the version of IP and the QoS parameters for the secondary management connection. The registration process is shown in Figure 7–2. The SS may wait for up to 3 s (timer T6 in Figure 7–2) for the registration response before resending the request or reinitializing the MAC.

Establish IP connectivity

A managed SS also needs to acquire an IP address (using DHCP), pull down a configuration file (using TFTP), and establish the time of day (using the Internet Time Protocol). Afterwards, the SS sends a message to report that the tasks were successfully completed.

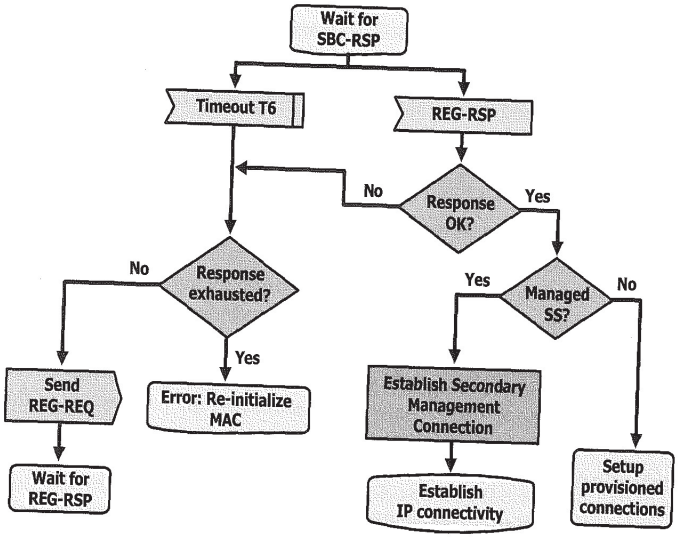


Figure 7–2: SS registration

Dynamic service establishment

The final step before actual data communication can begin is dynamic service establishment. In the case of a managed SS entering the network, the reception of the TFTP-CPLT ([configuration file] TFTP complete) message triggers the BS to start connection setup. When dealing with unmanaged SSs, the successful completion of registration serves as the trigger. The service flows for a given SS are assumed to have been preprovisioned by the BS. To activate them or take them to the admitted state, the BS initiates a dynamic service addition (DSA) exchange, as illustrated in Figure 7–3. After at least one service flow has been activated, the SS is capable of sending and receiving user data. The DSA signaling can also support setting up the classifiers and PHS on the CS.

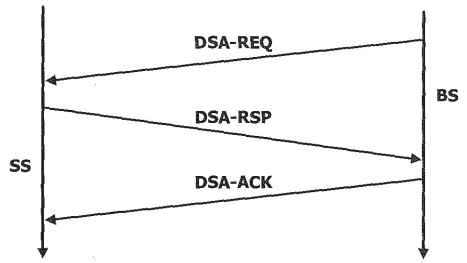


Figure 7–3: DSA message flow (BS-initiated)

PHY MAINTENANCE

Using IEEE Std 802.16, the PHY parameters used to communicate between the BS and an SS may change rapidly. These changes may be caused by myriad events that may differ depending on the operational frequency and choice of PHY specification. For instance, at 28 GHz the WirelessMAN-SC PHY is susceptible to rain fade, while at 2.5 GHz the WirelessMAN-OFDM PHY is susceptible to changes in multipath. How often PHY quality needs to be assessed and how often PHY parameters may need to change often depend on the environment of the system. For instance, at 28 GHz, the rain region may dictate the maximum rain fade, while at 2.5 GHz, population densities and urban or rural physical settings may impact multipath change. Many PHY adjustments are possible. For instance, power control may be applied, or the FEC or modulation scheme may change. Adjustments may occur in real time. Some of these changes, most notably FEC and modulation, impact the availability of bandwidth in real time. This aspect of IEEE Std 802.16, coupled with a need for true QoS, has blurred the lines between the classical partitioning of the MAC and the radio link control (RLC) layer. In the IEEE 802.16 standard, little has been done to prevent this blurring. In fact, the standard does not specifically discuss an RLC layer, but merely handles it as an extension of the MAC. Some messages, such as MAPs, span both layers by simultaneously allocating bandwidth and modifying PHY parameters.

MAPs and channel descriptors

In IEEE Std 802.16, all four currently defined PHYs share some common attributes:

- They are all framed PHYs, providing deterministic, periodic opportunities for resynchronization and centralized scheduling.
- They can be used in either TDD or FDD mode.
- In FDD mode, SSs may be half-duplex (H-FDD).
- They all use MAPs to describe the bandwidth usage and allocations in both the DL and the UL.
- They all use channel descriptors to associate sets of PHY parameters to DIUCs and UIUCs allowing a shorthand to make the MAPs more efficient.

In both the TDD and FDD cases, the PHY frame duration is selected by the BS from one of the allowed frame durations. Typical durations, depending on the PHY, range from 0.5 ms to 20 ms, with 1 ms being common for WirelessMAN-SC systems and 5 ms being typical in the case of WirelessMAN-OFDM. The duration is expressed to the SSs or discovered by them in a PHY-dependent way.

The PHY frame is further divided into a DL subframe and an UL subframe. In the TDD case, the frame starts with the DL subframe and ends with the UL subframe. Between the two is a time gap (TTG) allowing devices the time to shift from transmit to receive, or vice versa. Similarly, another time gap (RTG) falls between the end of the UL subframe and the start of the next DL subframe. When scheduling the end of the DL subframe and the beginning of the UL subframe, the BS must take care to respect the half-duplex nature of TDD devices. Two additional time durations besides the TTG and RTG are of importance, as well, as described in the following two paragraphs.

First, the BS must be aware of round-trip delay because SSs advance their transmissions so that they appear to be collocated with the BS from a timing point of view. The BS has two choices, depending on the complexity of the scheduling within the BS. Either it may use the concept of a maximum round-trip delay based on the planned cell size, or it may track the round-trip delay of each SS individually and use this when scheduling. The maximum

round-trip delay is a useful simplification in fixed systems, but it can cause problems if nomadic or mobile stations are allowed. It may also complicate adding an additional subscriber just beyond the current cell boundary. The use of a tracked round-trip delay is more robust to changing conditions, but it can add complexity to the BS scheduler.

The second important time duration is T_{proc}. This is the amount of time allowed for an SS to process the UL MAP after receipt. In systems that allow the UL MAP to apply to the current frame (WirelessMAN-SC and WirelessMAN-OFDM, but not WirelessMAN-SCa or WirelessMAN-OFDMA), the BS further needs to ensure that the SS has had sufficient time to receive and process the UL MAP before it is required to transmit in the UL. More on the time relevance of MAPs will be given in a few paragraphs. An example of a TDD PHY frame for a TDM system, taken from the WirelessMAN-SCa PHY, is shown in Figure 7–4.

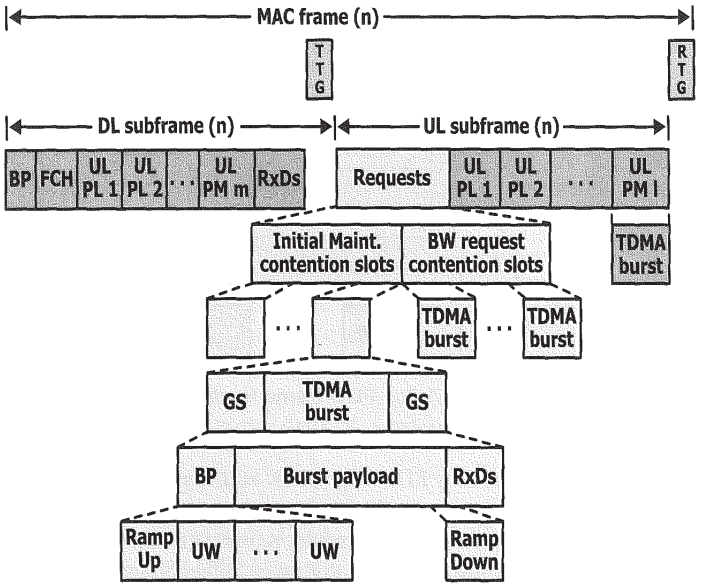


Figure 7–4: TDD frame format

In the FDD case, the PHY frame can be viewed as being the same as the TDD case, except that the DL subframe is stretched to end later and the UL subframe is stretched to start earlier. Therefore, the two effectively overlap. The structure of each subframe stays basically the same as the TDD case. In FDD, the standard allows two types of SSs: full-duplex and half-duplex. From the BS point of view, full-duplex SSs are relatively easy to schedule since there is no need to worry about receive-transmit exclusions. An example of a FDD PHY frame for a TDM system, again taken from the WirelessMAN-SCa frame, is shown in Figure 7–5.

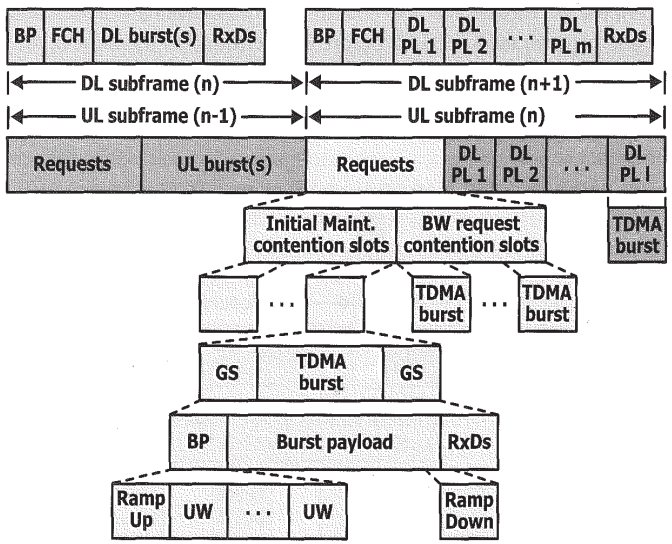


Figure 7–5: FDD frame format

However, we expect that many FDD SSs will be half-duplex (H-FDD) to save cost. The BS scheduler must, as in the TDD case, respect the half-duplex nature of the SSs. Therefore, the BS scheduler must be aware of the round-trip delay, turn around times, etc., to ensure that an H-FDD SS is not requested to transmit and receive at the same time. Additionally, if an H-FDD terminal is

scheduled to receive later in the frame than it transmits, it may lose synchronization during transmission and, therefore, may require a method to resynchronize with the DL prior to reception. Due to certain benefits of the OFDM symbol structure, this is usually of concern only for WirelessMAN-SC and WirelessMAN-SCa. For these two single-carrier PHYs, resynchronization points in the form of additional preambles provide the midframe resynchronization. Figure 7–6 shows the modified DL frame structure, taken from WirelessMAN-SC, for FDD with a preponderance of H-FDD SSs.

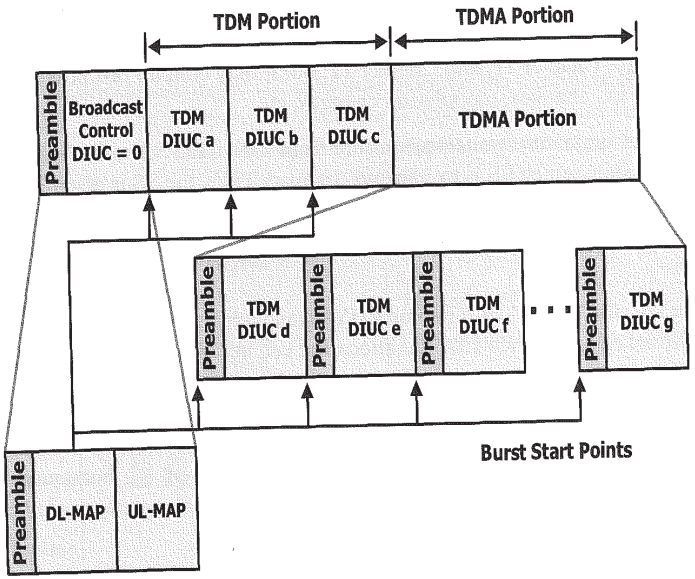


Figure 7–6: FDD DL subframe structure

For both the DL and UL subframes, the BS transmits MAPs that define the allocations of bandwidth in the subframe. The MAPs can take a variety of forms depending on the PHY and the options enabled. In the simplest form, there is a single DL MAP and a single UL MAP for a PHY frame. Certain

PHYs may add additional concepts. For instance, private MAPs are used for AAS zones, and a number of other PHY- and option-specific MAP variants are specified. Some of these MAP variants are caused by PHY-related issues concerning whether an SS can hear broadcast MAPs. In all cases, the basic concept is that the DL and UL subframes are described by MAPs.

At first glance, this appears to be merely a scheduling exercise determining the timing of the transmissions. But the content of the MAPs reaches beyond simply scheduling the time (and subchannels of OFDM-based PHYs) on which an SS sends or receives. The MAPs contain content impacting the PHY. Exactly what PHY information may be contained in the MAPs depends on the PHY and is described in more detail in the PHY-specific chapters, Chapter 10, Chapter 11, and Chapter 12; but it may contain fast power control, FEC changes, modulation changes, etc.

In the UL, since most transmissions are from individual SS's, the MAP elements state which SS is to transmit. In every PHY, the SS is also instructed which modulation and FEC to use. The SS reacts to any parameter changes, using the new information for the time (and subcarriers) covered by the MAP entry.

In the DL, the DL MAPs for some PHYs may address MAP elements to individual SSs, in which case the allocation can act as a quick change of DL burst profile. In these cases, the SS typically has negotiated with the BS a most robust DL burst profile to ensure the signal quality it needs for reliable operation. For some PHYs, the DL MAP entries are not directed to a specific terminal. In these cases, the SSs have a negotiated current DL burst profile and listen to all DL portions they can, including those at that negotiated burst profile and those that are more robust.

DL MAPs are always valid for the DL subframe in which they are carried. This can require some very fast processing, and PHY-specific features may be specified to allow the processing to occur fast enough. For instance, the WirelessMAN-OFDM PHY uses a DLFP in each frame, followed in only some frames by a DL MAP message. For the WirelessMAN-SC PHY, the DL MAP message is always transmitted at the beginning of the DL subframe. The start of the WirelessMAN-SC DL subframe is structured so that at least two noninterleaved FEC blocks of a well-known modulation and FEC will be

present. The first DL burst profile transition that is expressed in the DL MAP is always included in the first of these FEC blocks, allowing the SS the duration of the second FEC block for decoding, processing, and taking action.

The timing of the UL MAP is more complicated and is related to the burst profile, the duplexing scheme, and possibly the network setup. At the simple level, the UL MAP describes the UL allocations for a period starting at a point called the *allocation start time*. This is relative to the start of the DL subframe in which the UL MAP was received.

For WirelessMAN-SC and WirelessMAN-OFDM systems, the limits on the allocation start time are slightly different, depending on whether the system is TDD or FDD, but they basically have the same result: The UL MAP can describe the allocations in the UL subframe of either the current or the next PHY frame. For TDD systems, in other words, the MAP allocates the use of the time just after the TDD split of either the current or next frame, as shown in Figure 7–7.

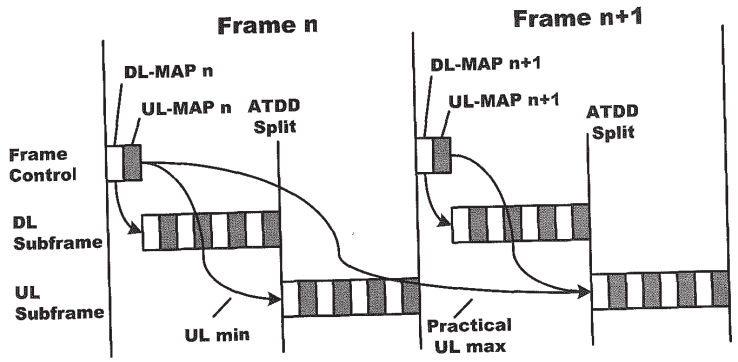


Figure 7–7: Practical TDD UL minimum-maximum relevance

For FDD systems using one of these two PHYs, the situation is a bit more complicated. The minimum delay is as small as the round-trip delay plus

T_{proc}. The latest allocation start time is the beginning of the next frame. Using this latest possible start time, the UL subframe period is aligned with the DL subframe following the one carrying the relevant UL MAP. With an earlier start time, the UL subframe is offset in time from the DL subframe. This is shown in Figure 7–8.

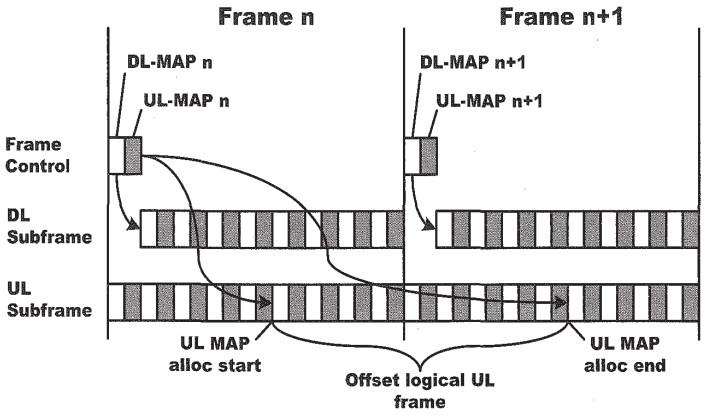


Figure 7–8: FDD logical offset

For WirelessMAN-SCa and WirelessMAN-OFDMA, the concept is similar, except the UL MAP messages are relevant for the next frame in TDD systems and the next one or two frames later in the FDD case. It is also possible that the UL MAPs sent in two different frames overlap. In such a case, the most recently received UL MAP takes precedence.

The reasons for the differences in UL MAP relevance between the PHYs are historical. The WirelessMAN-SC and WirelessMAN-OFDM systems can offer a shorter turnaround time on protocols requiring responses and can offer a shorter request/grant cycle. On the other hand, the BS scheduler and the implementation of UL MAP processing at the SS must be correspondingly more complex in able to respond with due speed.

Because there are so many options for how to configure the PHY in a given implementation, a shorthand was developed to ensure that the length of the UL and DL MAPs did not get out of hand from repeating the PHY parameters. As mentioned above, the MAPs contain codes that indicate a set of PHY parameters in the compact form of DIUCs or UIUCs. Periodically, the BS broadcasts DCD messages and UCD messages. These messages associate the chosen set of PHY parameters with a DIUC or UIUC, respectively. Then, these DIUCs and UIUCs can be used in MAP messages as if the MAP message contained the entire set of PHY parameters. While the sets of PHY parameters for a particular network are generally not expected to change, a protocol is defined for distributing new sets of parameters and coordinating the change from the old set to the new set.

Periodic ranging

In order to ensure effective UL communication, ongoing maintenance is required regarding the parameters adjusted during initial ranging, such as the transmit time advance, the power level, and the burst profile. This process is called *periodic ranging* and is used to compensate for clock drift, changing environmental conditions, etc. For most IEEE 802.16 PHYs, this involves the BS periodically granting UL bandwidth to the SS. If the BS senses that the SSs PHY parameters need to be changed, it can generate an unsolicited RNG-RSP message that can direct certain changes. Timing advance and power adjustments are fairly straightforward. The actual process for changing the burst profile is a bit more complex and is described in the next section. The process generally proceeds as follows:

- The BS must determine how often it needs to hear an UL transmission from each SS in order to maintain the UL within limits that allow continued communication. This maximum time between UL transmissions can be based on a number of factors, such as frequency band of operation; the likelihood of change in environmental conditions (e.g., rain region for WirelessMAN-SC); whether the setting is urban, suburban, or rural; etc.
- Any UL transmission from the SS should suffice. Therefore, as long as the SS has UL activity for any reason (e.g., user data, other management messages), there is no need for the BS to grant bandwidth specifically for

periodic ranging. The BS merely takes measurements from the scheduled UL transmissions, resetting its periodic ranging timer for the SS with each UL transmission. If, however, the SS has no other need for a UL transmission opportunity within the time necessary for UL PHY maintenance, the BS grants it UL bandwidth anyway. Since this bandwidth is not in response to a direct request to transmit user data or management messages in the UL, it is termed an *unsolicited grant*. The process of using normally granted UL bandwidth and unsolicited grants to manage the timing of unsolicited grants is shown in Figure 7–9.

- If, upon analysis of a UL transmission, the BS determines that the SS's UL transmission parameters need to be adjusted, it sends an RNG-RSP message to the SS instructing it to adjust its parameters. If this RNG-RSP indicates "continue," the SS responds with a RNG-REQ to keep the iterative PHY maintenance process in this quick response state until the BS issues a RNG-RSP message indicating "success."
- If the SS has not been granted UL bandwidth in a sufficient amount of time or if the BS sends it a RNG-RSP message indicating "abort," the SS must fall back to periodic ranging to avoid detrimental impact on the UL transmission of other SSs until it gets its UL transmission parameters back with a safe window.

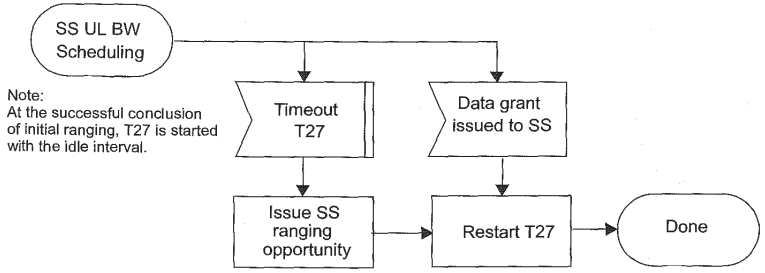


Figure 7–9: Periodic ranging opportunity allocation at BS

The BS side of the UL periodic flow is shown in Figure 7–10. The SS side is shown in Figure 7–11.

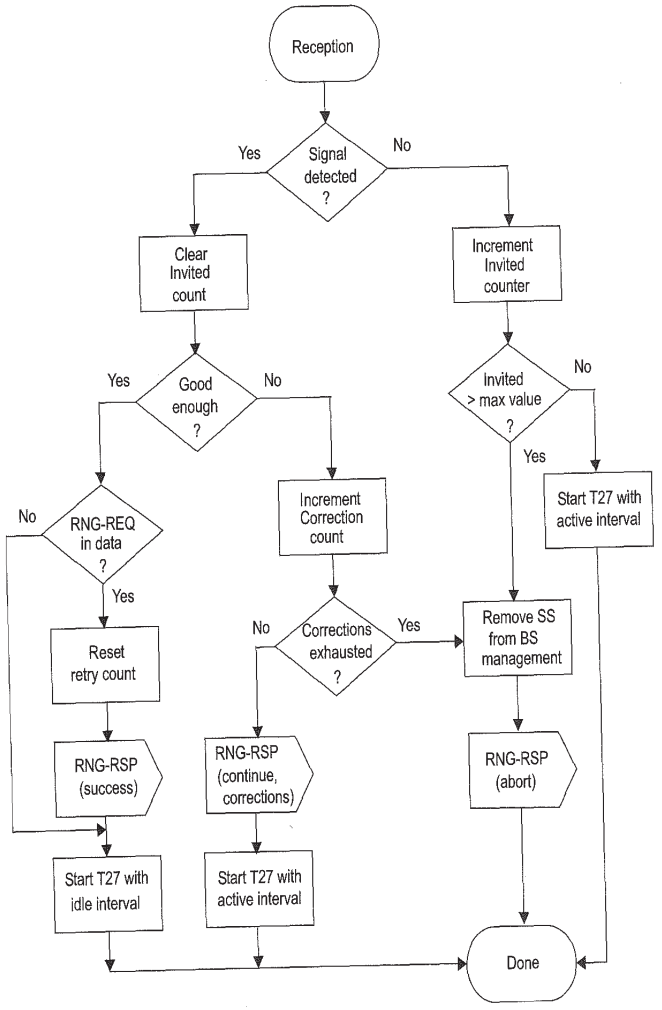


Figure 7–10: Periodic ranging receiver processing at BS

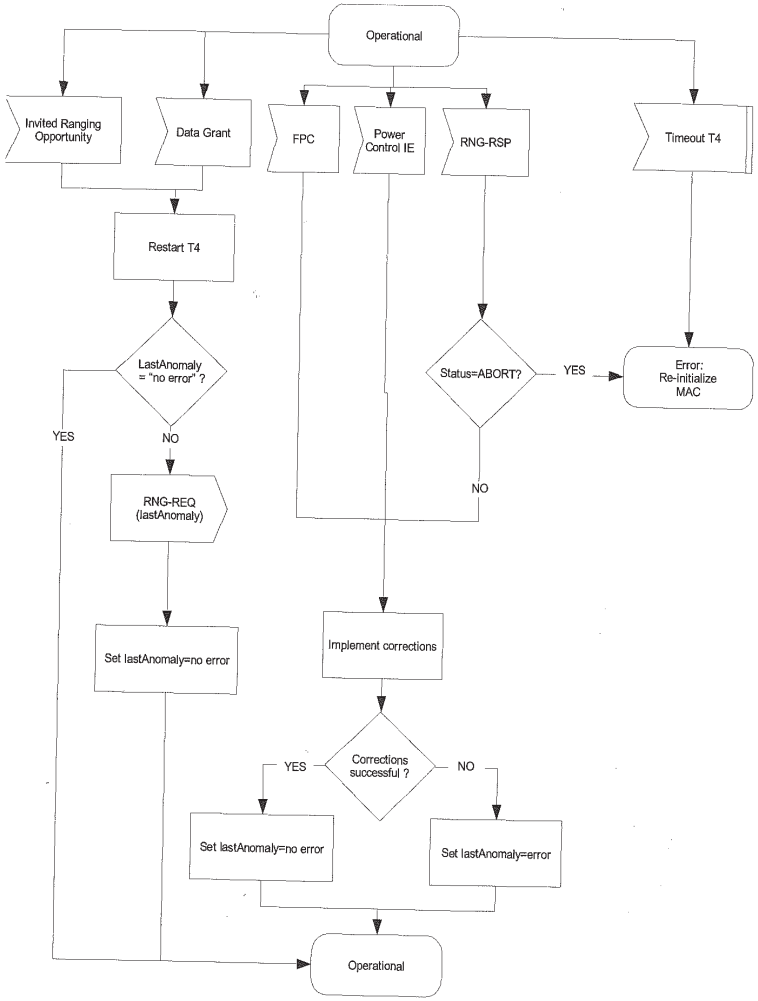


Figure 7–11: Periodic ranging at SS

For the WirelessMAN-OFDMA PHY, while the BS can initiate ranging as described, the SS also has the option of initiating CDMA-based ranging. This allows a distribution of the timer function to the SSs, easing BS requirements while taking some control away from the BS in the management of the system. In CDMA periodic ranging, the SS sends a randomly chosen ranging code in a randomly chosen contention-based ranging slot. Since the BS does not know which SS sent the code, it broadcasts a RNG-RSP advertising the code rather than directing the RNG-RSP to the specific SS. This method can be used to adjust the SS's power and timing advance. However, since the BS does not know the identity of the SS, it must still use measurements on the UL allocations to determine whether a burst profile change is necessary. The UL CDMA periodic ranging process is shown in Figure 7–12.

Burst profile changes

When the BS is communicating with an SS, it may use a UL burst profile differing from that of the DL. The burst profile used for communication with a particular SS is selected by the BS based on a number of considerations, such as the carrier to interference-plus-noise ratio (CINR), or the bit error rate (BER), or even the packet error rate (PER). The burst profile is always selected from the set described by the most recent DCD and UCD messages.

The initial burst profile is normally negotiated during the initial ranging process. The BS and SS determine the initial quality of the link, with the SS requesting a DL burst profile using the RNG-REQ message and the BS commanding a UL burst profile using the UL MAP. As time goes by, the burst profile used for transmissions to or from a particular SS may need to change. This may be due to changing environmental conditions such as rain or changes in multipath. The change may be either for the better—the conditions have improved to where a more efficient burst profile is practical—or for the worse—the conditions have degraded to the point where a more robust burst profile is necessary.

How these changes are implemented is different for the UL and the DL. In the UL, the BS monitors the UL quality from each of the registered SSs. The BS must give each SS a transmit opportunity at least often enough to adjust the parameters fast enough to keep up with environmental changes. For instance,

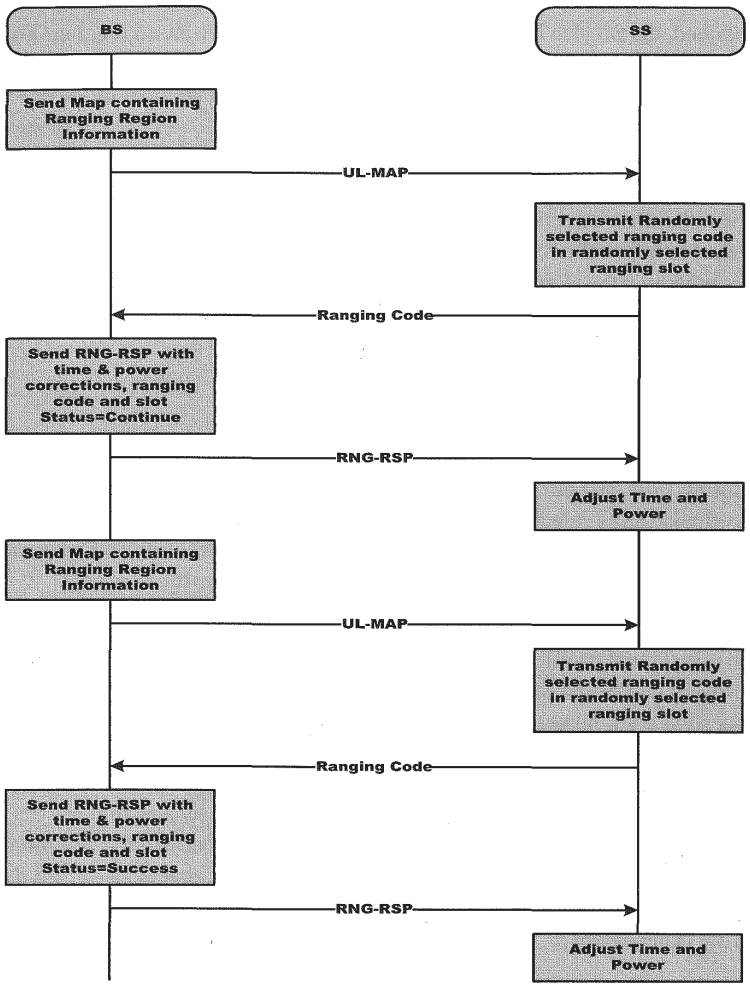


Figure 7–12: CDMA periodic ranging

for WirelessMAN-SC, where rain fade is the biggest physical degradation culprit, the BS generally needs to receive something from an SS every 80 or so milliseconds; otherwise, it may drop back to the most robust burst profile to ensure it can be heard. This monitoring constraint is a small burden, but the actual change is easy. Since each allocation in a UL MAP specifies the SS and the UIUC, the BS commands a change by simply allocating bandwidth for the SS using a new UIUC, from the set previously defined in the UCD message, in the UL MAP.

The DL can be more complicated. In the DL, while the SS has the advantage (except in certain AAS cases) of always hearing the broadcast portion of the DL, it constantly updates its measure of the DL quality from its point of view. The problem is informing the BS of a need for a change. Typically, the SS will request a change via either a downlink burst profile change (DBPC) message, if it has bandwidth allocated on the UL, or using an RNG-REQ message in a ranging interval, if it does not have its own UL allocation. For some situations with some PHYs, the BS may also periodically request PHY metrics from the SS to fine-tune such issues as which subchannels to use. For MAP efficiency purposes, the DL MAP, unlike the UL MAP, for most IEEE 802.16 PHYs does not direct a DL interval to an individual SS. The SS merely decodes all data that are as robust or more robust than the last negotiated DL burst profile. This is an important point. It allows the BS to send data to the SS using a more robust burst profile before first communicating its intent to switch, as shown in Figure 7–13. It also allows the BS to tell the SS to start listening to a less robust burst profile and then immediately begin sending data with that profile, as shown in Figure 7–14. This avoids complicated synchronization of the messages and the burst profile changes. In some cases, with some PHYs, the DL bursts are identified in the DL MAP as being directed to a specific SS. In these cases, the process works much like the UL process.

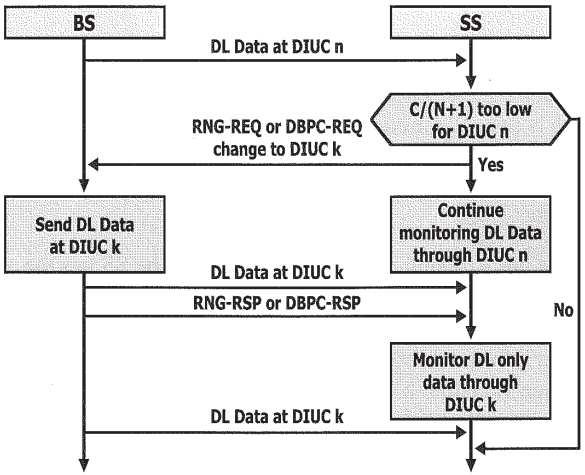


Figure 7–13: Transition to more robust burst profile

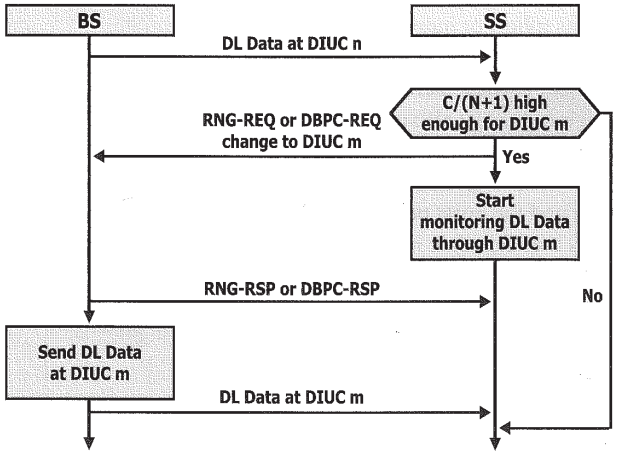


Figure 7–14: Transition to less robust burst profile

QoS AND SERVICE FLOWS

Dynamic service establishment and deletion

Besides the BS-initiated DSA described in the network entry and initialization section at the beginning of this chapter, other mechanisms are defined for SS-initiated DSA, as illustrated in Figure 7–15.

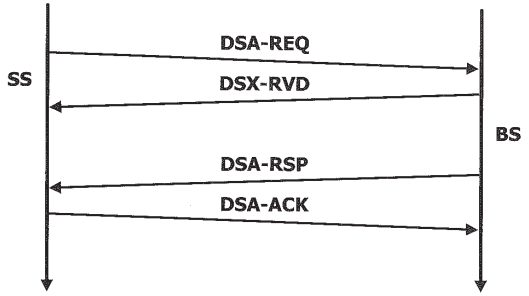


Figure 7–15: DSA message flow—SS initiated

The standard does not go into details on what actually triggers the DSA; triggering is just assumed to happen, stimulated by the upper layers when needed. For example, a nomadic device may be powered on in a network, similar to Wi-Fi networks that might be found in an airport or other public place today. The SS, being unknown to the network, asks for service. Authentication includes validating the user's credit card. Looking forward to IEEE Std 802.16e and mobility, another example would be a mobile VoIP call initiated at the MS.

One noteworthy detail in the procedure is the DSX-RVD (dynamic service addition, change, or deletion received) message that only acknowledges the reception of the DSA-REQ (dynamic service addition request) message by the BS. This allows the BS to take its time determining whether to admit the service flow without triggering unnecessary retransmissions of the DSA-REQ.

There may also be a need to change the parameters and state of the existing service flow in response to external circumstances. This can be done with the dynamic service change (DSC) procedure that is almost identical to the DSA.

Service flows are deleted using the DSD-REQ (dynamic service deletion request) message and the DSD-RSP (dynamic service deletion response) message. Deletion can be initiated by either BS or SS. Only one service flow can be deleted per message exchange.

QoS model

When IEEE Std 802.16 was in its early stages, the primary focus was from companies targeting last-mile extension of fiber and the PMP backhaul of cellular sites. At that time, IP core networks were considered interesting, but ATM core networks were the norm. Furthermore, ATM was highly advanced in its QoS support. The result was that QoS in IEEE Std 802.16 was modeled after ATM. Some of the concepts were modified based on input from proprietary system proposals and based on proposals for "wireless DOCSIS." As it turned out, ATM, DOCSIS, and the proprietary systems are compatible in their QoS concepts.

While extension of ATM networks was an early market driver, IP was considered to be the future, and the developers of IEEE Std 802.16 did not want to tie the standard too closely to ATM. DOCSIS, as it turned out, was forward-looking in its application of QoS and traffic parameters and was compatible with both ATM and IP services. Because of this, the QoS description and naming conventions from DOCSIS were chosen as the basis for describing QoS in IEEE Std 802.16. Other parts of the DOCSIS model were adopted as well, including the three-state model in which a service can be provisioned, admitted, or active. The state impacts the actual reservation state of the resources necessary for the service.

In order for QoS and traffic parameters to be fully applied to services, IEEE Std 802.16 MAPs services, including connectionless services, to an IEEE 802.16 MAC connection. This is done in basically the same manner as in ATM and in the UL of DOCSIS. Similar concepts are employed in Ethernet virtual circuits in metro Ethernet. Typically, a service is bidirectional,

resulting in the mapping to separate UL and DL IEEE 802.16 MAC connections.

The benefit of connection-oriented mapping of services is that it provides a "handle" with which the MAC, in particular the scheduler and QoS engines, can associate a packet with the full realm of QoS and traffic parameters, instead of just simple individual packet parameters such as a priority. Since no packets are expected until a service is active and no resources are reserved until a service is at least admitted, services in the provisioned state are not mapped to IEEE 802.16 connections. The services are mapped only when they become admitted or active.

QoS and traffic parameters

A number of QoS and traffic parameters, which the IEEE 802.16 standard calls *service flow encodings*, are available to help in the implementation of QoS within an IEEE 802.16 system.

SFID and CID

The SFID and CID are important concepts at the foundation of IEEE 802.16 QoS. The SFID identifies a service, which in turn identifies the right of an IEEE 802.16 SS to certain system resources, and also defines which of a user's packets will be mapped to the corresponding IEEE 802.16 MAC connection. The CID is then used internally in the IEEE 802.16 MAC for determining which SS receives the data on the DL, which service it maps to on the UL, and, most importantly, what QoS and traffic parameters are associated with the traffic. The mapping of potentially connectionless traffic to MAC connections is called *classification*. It typically involves some amount of packet inspection and can work much like a combination of a router and a bridge. Classification of IP, VLAN, Ethernet, and ATM to IEEE 802.16 MAC CIDs is currently defined in the standard.

Service class name

The service class name is a parameter that implies a predefined set of other QoS parameters. There is not much demand to define such common service

types yet. Once specific methods for handling certain services is agreed upon (if ever) by industry, service classes will most likely be used as an efficient shorthand for sets of parameters.

QoS parameter set type

The QoS parameter set type indicates whether the current set of parameters is the provisioned, admitted, or active set, as shown in Figure 7–16 and Figure 7–17.

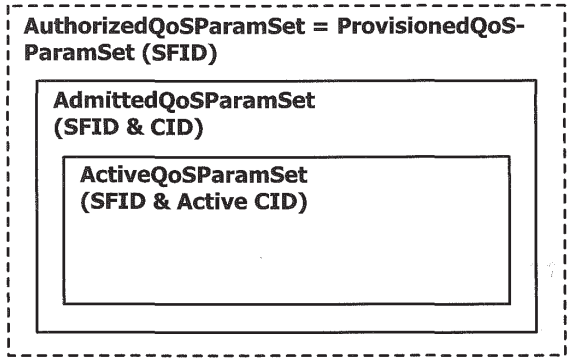


Figure 7–16: Provisioned authorization model

The differences between parameter sets are as follows:

1. The provisioned set is one that is allowed; i.e., when a service-level agreement (SLA) was set up, this is what the users paid for. But the provisioned set is neither reserved nor allocated. Depending on the implementation, the provisioned set may be ignored by the BS's call admission control (CAC) function. For instance, in a mobile system, the users have paid for a service and have a right to certain resources. However, if their phones are turned off, they have no resources allocated, and the BSs ignore them from a CAC point of view because they may never turn on or roam into their field of view.

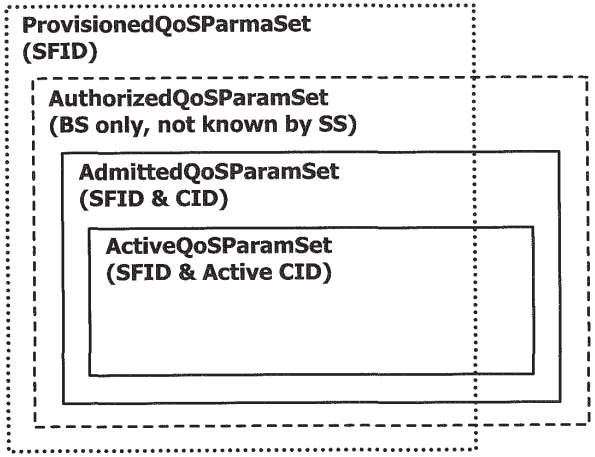


Figure 7–17: Dynamic authorization model

2. The admitted set is one that is allocated but not being used. IEEE Std 802.16 is a bandwidth-on-demand system in which users gets their guaranteed bandwidth only if they ask for it. Allocated but unused bandwidth will cause CAC to keep the resources for this service in reserve so they are available, but will allow other services that are peaking above their guaranteed bandwidth to use the resources on a BE basis. This could be useful in a handover situation, with the necessary resources "admitted" on the new BS prior to the handover.
3. The active set is one that is currently reserved and being used by a service. For bursty services, there is little to distinguish it from the admitted set, since they both consume resources from a CAC point of view, but bandwidth is not used until requested. Technically, the service can't ask for the bandwidth until it transitions to the active state; therefore, the admitted set would rarely be used for bursty services.

These parameters have some interesting potential above and beyond simple resource reservation. For instance, SIP-based VoIP could be modeled as a low-bandwidth (but guaranteed), bursty, nonreal-time polling service (nrtPS) for the SIP signaling, coupled with a fixed-bandwidth UGS [more likely extended real-time polling service (rtPS) in IEEE Std 802.16e] that toggles between provisioned and active, where the active bandwidth may be modified from the provisioned due to choice of codec.

Traffic priority

Traffic priority is a value from 0 to 7 that describes the priority of traffic. Priority alone, in a situation of frequent congestion, does not provide QoS but rather causes starvation. Therefore, this parameter should be used sparingly. It should not override the minimum reserved traffic rate of a connection. It can be useful for determining how to distribute excess bandwidth among connections that peak above their minimum reserved traffic rate. Since the parameter is on a service-by-service basis, rather than a packet-by-packet basis, it cannot be used to differentiate between two packets from the same service for the purpose of discard on congestion.

Maximum sustained traffic rate

The maximum sustained traffic rate is the peak data transmission rate allowed for a service. In the absence of bandwidth demands from other services, a service would expect to be able to transmit at this rate continuously. Since the time base over which this is statistically guaranteed is not specified, it is left to vendor design to determine how smoothly this is guaranteed. It is expected that client devices will police traffic at the ingress and not request more bandwidth than the maximum sustained traffic rate, again statistically determined over some period. BSs can generally assume that the traffic has been policed before it is received across the backhaul from the core network. If the bandwidth demand is greater than the link can bear, services should first receive their minimum reserved traffic rate. Excess bandwidth above the sum of the minimum reserved traffic rate should be distributed proportional to the difference between the minimum reserved traffic rate and the maximum sustained traffic rate. Traffic between the minimum reserved traffic rate and

the maximum sustained traffic rate is not guaranteed to be transmitted and may be discarded. Intelligent discard algorithms that understand the relative importance of individual packets of a service can enhance the QoS perception over general discard methods like weighted random early detection.

Maximum traffic burst

The maximum traffic burst is the largest burst expected at the ingress port for the service. Its main purpose is to ensure adequate buffering in case of temporary congestion or delays between receiving user traffic on the ingress and transmission over the air.

Minimum reserved traffic rate

The minimum reserved traffic rate is statistically guaranteed to a service over time. QoS algorithms should be designed to give services their minimum reserved traffic rate before any services receive bandwidth in excess of their minimum reserved traffic rate, even if they are a higher class of service (CoS) and the traffic is within the bounds of their maximum sustained traffic rate. This parameter is used to prevent starvation of data from connections that are less real-time than others but have guaranteed rates as part of their SLA. Some systems may chose to oversubscribe the minimum reserved traffic rates when performing CAC, assuming that not all services will request bandwidth simultaneously. Typical QoS algorithms use either a credit or a debit scheme with some finite duration memory in determining the bandwidth available to a service under this parameter. For instance, a service that is idle for hours typically cannot expect to wake up and instantly transmit data equivalent to the minimum reserved traffic rate accumulated during those hours.

Minimum tolerable traffic rate

The minimum tolerable traffic rate is related to the Minimum reserved traffic rate. For certain services, it may be convenient to discuss the service paid for—the minimum reserved traffic rate—and the worst service the user will tolerate—the minimum tolerable traffic rate. While CAC should normally use the minimum reserved traffic rate, there may be circumstances in which resource availability changes after a call has been admitted. For instance, the

signal quality may degrade due to some interference or, in a mobile system, due to the client moving to an area of degraded coverage (e.g., farther away or indoors). Alternatively, a mobile call accepted in one cell may move into the neighboring cell that may already be at capacity with respect to minimum reserved traffic rates. The minimum tolerable traffic rate allows calls to be retained at a degraded level even when the call would not normally be accepted by the CAC algorithm.

Vendor-specific QoS parameters

The vendor-specific QoS parameters provide a QoS backdoor allowing two devices to share proprietary information with the expectation that they can do a better job distributing bandwidth without using general algorithms based on generic QoS and traffic parameters. Care must be taken, however, to maintain the ability to interoperate with devices from other manufacturers that do not use the same parameters. In general, an unknown device is handled by dropping back to exclusive use of the parameters defined in the IEEE 802.16 standard.

Service flow scheduling type

The service flow scheduling type defines the CoS.

Converted to ATM terminology, UGS is equivalent to constant bit rate (CBR) and usually has a fixed packet size and periodicity. The BS allocates bandwidth at the desired rate automatically, without any BW request from the client device. Keep in mind that the BS may "clump" the allocation into a portion of the PHY frame (unavoidable for TDD links), increasing the jitter and possibly causing the need for jitter buffering.

Real-time polling service (rtPS), nonreal-time polling service (nrtPS), and best-effort (BE) service are all variable-rate packet-based service classes. The main reason for differentiation is that there are different rules governing how a client device is allowed to request bandwidth for the different CoS. Also, when resources are tight, rtPSs may receive their minimum reserved traffic rate in preference to nrtPSs, which in turn may have priority over BE services.

CoS may impact the allocation of bandwidth above the minimum rate (generally policed to the maximum sustained traffic rate).

Generally, a UGS will have minimum reserved traffic rate set equal to maximum sustained traffic rate. A BE service will have its minimum reserved traffic rate set to zero, but have a nonzero maximum sustained traffic rate. rtPSs and nrtPSs will generally have nonzero minimum reserved traffic rates that are less than, but not equal to, their maximum sustained traffic rates.

Request/transmission policy

The request/transmission policy parameter is slightly misplaced in being considered part of the service flow encodings. While these parameters can be set on a per-service basis, most are really device capabilities and would normally be set to the same values for all services of the same CoS to or from an individual client device. They include the use of broadcast BW request opportunities, piggybacked requests, fragmentation, PHS, packing, and MAC CRC.

Tolerated jitter

The tolerated jitter parameter indicates the jitter that is allowed for a given service. It is quite often assumed that the receiving application has jitter buffering or other means to mitigate the problem. It is not clear from the IEEE 802.16 standard whether the tolerated jitter is meant to be total jitter from all sources or that due to the IEEE 802.16 link of the data's path. It could be used as the basis of jitter buffering but, if other points on the path do jitter buffering as well, unnecessary delay can result. This is a worrisome parameter because certification bodies may impose tests that would require jitter buffers where they may be of little value in the end-to-end system.

Maximum latency

Maximum latency is the maximum delay between receipt of a packet from the network and forwarding it on the air interface. A system is generally designed to minimize latency anyway. This parameter can, however, be used in times of congestion to give packets of similar CoS/QoS relative priority with respect to

each other. Since the other QoS and traffic parameters have more influence on which packets are transmitted and when, it can also be used to age out old packets.

Fixed-length vs variable-length SDU indicator

The fixed-length vs variable-length SDU indicator specifies whether the service carries fixed-length SDUs (like ATM cells) or variable-length SDUs (like IP packets). This parameter is usually meaningful only for QoS in the case of UGS connections, where the periodic grants should be a multiple of the natural fixed packet length of the CBR service, such as an ATM T1. It can, however, be useful in applying packing and fragmentation more intelligently for variable-rate services that use a fixed packet size.

SDU size

If the fixed-length vs variable-length SDU indicator parameter is set to fixed, the SDU size must be used to indicate the size, allowing scheduling algorithms to take advantage of the known fixed size in whatever way is appropriate.

One interesting omission in IEEE Std 802.16 is the equivalent of the cell loss ratio (CLR) of ATM systems. This omission probably stems from the fact that this parameter is used not so much in the scheduling aspects of QoS, but in the link quality aspects. Typical systems would calculate and report a BER or PER that could be used to determine the need to transition to a more robust PHY mode if and when the loss rate is too great for the ToS. Unfortunately, without a parameter, this decision is left to the vendor or to vendor-specific parameters.

A common fallacy is that IEEE Std 802.16 describes only UL QoS. This is an unfortunate artifact of the naming conventions for services and the QoS model description coming from DOCSIS. DOCSIS assumed that since the head end (the cable modem equivalent of a BS) was in control, the cable modem didn't need to know how QoS was handled on the DL. This works in general but, for some issues like jitter, it's best if the receiving end knows what's going on as well. Unlike in DOCSIS, service setup in IEEE Std 802.16

associates QoS and traffic parameters with the DL as well. Therefore, even though IEEE Std 802.16 uses service names like "rtPS," the QoS concepts apply equally to the DL and the UL.

Interactions between QoS, CAC, and adaptive PHY

Determining available bandwidth

BWA systems are normally intended to provide very reliable service, as an extension of fiber optics or some other high-reliability network. When BWA systems are installed, frequency planning is normally performed to minimize interference while maximizing frequency reuse and, therefore, available bandwidth. Cell sizes are chosen based on some availability goal such as 99.99% or 99.999% (often referred to a "four 9's" or "five 9's").

Furthermore, in systems with a subscriber-level adaptive PHY, each SS can be assigned a "planned" PHY mode. The SS is expected to be able to operate at or above the target BER of the system using the planned PHY mode, or a more efficient one, a fraction of time equal to or exceeding the desired availability. For example, if the system availability goal is 99.99% and an SS's planned PHY mode is 16-QAM, then the SS is expected to operate at 16-QAM or 64-QAM the entire year except for 52 min and 33.6 s. The choice of planned PHY mode for a given availability is generally a function of the distance from an SS to the BS and the LOS conditions. Additionally, certain SSs, such as those at the edges of sectors, may be capped so that they cannot use certain more efficient PHY modes due to interference issues.

As a system operates, the physical environment may also change. Most of the time, most SSs will be able to operate in both the UL and DL in not just their planned PHY mode but also in a more efficient one. For LOS systems in the 10-66 GHz frequency range, rain fade is the main variable in link quality. During rain, an SS may not be able to operate above its planned PHY mode. In fact, during heavy rains, an SS may need to drop to a PHY mode more robust than its planned PHY mode to maintain sufficient link quality to communicate with the BS. For lower frequency systems, multipath variation, due for instance to moving objects such as vehicles, is the main culprit.

The CAC algorithm for modern, third-generation BWA systems must deal with the issues created by these dynamic environments. Many fixed BWA services are established as permanent connections and remain active (in operation) for a long time—months and sometimes years. Rain fades, on the other hand, tend to be relatively short in duration. Therefore, the BWA CAC need not consider short-term changes in system capacity. Those are best handled by the queuing, scheduling, and discard algorithm. CAC must, however, consider the planned PHY modes of the SSs, as these determine the amount of bandwidth that can be guaranteed while still meeting the system's availability goals. The BWA CAC algorithm must address three issues:

- What aspects of a service's QoS and traffic parameters are mapped to the bandwidth availability determined by planned PHY modes to guarantee contractual and QoS requirements?
- What should be done with the additional bandwidth available when a terminal can operate at a more efficient PHY mode, which may be most of the time?
- What should be done when a terminal is unable to operate at its planned PHY mode and must use a less efficient one?

Bandwidth on demand setting the basis for CAC

The solutions to the first two of the problems above are based on the concept of bandwidth on demand. While some BWA systems will be expected to carry legacy T1/E1 TDM voice traffic, demand for variable-rate, bursty services such as Internet access is growing. Even for voice, there is a growing interest in using compressed voice, which results in variable bit rate rather than CBR used by traditional TDM connections. These variable-bit-rate services allow more users and more voice calls to be multiplexed into the same bandwidth by taking advantage of not only compression but also statistical multiplexing.

While third-generation BWA systems can interface to a plethora of user services, such as Ethernet, IP, TDM voice, and ATM, it is ATM that currently has the most mature concept of QoS and the richest QoS and traffic parameter terminology. Therefore, while the concepts apply equally well to IP-based systems, they will be explained using ATM terminology and mapped to the more generic WirelessMAN terminology.

The ATM service architecture defines six CoSs, as shown in Table 7–1:

- ***CBR:*** Constant-bit-rate services require a static amount of bandwidth that is available for the lifetime of the connection and impose stringent control on delay and jitter. CBR services are intended to emulate legacy TDM services, such as voice and video, and do not generally support statistical multiplexing.
- ***rtVBR:*** Real-time variable-bit-rate services are bursty, but like CBR also require tight control of delay and jitter. rtVBR services are a good choice for VoIP or streaming video. Because of the bursty nature of rtVBR services, statistical multiplexing may be supported.
- ***nrtVBR, GFR, UBR, and ABR:*** Nonreal-time variable bit rate, guaranteed frame rate, unspecified bit rate, and available bit rate are categorized as nonreal-time services. These services may be very bursty and have a large difference between average and peak usage rates. GFR, nrtVBR, and ABR offer a minimum guaranteed rate while UBR is simply a BE service.

Table 7–1: Mapping of ATM CoS concepts to WirelessMAN

ATM CoS	WirelessMAN CoS
CBR	UGS (unsolicited grant service)
rtVBR	rtPS (real-time polling service)
nrtVBR, GFR, ABR	nrtPS (nonreal-time polling service)
UBR	BE (best-effort service)

These service categories relate traffic descriptions and QoS parameters to network resources. Queuing, scheduling, and CAC are structured differently for each service category. The purpose of CAC is to ensure that the system has the necessary resources (e.g., bandwidth and buffers) to support a new connection without impacting any preexisting connections. CAC uses the following QoS parameters and traffic descriptors to calculate resource availability.

- ***QoS parameters***
  - ***End-to-End Transit Delay:*** This parameter is the amount of delay the service can tolerate. Information delayed longer than the specified delay is considered to be less useful to the application and thus may be discarded by the network. In WirelessMAN, this is the maximum latency parameter.
  - ***Peak-to-peak delay variation (or jitter):*** This parameter is used to size buffers at various multiplexing points within the system. It also impacts the overall delay experienced by the application. In WirelessMAN, this is the tolerated jitter parameter.
  - ***CLR:*** The cell loss ratio parameter is used in sizing the buffers allocated to the connection and in the level of statistical multiplexing gain that can be supported by the BWA system.

    It should be noted that each BWA system has inherent end-to-end transit delay and peak-to-peak delay variation characteristics that are directly related to system design choices.
- ***Traffic description parameters***
  - ***PCR:*** Peak cell rate (or peak bit rate for non-ATM systems) is the maximum data rate allowed for the service. In WirelessMAN, this is the maximum sustained traffic rate parameter.
  - ***SCR***: Sustained cell rate is the average data rate expected to be available to the service. The minimum cell rate of the GFR service would map to this as well. In WirelessMAN, this is the minimum reserved traffic rate parameter.
  - ***MBS:*** Maximum burst size is used in buffer allocation and statistical multiplexing calculation. In WirelessMAN, this is the maximum traffic burst parameter.

The focus of this section is the impact of adaptive PHYs on CAC. Therefore, we will ignore for clarity and brevity the impact of CLR and MBS, since they affect mostly buffer allocation calculations. Likewise, as long as a system is not oversubscribed, the delay requirements of a service are impacted much more by the bandwidth allocation scheme and the implementation of fairness algorithms.

So what does it mean for a system with a subscriber-level adaptive PHY to not be oversubscribed? The answer to this question lies in the definition of what is guaranteed to a service. For a CAC algorithm to work, certain rules should be obeyed. In particular, policing should be performed at the ingress point to the network. Services should not, in general, be allowed to transmit more data than their maximum sustained traffic rate. Granted, this must be done in a constructive way, since the service should not be penalized for the fact that the physical line rate of its input port is greater than its maximum sustained traffic rate. For BE services, it is not uncommon for the maximum sustained traffic rate to equal the line rate of the port.

SLAs typically stipulate the grade of service the end customer is getting from the service provider. The SLA's attribute that is of interest to CAC is the minimum reserved traffic rate. The customer is typically guaranteed that this rate will be available. This guarantee is, however, a statistical guarantee, meaning that, over time, the customer will get the minimum reserved traffic rate, but that there may be short periods in which that rate is not available. In general, the following holds:

- For UGSs, minimum reserved traffic rate = maximum sustained traffic rate.
- For rtPSs, minimum reserved traffic rate < maximum sustained traffic rate.
- For nrtPSs, minimum reserved traffic rate << maximum sustained traffic rate.
- For BE services, minimum reserved traffic rate = 0.

Bandwidth demands in excess of the minimum reserved traffic rate can be viewed as not guaranteed and are candidates for statistical multiplexing of services. For instance, as any cable modem user knows, BE customers can be continually added until they complain.

CAC can determine whether the minimum reserved traffic rate of a service can be guaranteed, by comparing that rate against the bandwidth available in the airlink after subtracting the minimum reserved traffic rates of the other services already admitted. This is easier said than done. First, there is the issue of jitter. If a service is at all bursty, even if it uses exactly its minimum reserved traffic rate, it will need less at one instance and more at another.

Therefore, some additional bandwidth may be necessary to compensate for jitter. Alternatively, an operator's policy may be to oversubscribe somewhat the minimum reserved traffic rate. Therefore, CAC should base a service's guaranteed bandwidth on a function of minimum reserved traffic rate and jitter. The actual function used can be tailored to take into account oversubscription policy and may also account for the responsiveness of the bandwidth allocation algorithm in the BS, since this can affect jitter.

Second, there is the issue of the adaptive PHY. The same minimum reserved traffic rate for two different services, each to a different SS, may use substantially different resources. For instance, an SS operating at QPSK will use three times as much physical airlink resources for the same amount of data as a terminal operating at 64-QAM and the same FEC. Therefore, the CAC algorithm must be aware of the planned PHY mode of each SS and normalize the bandwidth requirements of services to common units to determine actual PHY resources consumed. The planned PHY mode is used because the CAC algorithm is concerned with guarantees, and it guarantees that contractual obligations can be met, but only within the bounds of system availability. The planned PHY modes dictate the bandwidth guaranteed available a sufficient fraction of the time to meet the operator's availability goals.

Since SSs quite often have environmental conditions allowing them to operate above their planned modulation (just not often enough to meet availability goals), there is quite often more airlink bandwidth available than CAC uses in admitting services. This bandwidth can be used for BE traffic. It can serve this purpose in two ways. For rtPS and nrtPS connections, this excess bandwidth is normally available to service demands in excess of the minimum reserved traffic rate, up to the maximum sustained traffic rate. Similarly, the excess bandwidth is available for BE services. The high probability of excess bandwidth in the system allows for substantial oversubscription of maximum sustained traffic rate. This can be especially useful for selling services that can peak at their port rate (for example, 100BaseT Ethernet) if bandwidth is available.

Adaptive CAC philosophies for adverse conditions

Once a function is determined for allowing services into the system based on the planned PHY modes of the terminals, there remains a need for a plan of action for when the system is faced with adverse conditions, e.g., when it rains hard enough for terminals to drop below their planned PHY modes to more robust but less efficient PHY modes.

First, it must be determined whether a terminal dropping to a more robust PHY mode actually causes oversubscription. The airlink on which the terminal operates may not be fully subscribed (in the sense defined in the previous section) with guaranteed traffic. Also, if the terminal in question has only BE services, there is no change in the ability to accommodate the guaranteed traffic within the system. Finally, especially in uncorrelated fades, some other terminals may be operating at more efficient PHY modes, offsetting any bandwidth availability problem.

Second, if it is determined that the actual PHY mode dropback causes an oversubscription problem, some action must be taken to retain some amount of services at adequate QoS. This allows the system to maximize the benefit of one of the major features of adaptive PHYs—the ability to still retain the link, albeit with reduced capacity, even in severe weather conditions.

Care must be taken in maintaining QoS and system stability on the temporarily oversubscribed airlink. Initially, it might be tempting to just let the system run with reduced capacity, increased delay, and reduced QoS because this condition may exist only a few minutes a year. In systems with no UGS or rtPS, this extremely simple approach may be very reasonable. Customers can be taught to accept slower response during extreme weather. But UGS, and to some extent rtPS, is very intolerant to change. If the system has a high enough ratio of UGSs, the T1/E1 TDM systems will start alarming due to insufficient bandwidth. A minor oversubscription situation could ultimately bring down all the phone lines in the sector. Obviously, this is not good.

Another simple approach is to temporarily suspend the services with guaranteed bandwidth for the terminal that is below its planned PHY mode (keeping in mind that, for VBR connections, suspending may be merely

reducing the SCR). After all, the reason it is in this predicament is that it's raining hard enough to be in the 52 min a year (four 9's) when it's not guaranteed to be available anyway. This works and meets contractual obligations; in fact, the system may not need to suspend all of the services at that terminal, but just enough to resolve the oversubscription situation. That effectively gives the remaining services an availability slightly higher that the system goal.

This concept can be taken further and can lead to system availability greater than the goal, or it can be traded off for additional capacity by slightly increasing the size of the higher PHY mode regions within a sector. This is especially true in light of the fact that, while affecting only the services on the SS at lower modulation meets contractual guarantees, it's not really fair.

There are two main points that make this simple solution unfair. First, consider that the oversubscription situation may have come into being not because of this single SS, but because a sequence of SSs dropped below the planned modulation. This one just happens to push the system below the oversubscription threshold. Why should its customers be the ones to suffer? The second reason the solution is unfair is best explained pictorially.

In Figure 7–18, two SSs within the same sector have the same planned PHY mode. SS 1 is not quite close enough to the BS to have a more efficient PHY mode and still meet availability goals. SS 2 is in the opposite situation; it is barely close enough to the BS to meet availability with its planned PHY mode. In this situation, SS 2 will normally need to drop to a more robust PHY mode long before SS 1. In fact, SS 1 may never need to drop to a more robust PHY mode. With the previous simple solution to the temporary oversubscription problem, SS 1will be unfairly penalized for its relative geographical proximity to the BS.

There are numerous methods to make the situation fairer. In applying them, the CAC algorithm can rely on the fact that the availability of services has a minimum, as expressed by the availability goal, but that many terminals actually enjoy much greater availability. So the algorithm can "spread the pain," keeping the average availability the same while decreasing the worst-case availability. One method is to randomly, or in a round-robin fashion, suspend services with guaranteed bandwidth until the oversubscription

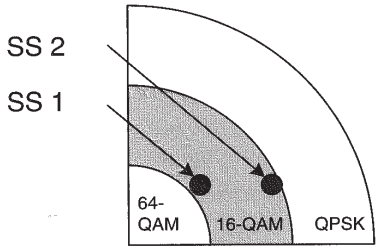


Figure 7–18: Two terminals with same planned PHY mode

problem is alleviated. This is more fair than the previous approach, and it exceeds the availability goals of the system. However, it has the drawback that some of the suspended services may have been critical.

The problem of critical services can be handled by assigning services that have guaranteed bandwidth a precedence level relative to each other. Those with lower precedence are suspended first until the oversubscription problem is alleviated. Within a precedence level, the random or round-robin suspension algorithms may be used. The term *precedence* is used intentionally and should not be confused with CoS or QoS. It may not be acceptable, for instance, to suspend nrtPS first, followed by rtPS, with UGSs getting the axe last. Take, for instance, the business with both its phone lines and its point-of-sale (PoS) transactions going over the BWA system. While maintaining some phone lines during severe weather may be necessary, it is probably more acceptable to have reduced phone capability while retaining PoS capability than to retain all phone lines at the expense of PoS.

Multicast connections

The IEEE 802.16 standard supports multicast connections that allow a service on the DL to be transmitted to multiple SSs on the same airlink without duplicating the transmission to each subscriber. The transmission would be performed at the lowest common denominator of the PHY modes currently used on the DL for the set of subscribers receiving the service.

Multicast introduces an additional complication regarding security. IEEE Std 802.16 typically uses a pair of encryption keys per client so that one client's data cannot be read by another client on the same airlink. For multicast data, e.g., a TV program being broadcast to a number of subscribers simultaneously, this policy of one set of keys per client would effectively require the service to be transmitted individually to each client so it could be individually encrypted for each client. The IEEE 802.16 standard, however, allows per-service encryption keys shared by all recipients of the service, allowing the service to be broadcast once and received by many, but still by only those who subscribe. Clients that support this must be able to have multiple SAs. The target security association identifier (SAID) parameter associated with the service is used to determine which set of keys and which encryption method to use. The SSs' control and any unicast services must use the SA unique to the client. Any multicast or broadcast services are mapped to SAs unique to the service.

Multicast connections also have a unique interaction with ARQ. For multicast service, per-service MAC ARQ is not allowed. Further, since PHY HARQ is on a per-client basis, SSs receiving a multicast connection must be configured to not enable HARQ.

BW request/grant

The BW request/grant mechanism for the IEEE 802.16 standard was chosen to be efficient, low-latency, and flexible and to dovetail with QoS. Requests are made on a per-connection basis to ensure they can be properly used in fairness algorithms in the BS's UL scheduler. But grants are made to the SS, not to the connection. This increases efficiency; e.g., the MAPs are smaller due to an entry for the terminal, rather than each of the terminal's connections. Latency can also be reduced. The terminal can use the bandwidth for the best current need rather than for a reason dictated by the BS. For instance, if an SS asked for bandwidth for a BE connection, but an urgent need to go to a more robust DL PHY mode arises, the SS can use the bandwidth to send a DBPC-REQ (DL burst profile change request) message. This very important message is sent with lower latency than if the SS needed to wait for bandwidth explicitly granted for that purpose.

Also, there are no explicit acknowledgments (positive or negative) of requests. Either the SS gets a grant, or it does not. This saves the bandwidth that would have been used for the acknowledgments. Latency is reduced because the SS does not need to wait for an acknowledgment, just for a grant.

Giving grants to SSs rather than connections can result in inconsistent behavior. Even if the SS does not intentionally use the bandwidth differently from its requests, an SS with multiple services does not know exactly how much of a grant the BS intended for one service vs the other. Additionally, the lack of acknowledgments can cause the BS's perception of the SS's queue status to deviate from the actual situation.

The reality at the SS and the perception at the BS can get out of sync for a number of reasons:

- The BS does not hear a BW request (unrecoverable PHY error).
- The SS does not hear the allocation in the MAP (unrecoverable PHY error).
- The BS scheduler decides it does not have bandwidth right now for the particular service for which the SS requested bandwidth.
- The SS used bandwidth for a purpose different from that originally requested.

To rectify any discrepancy between the SS's reality and the BS's perception, the BW request/grant scheme is designed to be self-correcting. After a period, potentially based on the QoS of the service in question, if the SS still needs bandwidth for a service, it simply asks again. To avoid the BS's perception becoming further askew from reality by duplicate requests, the SS issues an *aggregate request*. An aggregate request tells the BS that this is the current state of the SS's queue for that service, allowing the BS to reset its perception of that service's needs.

Aggregate requests are necessary to guarantee the self-correcting nature of the request/grant scheme. However, there is a chance that a repeated aggregate request crosses the grant for that same bandwidth in the same frame. While this is very unlikely to happen with any regularity, it can, if the request is large, cause wasted allocations to the SS that would have been better served going to a different SS. Therefore, while aggregate requests are both

necessary and sufficient for the request/grant mechanism to operate,[4] there is a potential waste that can be easily avoided by adding the concept of *incremental requests*. With an incremental request, the SS is simply saying "I need more bandwidth" for a particular service. The BS would add this BW request to its current perception of the bandwidth needs for that service rather than using the BW request to reset its perception of the services needs. In general, the airlink should be reliable. Therefore, most BW requests typically would be incremental, with only periodic aggregate requests to ensure the BS does not deviate too far from reality.

If a BS does happen to allocate bandwidth to an SS without a need, the SS can be polite and send an aggregate request for zero bandwidth. This deterministically informs the BS that the SS does not need bandwidth. Alternatively, the SS can just send fill bytes. The BS can key into the presence of fill bytes to modify its perception of the SS's bandwidth needs. Obviously, keying in on fill bytes can be dangerous if the SS implementation can generate fill bytes when the SS has data queued. But, then again, the request/grant scheme is self-correcting, and this situation will correct itself as well.

A few other BW request options exist as well. For UGS connections, which would typically be used for such services as legacy TDM voice, there is no need for BW requests since both parties know the recurring bandwidth needs of the service. However, clock skew can be a problem. To address this, if the SS sees UGS data starting to back up, it can set the SI bit in the GMSH, thereby requesting the BS to slightly increase the rate at which it automatically allocates bandwidth to the SS. This increase would continue until the SS stops setting the SI bit.

If an SS needs more bandwidth for a connection that is currently granted some over the airlink, the SS can piggyback an embedded incremental BW request in a GMSH of an MPDU for that connection.

Similarly, if the SS is receiving bandwidth that is sufficient only for its UGS connections, the SS can request to be polled, indicating it has a bandwidth

[4] ETSI BRAN HiperACCESS uses the same request/grant scheme, but with only aggregate requests.

need on another connection, by setting the PM bit in the GMSH of an MPDU of a UGS connection. When the BS sees the PM bit set, it knows the SS needs to make a BW request and may choose to short-circuit its polling cycle for that SS and poll it immediately. In fact, if an SS has bandwidth needs for its UGSs sufficient large that it receives an unsolicited allocation every frame, the BS can refrain from polling the SS until it sets the PM bit, saving UL bandwidth.

Scheduling

Scheduling algorithms were intentionally left outside the scope of the IEEE 802.16 standard. The set of service offerings combined with scheduling algorithms is a major area for vendor differentiation. However, some concepts are common to all implementations, and some ideas were intended even while not explicitly made a part of the standard.

An IEEE 802.16 system has three main scheduling points. The most obvious one, and the one that most often comes to mind first, is the scheduling of the UL by the BS. The BS receives requests from SSs, processes those requests, and then creates and distributes UL MAP messages. This process must consider not only the outstanding BW requests but also polling, PHY maintenance needs, and the current UL PHY mode for each SS. The current UL PHY modes impact the amount of bandwidth, from a data rate perspective, available in the UL.

But, as previously mentioned, the UL MAP messages allocate bandwidth to SSs, not connections. This leads to the second scheduling point. Once the SS receives an allocation, it must decide which of its data to transmit during that allocation. To do this, it runs a scheduling algorithm analogous to that in the BS, but limited in scope to its own services, MAC management messages, and its individual allocation at its current UL PHY mode.

The third major scheduling point is the creation by the BS of the DL subframe. Like the BS scheduling of the UL, this task must take into consideration not only the DL data awaiting transmission but also the current DL PHY mode of each SS. DL MAC management messages must be included in this scheduling. The end product of this scheduling point is a DL subframe, complete with a DL MAP message indicating PHY mode transitions within

the DL subframe, and the UL MAP message resulting from the BS UL scheduling.

The intent of IEEE Std 802.16 was that the DL MAP and UL MAP messages be transmitted using the most robust modulation currently in use on the airlink. This aids new SSs entering the network, and it guarantees that the MAP will be received with an low likelihood of error. Since the DL MAP is of variable length, it needs to be relevant to the current PHY frame; i.e., the frame containing it. If it instead mapped the PHY mode transitions of the next frame, it would need to know the length of the next DL MAP so it could map the first PHY transition. This is a nondeterministic problem. It requires an FEC structure for transmitting the MAPs that guarantees that the first DL PHY mode transition is contained in an FEC block that is not the last FEC block of the initial PHY mode. This guarantees at least one FEC block buffer for processing the first DL PHY transition. For most practical implementations, this will require the DL MAP to be processed in hardware or firmware rather than software.

For TDD systems, the UL MAP message can be for either the current frame or the next frame, as shown in Figure 4–7 and Figure 4–8 (in Chapter 4). The BS must choose one of these at the start and not change its choice. When the UL MAP is for the current frame, the latency of the request/grant cycle is reduced. This comes at an expense, however. The UL MAP cannot be for any earlier than the round-trip delay of the first SS scheduled in the UL plus the time required to be allotted for processing the MAP. This bounds the minimum DL subframe duration, reducing flexibility compared to when the UL MAP is for the following frame.

For FDD systems, the UL MAP can map allocations starting as early as the start of the current frame plus the time allotted for processing the MAP plus the round-trip delay. Alternatively, it can map allocations starting with the following frame. Anything within these two extremes is possible, as shown in Figure 4–9 and Figure 4–10 (in Chapter 4).

When an SS has UGSs, it does not need to request bandwidth for those services. In fact, it should not make BW requests for those services. The BS already knows the bandwidth need of the UGSs and allocates the bandwidth unsolicited, hence the name *unsolicited grant* services. Based on the frame

duration and the maximum sustained traffic rate parameter, which should be the same as the minimum reserved traffic rate parameter for UGSs, the BS can determine how much periodic bandwidth the SS will need each frame. In granting the allocations, the BS must also take into account the ToS being offered and other parameters of the service.

For example, assume a WirelessMAN-SC backhaul for a cell site of a global system for mobile communications (GSM) network in the United States. The IEEE 802.16 system would most likely be backhauling TDM voice circuits and using a 1 ms PHY frame. A convenient way to backhaul such cell sites is with an ATM adaptation layer virtual circuit. There are various modes for such a circuit, but a popular mode generates 4192 ATM cells per second. Therefore, the circuit needs bandwidth equivalent to 4.192 ATM cells each PHY frame. Since the ATM CS does not allow ATM cells to be fragmented, the BS must, over the course of each second, allocate the SS at least 4 ATM cells in 808 frames and at least 5 ATM cells in the other 192 frames. How the BS distributes these allocations is implementation-dependent and outside the scope of the IEEE 802.16 standard, but it should be obvious that doing a better or worse job can have a big impact on jitter and on buffer requirements at the SS. In general, the BS should try to distribute the allocation of the fifth cell evenly across time.

Even with very good bandwidth allocation algorithms at the BS, extra care must be taken with UGS connections. For instance, the bandwidth allocation for a UGS connection will likely be grouped into a single burst for each UL subframe rather than distributed across the frame. There are a number of reasons for this:

- Reduced complexity of SSs that need to transmit only once per frame.
- Reduced overhead of a single burst rather than multiple smaller bursts (e.g., packing, burst preambles).
- The fact that, unless all you're doing is nailing up T1s or E1s, the scheduling algorithm will never be able to perfectly distribute the allocations within a frame if there is more than one SS in the system.
- The fact that, in TDD systems, the allocations must be compressed into a partial frame anyway.

The "clumping" that results is really the equivalent of adding one PHY frame duration of jitter into the connection. In our example, this is 1 ms. Jitter buffers at both ends of the airlink should be sufficient to remove at least one PHY frame duration of jitter. Be careful, however, because the way to eliminate jitter is to turn it into delay.

Since UGS connections are the preferred choice for transferring constant data rate services, it is important to provide mechanisms for each side to maintain synchronization. Potential problems may arise. The clock used by the BS for PHY frame generation may drift with respect to the clocks used by the network to clock the UGS data. This can lead to data in the UGS queues at the SS building up at a faster rate than the allocations to the SS can drain the queues. When this situation happens, the SS notices that its UGS queues are above an implementation-dependent threshold. The SS then sets the SI bit in a GMSH, indicating that, due to a clock mismatch, the BS should grant it slightly more bandwidth than agreed. Additionally, many types of UGSs need their clock at the SS to match the network clock used for the service on the backhaul. The SS has the airlink clock available, but this may not be locked to the network clock. Whenever the BS is providing UGSs, it also provides the CLK-CMP ([SS network] clock compare) message, which allows the SSs to derive the network clock, or multiple clocks, from the airlink timing. The method for deriving the network clock is described later in this chapter.

rtPS, nrtPS, and BE service all differ from UGSs in that they do not expect continuous, constant bandwidth allocations. They typically carry bursty traffic such as web browsing or VoIP. From the end user's point of view, they may be the aggregation of multiple bursty services. For instance, a single IEEE 802.16 rtPS may allow some number of simultaneous VoIP calls. For rtPS and nrtPS, the minimum reserved traffic rate should be statistically guaranteed. The parameters of such a guarantee are outside the scope of the IEEE 802.16 standard; but in general, if a service is requesting its minimum guarantee, it should receive it. On the other hand, if a service does not use its guaranteed minimum for some period of time, it should not build up an infinite credit. Care must be taken to strike a balance between guaranteeing a minimum rate on average for bursty traffic and building up too much credit, which could starve other services.

The main difference between rtPS and nrtPS is the expected response time. This is especially true once a service becomes active. For instance, a VoIP service may tolerate greater delay during initial setup of a call than during the life of the call. BE services, however, differ from rtPS and nrtPS in that they have no minimum reserved traffic rate. Depending on the CAC, fairness, and scheduling algorithms used, the guaranteed part of other CoSs could effectively starve BE services. If this condition is more than an infrequent transient phenomenon, the system is oversubscribed, and more channels should be made available if possible.

Along with satisfying minimum guarantees and distributing bandwidth in excess of guarantees, another aspect of scheduling is policing. In some networks, services may be sold under agreement guaranteeing some (or no) minimum rate, bursting to some peak rate. Network operators may not want to allow users to become accustomed to having access to all the system's bandwidth if no one else is using it. UGSs typically have a fixed data rate, but rtPS, nrtPS, and BE service can burst with a peak rate equivalent to the line rate of their input port to the SS. For instance, a service connected to the SS via a 100BaseT port can burst data upwards of 96 Mbit/s. At ingress points to the network, policing algorithms should use a function of the maximum sustained traffic rate and the maximum traffic burst parameters to allow instantaneous bursts at the line rate of the input port while statistically limiting the traffic over time to at most the maximum sustained traffic rate. Obviously, if it is desired to have no upper limit on the bandwidth a service can use when available, the maximum sustained traffic rate can be set greater than or equal to the line rate of the port or the maximum rate of the airlink.

Unicast polling

The first thing to understand about unicast polling in IEEE Std 802.16 is exactly what it means. The basic intention of unicast polling is to give the SS a contention-free opportunity to tell the BS that it needs bandwidth for one or more connections. Since the IEEE 802.16 bandwidth allocation scheme allows the SS to use any bandwidth allocation as it sees fit, the SS simply sees extra, unsolicited bandwidth. If the bandwidth is available, unicast polling is generally preferred to broadcast or multicast polling. It can be used to minimize delay and jitter as well as to provide more deterministic delay and

jitter than broadcast and multicast polling, both of which have the potential for collision between two BW requests.

In keeping with the main purpose of the unicast poll, the BS has two primary options for how much additional bandwidth to give an SS. Which option a BS uses is a function of many things, including the SS's service mix, the connection activity, and the complexity of the BS's polling algorithms. If an SS has only a single, active connection, the BS may opt to give the SS only the additional 2 bytes necessary to piggyback a BW request in a GMSH. Alternatively, if the SS has no outstanding BW request or if it has multiple connections, the BS could opt to allocate the 6 bytes necessary for a BW request header. These two options represent the minimum necessary unsolicited allocations for unicast polling. The BS is free to grant more unsolicited bandwidth to an SS. This could allow the SS to transmit multiple BW requests for different services. There are additional reasons why the BS may want to vary the unsolicited grant size. These additional reasons will become more clear when discussing the circumstances that may stimulate the BS to grant unsolicited bandwidth.

In keeping with the main purpose of unicast polling, the BS would periodically grant unsolicited bandwidth at a rate it deems sufficient for the QoS requirements of the SS's services. The frequency can depend on many factors. The BS could merely grant unsolicited bandwidth on a periodic basis, such as every *n* frames. Alternatively, the BS could consider the CoS of the SS's services. For instance, an SS with an rtPS may be polled more frequently than one that has only nrtPS or BE service. Another factor that could be considered is activity. A BE service that has been idle for a few hours has much less likelihood of needing bandwidth soon than one that has transmitted data within the last second. The BS is free to take into consideration all these issues and any others the implementor thinks are important.

There are a few secondary purposes of unicast polling. First, for MAC management message exchanges on the basic or primary MAC management connections, the BS can predict the SS's need for response bandwidth. Such a predictive ability can greatly speed up MAC management protocols. Of course, the SS has the freedom to either use the bandwidth for its part of the protocol exchange or request bandwidth for its part of the exchange. The

timing of these unsolicited allocations depends on the MAC management exchanges in which the BS and SS are engaged.

Additionally, since IEEE 802.16 systems have subscriber-level adaptive PHYs, the BS needs to periodically assess the UL quality from each SS individually. Periodic ranging is used for this purpose. Any unicast bandwidth allocation is considered a PHY maintenance opportunity for periodic ranging. Because the SS must transmit at least filler, even when it has no other data or requests to transmit, any unicast allocation gives the BS an opportunity to measure the receive signal quality of the SS's UL transmissions. Additionally, the SS may use any UL allocation to transmit a DBPC-REQ message for maintenance of the DL. Therefore, the BS must periodically send a unicast poll to the SS for the purpose of PHY maintenance. The frequency of this type of unicast poll is PHY-dependent. For example, to ensure proper PHY maintenance with 10 dB/s fading in a WirelessMAN-SC system, the BS should give the SS a PHY maintenance opportunity about once every 80 ms. If the SS's services are idle, this PHY maintenance opportunity takes the form of a unicast poll.

Finally, at the BS's discretion, any unused UL bandwidth in a frame may simply be allocated to various SSs, effectively polling them in a unicast fashion. This can be particularly effective as a semipredictive way of reducing the acknowledgment latency of TCP/IP transfers, resulting in better response time, such as faster File Transfer Protocol (FTP) transfer rates.

Broadcast polling

During times of congestion, it may be better for certain CoSs to reduce the UL overhead associated with unicast polling and poll SSs on a contention basis as groups. This is especially true of any SS that has only BE connections. Additionally, SSs whose services have been idle for some period would be good candidates for being polled as a group. When an SS enters the network, it is automatically part of the broadcast polling group. At the discretion of the BS scheduling algorithm, the BS may chose to provide contention-based BW request opportunities open to the entire set of SSs on the channel. These opportunities appear in the UL MAP as a special DIUC indicating a BW request opportunity, addressed to the Broadcast CID 0xFFFF.

The method of sending a BW request in such an allocation, common to all PHYs, is the transmission of the BW request header. However, for the OFDM and OFDMA PHYs, an additional anonymous BW request method exists, for which the allocation must be separately provided by the BS. Different techniques are used in OFDM and OFDMA to support anonymous BW requests. However, in both cases, in addition to the BW requests being anonymous, the techniques allow the transmission of simultaneous BW requests from multiple SSs in the same UL allocation.

For OFDM, this so-called *focused contention BW request* (as compared to the full BW request, which uses the BW request header) is achieved by sending a randomly selected 4-bit code on four out of all data subcarriers for the duration of two OFDM symbols, while sending no energy on all other subcarriers. A random selection of a group of four subcarriers by multiple SSs allows a low probability of collision within the same BW request opportunity. For OFDMA, this so-called *contention-based CDMA BW request* is achieved by modulating a randomly selected spreading sequence over the ranging subchannel in the traditional CDMA fashion.

Upon reception of these codes or sequences, the BS can provide UL allocations; however, since the originator is unknown, it indicates in the allocation the received code or sequence and the BW request opportunity in which it was received rather than the Basic CID of the requesting SS. One drawback of this method is that SSs may sometimes select the same code or sequence during the same BW request allocation, resulting in a collision within the resulting UL allocation. Another is that it is not possible to indicate the size of the UL allocation needed in the code or sequence. These drawbacks are offset by the efficiency of the requests themselves.

Multicast polling groups

Sometimes it may be desirable to used contention-based opportunities for groups of SSs, but the entire group may be large enough that the possibility of collision is too high. For situations such as this, SSs can be divided up into multicast polling groups. Multicast polling groups should not be confused with multicast connections. Multicast connections, as described in Chapter 6, are DL connections that are received by multiple SSs simultaneously, without

duplicate transmissions on the same airlink channel. Multicast polling groups are groups of SSs that may respond to a BW request addressed to their multicast polling group address. Effectively, it is the same as the broadcast poll, but only selected SSs are members. This allows the BS to limit the number of SSs that may respond to a contention-based poll, which it may do for a variety of reasons.

- To reduce the possibility of collisions.
- To group SSs by like services, allowing certain groups to be polled more frequently than others, thus maintaining some determinism in QoS while still taking advantage of group polling.
- To provide the BS with a smaller set of SSs that it may decide to poll individually upon detection (assuming the BS has this capability) of a collision.

CLOCK COMPARISON

The clock compare function arose from the need to be able to support services that transferred data synchronously. In particular, synchronous T1s and E1s, while gradually being supplanted by asynchronous T1s and E1s, were relatively common when IEEE Std 802.16 was first approved and are still used today for certain voice backhaul applications. Also, the standard's developers noted that, since synchronous data transfer is generally faster, some future applications are likely to require synchronous transport. Such applications may include the streaming of audio to multiple devices. However, it is difficult to ensure that the clocks needed for synchronizing services across a network and the clocks used for generating the wireless transmission are the same or even synchronized with each other.

In TDD systems and in FDD systems that allow H-FDD devices, information is not provided continuously from a transmitter to the receiver, further aggravating the problem of synchronization of services. In IEEE Std 802.16, we add sophisticated PHY burst profile usage and MAC scheduling that can further disrupt the continuity desired by synchronous services.

Usually, but not always, the synchronous data are synchronized to the network clock that is also used to generate the backhaul. Theoretically, the symbol

clock for the wireless link could be generated from the network clock, but this is typically not done for a variety of reasons:

- The airlink typically needs a stratum 1 reference. While the network clock is probably of high quality, there is no guarantee that the network clock is better than stratum 3 because it may have accumulated errors through the network.
- Especially in TDD systems, it is advantageous to synchronize BSs within a network. Depending on the network, the backhauls to and from each BS may not be synchronized among themselves.
- Rather than build a variety of backhaul synchronization means, it is easier for BS vendors to design airlink synchronization around a known, readily available stratum 1 clock source, such as a global positioning system (GPS).

This creates a situation such as shown in Figure 7–19. In this example, the BS is synchronized to the 10 MHz reference typically provided by a GPS receiver collocated with the BS. It may further use the 1 pps (pulse per second) reference typical of a GPS receiver to synchronize PHY frame boundaries, but this is immaterial to the clock compare discussion. The BS generates the symbol clock for the airlink synchronized to the 10 MHz reference. The SS synchronizes its symbol clock to the airlink, effectively synchronizing to the 10 MHz reference from GPS.

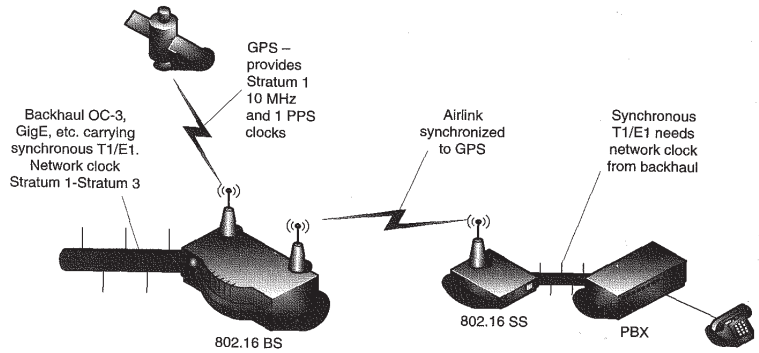


Figure 7–19: Clock comparison

However, any network clocks derivable from the backhaul or services on the backhaul are not guaranteed to be synchronized with GPS. They could be synchronized to a stratum 1 rubidium oscillator clock generator somewhere in the network or to some other clock generator of similar or worse quality, or they could be synchronized to GPS, but with accumulated errors. But, regardless of its quality or its relationship to GPS, one of these network clocks is the one to which the T1/E1 recreated by the SS must be synchronized.

The SS must regenerate the network clock for the service from the airlink clock. By itself, there is insufficient information to do this with accuracy sufficient to prevent bit slips on a synchronous T1 or E1. Therefore, the BS helps out. Every 50 ms, the BS notes the difference between the 10 MHz reference used to generate the airlink and an 8.192 MHz clock generated from the desired network clock. In the clock compare algorithm, 10 MHz was chosen due to the predominance of GPS as the choice for BS reference, and 8.192 MHz was chosen since it is a commonly occurring divisor of most network clocks.

Now the SS can quantify the difference between the two clocks at known intervals. The SS then adjusts the network clock it is regenerating so that the difference between the 8.192 MHz reference it creates for the network clock and the 10 MHz reference it generates from the airlink symbol clock tracks the difference between the two source clocks at the BS. This can be done with sufficient accuracy to prevent bit slips on a synchronous T1 or E1. It has applicability to any synchronously clocked services carried over the airlink.

ARQ OPERATION

ARQ parameters and timers

This section describes the set of parameters and timers required for ARQ operation. Table 7–2 provides the basic IEEE 802.16 ARQ parameters and their usage. An ARQ implementation may also define additional parameters. For example, an ARQ block may have a maximum retry limitation in addition to its lifetime. The acknowledgment generation at the receiver may also be controlled by an acknowledgment frequency parameter (specifying how often to generate ARQ feedback) and/or a maximum ARQ feedback size.

Table 7–2: ARQ parameters and timers

Parameter	Description
ARQ_BSN_MODULUS	The maximum number of unique BSN values (= 2^{11})
ARQ_WINDOW_SIZE	The maximum number of consecutive ARQ blocks in the sliding window maintained by the transmitter and receiver. In order for the SR ARQ algorithm to work correctly, this window size cannot be greater than half of the ARQ_BSN_MODULUS. The transmitter and receiver state machines use the following additional variables to manage the ARQ blocks within their respective sliding windows: • ARQ_TX_WINDOW_START: transmitter variable—All ARQ blocks with BSNs up to one less than this value have been acknowledged by the receiver. Therefore, this is the BSN of the first (unacknowledged or not yet transmitted or not yet available for transmission) ARQ block in its sliding window. • ARQ_TX_NEXT_BSN: transmitter variable—The BSN of the next ARQ block to send. Therefore, ARQ_TX_WINDOW_START <= ARQ_TX_NEXT_BSN <= (ARQ_TX_WINDOW_START + ARQ_WINDOW_SIZE) • ARQ_RX_WINDOW_START: receiver variable—All ARQ blocks with BSNs one less than this value have been successfully received. • ARQ_RX_HIGHEST_BSN: receiver variable—The BSN of the highest ARQ block successfully received, plus one. This variable is related to the window size as follows: ARQ_RX_WINDOW_START <= ARQ_RX_HIGHEST_BSN <= (ARQ_RX_WINDOW_START + ARQ_WINDOW_SIZE)
ARQ_BLOCK_SIZE	The ARQ block size negotiated during connection setup.

Table 7–2: ARQ parameters and timers (Continued)

Parameter	Description
ARQ_BLOCK_LIFETIME	The maximum amount of time an ARQ block will be held by the transmitter after the initial transmission of the block. A timer with this value is typically set for every ARQ block after the first transmission. If the block is successfully transmitted before its lifetime expires (i.e., upon reception of an acknowledgment), this timer is cancelled. If the transmitter is unable to successfully transmit the block within its lifetime subject to retransmissions, the ARQ block is discarded.
ARQ_RETRY_TIMEOUT	The minimum time between a transmission (or retransmission) and a subsequent retransmission of an ARQ block. A block is retransmitted if no acknowledgment is received within this timeout and the ARQ block is still eligible for retransmission.
ARQ_SYNC_LOSS_TIMEOUT	The maximum time the transmitter or receiver keeps its respective TX/RX_WINDOW_START variables at the same value (i.e., sliding window remains fixed) before declaring a loss of synchronization between the transmitter and receiver. This timeout is meaningful only if the connection is actively transferring data and the TX/RX_WINDOW_START has not moved. The transmitter and receiver have their own way of determining whether a connection is actively transferring data.
ARQ_RX_PURGE_TIMEOUT	The maximum time interval the receiver waits after successfully receiving an ARQ block that did not advance the ARQ_RX_WINDOW_START. After this timeout, the ARQ_RX_WINDOW_START is advanced by the receiver.
ARQ_DELIVER_IN_ORDER	If enabled, the ARQ receiver must deliver the MSDUs to the upper layers in order. If disabled, MSDUs may be delivered to the upper layer out of order.

ARQ protocol messages

Apart from a stand-alone message to transport the ARQ feedback, as described in Chapter 6, a few additional MAC management messages are necessary to exchange state information between the transmitter and receiver and keep their ARQ states in sync. These messages are transported in a management connection using the TLV format.

ARQ Feedback management message

The ARQ Feedback management message (Type 33) may carry any size of ARQ feedback up to the maximum allowed by the connection. It is used by the receiver when no other piggybacking opportunities are available to transport the ARQ feedback payload.

ARQ Discard management message

The ARQ Discard management message (Type 34) is used by the transmitter to instruct the receiver to skip a certain number of ARQ blocks. This message, of fixed size, carries the CID and BSN, where the BSN indicates the last ARQ block that the transmitter wants to discard. The receiver should skip up to the indicated BSN and update its sliding window variables accordingly. For an ARQ receiver, the reception of an ARQ Discard management message is equivalent to the successful reception of the ARQ blocks up to the specified BSN. Therefore, in response, the receiver generates an ARQ Feedback IE with the appropriate acknowledgment (i.e., cumulative acknowledgment with the BSN from the discard message).

ARQ Reset management message

The ARQ Reset management message may be sent by the transmitter or the receiver to reset the transmitter and receiver state machines. This message, which resets all variables in the transmitter and receiver to zero, is used when all other attempts to synchronize the state machines fail. The reset message carries the CID, a two-bit Type field, and a two-bit Direction field. The Type field indicates whether the reset message is the original reset message from the initiator, the acknowledgment from the responder, or the confirmation

from initiator. A three-way handshake is necessary to complete the reset operation. The Direction field indicates whether the CID in the message refers to an UL or DL CID. For transport connections, the Direction field is reserved. However, for the secondary management connection, the same CID is used in both the UL and DL for a specific SS. Therefore, if ARQ is enabled for the secondary management connection, the Direction field is used to identify the source of the reset message.

BSN comparison

Both the transmitter and receiver ARQ state machines need to take actions based on comparison of BSNs (e.g., comparing the BSN from a message or received MPDU with a locally maintained value). Since a simple numerical comparison is inaccurate, due to the size of the BSN space and sliding window and the rolling over of BSN values, the BSN values need to be normalized prior to comparison. The normalized BSN, BSN', is defined as BSN' = [(BSN – BSN_Base) *mod* ARQ-BSN-MODULUS], where BSN_base is ARQ_TX_WINDOW_START for the transmitter and ARQ_RX_WINDOW_START for the receiver.

ARQ transmitter

The primary functions of the ARQ transmitter are the following:

1. Determining whether one or more ARQ blocks are qualified for transmission and whether any ARQ blocks need to be retransmitted.
2. Processing ARQ feedback, extracting acknowledgment information, and updating the local state variables, including the ARQ_TX_WINDOW_START, if necessary.
3. Generating a discard message when an ARQ block's lifetime is expired without a successful transmission.
4. Generating a reset message whenever the ARQ_SYNC_LOSS_TIMEOUT is expired.
5. Processing reset messages from the receiver.

The first three functions of the transmitter state machine can be best illustrated using the state transition of ARQ blocks. Figure 7–20 shows the ARQ block states at the transmitter.

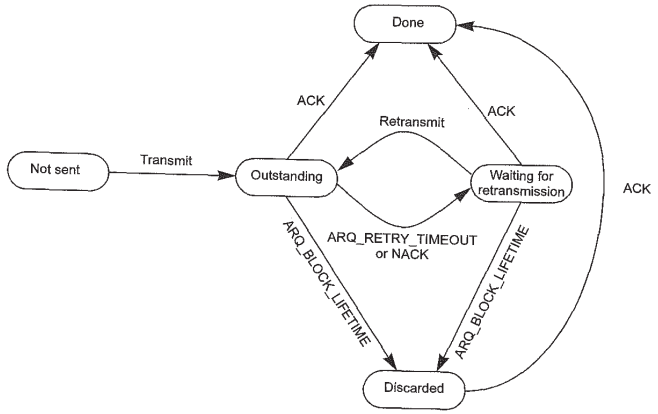


Figure 7–20: ARQ block states at the transmitter

As described in Chapter 6, an MSDU received from the CS is divided into one or more ARQ blocks based on the block size negotiated during the connection setup. An ARQ block starts at the *not-sent* state. An entity outside of the ARQ transmitter, typically the scheduler, is responsible for determining transmit opportunities for a specific connection, based on the QoS requirements and the number of queued ARQ blocks. The exact interface between the scheduler and ARQ is not defined in the IEEE 802.16 standard.

Based on the maximum number of ARQ blocks that can be transmitted for a given opportunity, the ARQ transmitter builds an MPDU of ARQ blocks comprising first transmissions and/or retransmissions. The ARQ transmitter has to follow the packing and fragmentation rules defined for this connection in constructing the MPDU. In order to qualify an ARQ block in not-sent state for transmission, the ARQ transmitter must make sure that the block falls within its window size. Once a block is qualified, it is assigned the ARQ_TX_NEXT_BSN as described in Table 7–2 and the ARQ_TX_NEXT_BSN is incremented by one.

Once an MPDU is constructed and sent for transmission, the ARQ blocks carried in that MPDU become outstanding, i.e., enter the *outstanding* state. Once in the outstanding state, the ARQ block remains in that state until one of the following four events occurs, whichever happens first: (1) Positive acknowledgment (ACK) is received, (2) Negative acknowledgment (NACK) is received, (3) ARQ_RETRY_TIMEOUT is exceeded, 4) ARQ_BLOCK_LIFETIME is exceeded. If a negative acknowledgment is received or an ARQ_RETRY_TIMEOUT occurs, the ARQ block is placed in *waiting-for-retransmission* state. If the ARQ_BLOCK_LIFETIME expires, the ARQ block is moved to the *discarded* state, irrespective of the current state (i.e., outstanding or waiting-for-retransmission).

Similarly, if an acknowledgment is received, irrespective of the current state, the ARQ block is considered successfully delivered. If the BSN of this block is equal to the ARQ_TX_WINDOW_START, then the ARQ_TX_WINDOW_START is advanced by one. Whenever the ARQ transmitter has the opportunity to build a new MPDU, it must give preference for retransmissions, i.e., blocks in the waiting-for-retransmission state must be transmitted before attempting to transmit any blocks in the not-sent state. If there is more than one ARQ block in the waiting-for-retransmission state, the block with the lowest BSN must be transmitted first.

Whenever an ARQ block is marked as discarded (i.e., ARQ_BLOCK_LIFETIME expires), the transmitter sends an ARQ Discard management message to the receiver. This message can be cumulative; therefore, multiple consecutive discard orders can be aggregated into a single discard message that includes the highest BSN to discard. Since it is still possible to receive an acknowledgment while in discarded state and the receiver waits up to ARQ_RX_ PURGE_TIMEOUT before advancing a stagnant ARQ_RX_WINDOW_ START (i.e., deciding to discard on its own), a discard message may be delayed up to (ARQ_RX_ PURGE_TIMEOUT + ARQ_RETRY_ TIMEOUT). In order to quickly recover from possible out-of-sync or other disconnects between the transmitter and receiver, the transmitter repeats discard orders at ARQ_ RETRY_TIMEOUT intervals until an acknowledgment is received for the discarded BSN.

ARQ feedback processing

Whenever an ARQ feedback (ACK or NACK) is received for a specific BSN, the transmitter checks to see whether the BSN falls within its window and whether it is less than ARQ_TX_NEXT_BSN. If either of those conditions is not met, the positive or negative acknowledgment is ignored as invalid.

All the acknowledgment types supported by the IEEE 802.16 standard can be decoded into a cumulative and/or selective acknowledgment. For example, the selective with cumulative acknowledgment and the sequence acknowledgment messages can be decoded into two separate acknowledgments: a cumulative and a selective.

Whenever a valid cumulative acknowledgment is received, the ARQ_TX_WINDOW_ START is set to BSN + 1 of the cumulative acknowledgment. For a selective acknowledgment bitmap, the ARQ_TX_WINDOW_START is advanced for every bit with a positive acknowledgment, whose BSN is equal to ARQ_TX_WINDOW_START.

There may be a delay between the reception of ARQ blocks at the receiver and the actual transmission of the ARQ feedback to the transmitter. This may create a situation in which the transmitter receives a negative acknowledgment for an ARQ block just after it was transmitted and successfully received by the receiver. In order to address this problem, an optional parameter RECEIVER_ARQ_ACK_ PROCESSING_TIME (specified in milliseconds) is defined. This parameter may be provided by the receiver (BS or SS) to the transmitter during connection setup or change. The transmitter may use this information to intelligently interpret ARQ feedback and make retransmission decisions.

ARQ receiver

The primary functions of the ARQ receiver are as follows:

1. Receiving ARQ blocks within its receive window and updating the appropriate state machine variables and timeouts.
2. Generating ARQ feedback (ACKs and NACKs) whenever necessary.

3. Generating a reset message whenever the ARQ_SYNC_LOSS_ TIMEOUT is expired.
4. Processing discard messages and generating acknowledgment (ARQ feedback) in response to indicate to the transmitter the reception of the discard message.
5. Processing reset messages from the transmitter.

Figure 7–21 shows the ARQ block reception procedure. As describer earlier, the CRC-32 is mandatory for ARQ connections. If the CRC check fails, all the ARQ blocks in that MPDU are considered invalid. The CRC check is typically performed outside of the ARQ receiver, and only ARQ blocks that passed the CRC check are handed over to the ARQ receiver. If the MPDU is packed, it needs to be unpacked prior to the examination of ARQ blocks. The reassembly occurs after all the blocks for a specific MSDU are received correctly.

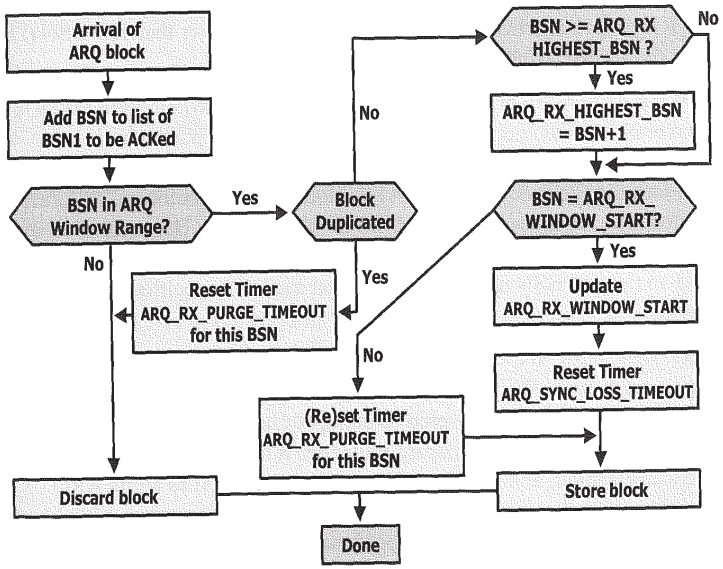


Figure 7–21: ARQ block reception

The first step after the reception of an ARQ block is to add the BSN to the list of BSNs to be acknowledged and to determine whether its BSN is within the ARQ receive window. The ARQ receiver also checks to see whether this block is a duplicate block and discards duplicate blocks after resetting the ARQ_RX_ PURGE_TIMEOUT timer corresponding to that BSN. If this is not a duplicate block, then the BSN is checked against the ARQ_RX_ HIGHEST_BSN. If the BSN is greater than ARQ_RX_HIGHEST_BSN, then ARQ_RX_ HIGHEST_BSN is set to BSN + 1. Then the receiver checks to see whether the BSN advances the ARQ_RX_WINDOW_START. If it is equal to the ARQ_RX_WINDOW_START, then the value is updated and ARQ_SYNC_ LOSS_TIMEOUT timer is reset.

Whenever an ARQ block that did not advance ARQ_RX_WINDOW_START is received, an ARQ_RX_PURGE_TIMEOUT is started for that block. Every time an ARQ_RX_PURGE_TIMEOUT expires, the ARQ_RX_WINDOW_ START is set to the BSN of the next block that has not yet been received. When an ARQ block advances the ARQ_RX_WINDOW_START, all ARQ_RX_PURGE_TIMEOUT timers set for BSNs lower than this BSN (i.e., outside the receive window) are cancelled.

Only if the ARQ receiver has all blocks corresponding to complete MSDUs is the reassembled SDU delivered to upper layers. If only partial blocks are available for an SDU and the ARQ_RX_WINDOW_START has already moved past the required BSNs, all ARQ blocks in the incomplete SDU are discarded. The receiver also discards all blocks up to the BSN specified in an ARQ Discard management message and advances the ARQ_RX_WINDOW_ START to the first BSN not yet received after the BSN specified in the discard message.

If ARQ_DELIVER_IN_ORDER is enabled, the reassembled MSDUs must be delivered in order. This requires that all ARQ blocks of all previous MSDUs have been either delivered to the upper layers or discarded due to a timeout or the reception of a discard message. When ARQ_DELIVER_IN_ ORDER is disabled, an MSDU may be delivered as soon as all its ARQ blocks are received correctly, irrespective of the state of previous ARQ blocks and MSDUs. Most higher layer protocols and applications require in-order delivery, although some may be able to accept out-of-order delivery.

ARQ feedback generation

The frequency of acknowledgment generation is not defined in the IEEE 802.16 standard. It is up to implementations to define a frequency or other innovative algorithms to decide when to generate an acknowledgment. However, all received ARQ blocks must be acknowledged.

All acknowledgments outside the receive window must be cumulative, irrespective of whether those blocks were actually received correctly. As described in earlier subsections, the ARQ_RX_WINDOW_START may be advanced for a variety of reasons: (1) the reception of an ARQ block with BSN equal to ARQ_RX_WINDOW_START, (2) the reception of a discard message, or (3) a purge timeout. In all of these cases, the ARQ receiver generates an ARQ Feedback IE with cumulative acknowledgment at least once to indicate the reception of all ARQ blocks less than ARQ_RX_WINDOW_START. For all other ARQ blocks received that fall within the receiver window, the generated acknowledgments may be cumulative (if applicable), or selective, or a combination of both.

ARQ state machine reset and resynchronization

The ARQ reset mechanism is used to recover from an out-of-sync state, with the transmitter and receiver having lost synchronization due to a sequence of lost messages or an incorrect implementation. The reset operation may be initiated by the transmitter or the receiver.

- ***Transmitter-initiated reset:*** A reset is initiated by the transmitter when the ARQ_TX_WINDOW_START has not moved for ARQ_SYNC_LOSS_TIMEOUT while the ARQ connection is active. The process of transmitter-initiated reset, along with the actions taken by the receiver, is given in Figure 7–22. The timer T22 used in Figure 7–22 is defined in the IEEE 802.16 standard as the "*Wait for ARQ-reset*" timer, with the default value of 0.5 s.

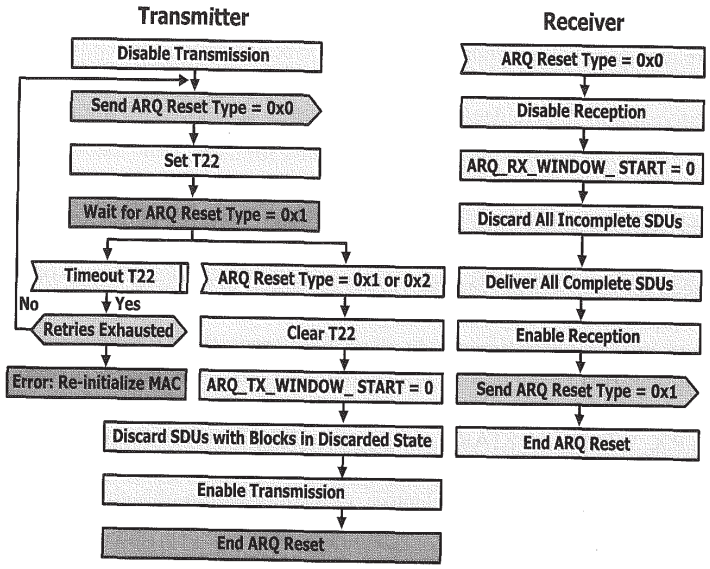


Figure 7–22: Transmitter-initiated reset

- ***Receiver-initiated reset:*** The receiver, like the transmitter, can initiate an ARQ reset. It does so when the ARQ_RX_WINDOW_START has not moved for ARQ_SYNC_LOSS_ TIMEOUT while the ARQ connection is active. Figure 7–23 shows the process of receiver-initiated reset. The timer T22 used here is the same timer used in transmitter-initiated reset.

Both the reset procedures take into account the case where both the transmitter and receiver independently initiate a reset. In the case of an ARQ reset error, the SS should reinitialize its MAC. The behavior of the BS is undefined, and it is left to the implementation to take the appropriate action when an ARQ reset error occurs.

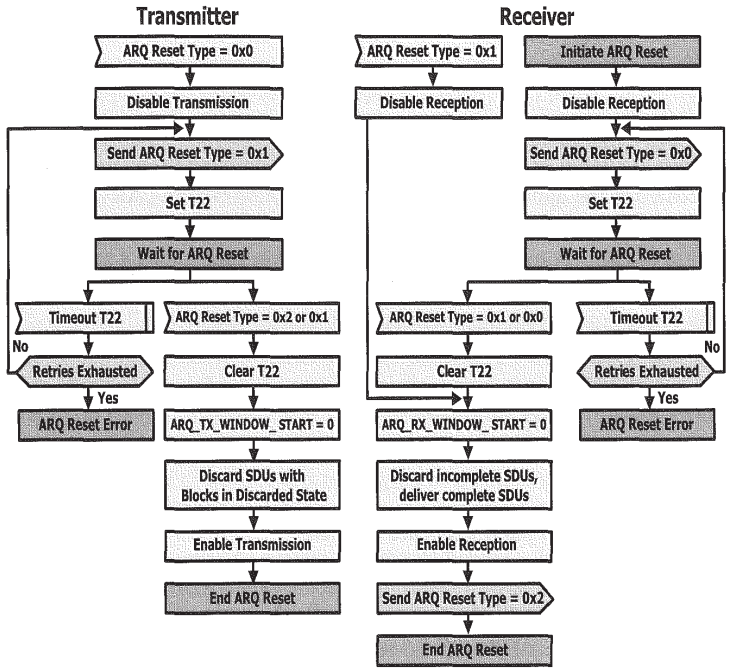


Figure 7–23: Receiver-initiated reset

Interaction with scheduler

The interactions between a scheduler and ARQ are not defined in the IEEE 802.16 standard. It is up to the implantation to define the exact behavior, including the following:

- *How to count retransmissions?* The ARQ protocol requires that the retransmissions be given preference over first transmissions. However, a MAC-level QoS scheduler has to decide whether to fully or partially count the retransmissions or count only the first transmissions toward the connection's specified limit. For a higher layer protocol or application,

only the successfully delivered MSDUs count toward its perception of QoS. Moreover, delay and jitter management also become challenging with retransmissions.

- *How to allocate bandwidth for ARQ feedback?* Although ARQ Feedback IEs may be piggybacked on another connection, such opportunities are not always available or sufficient. In such cases, bandwidth may need to be allocated in the opposite direction exclusively for carrying ARQ feedback. A BS may be able to allocate and transmit ARQ feedback going toward an SS on demand from the ARQ receiver. However, for transporting ARQ Feedback IEs from an SS to the BS, either an SS has to make an explicit BW request or the BS has to proactively and periodically allocate bandwidth for this purpose when ARQ connections are active.

HARQ OPERATION

HARQ parameters

The following HARQ parameters are defined:

- ***SPID:*** Subpacket identifier for the four subpackets of the encoder packet, which has a predefined range of 0b00 to 0x11.
- ***ACID:*** ARQ channel identifier, which specifies the subchannel for the subpacket transmission.
- ***AISN:*** ARQ identifier sequence number, which is toggled between 0 and 1 on successful transmission of an encoder packet.

The following subsections describe the usage of these parameters by HARQ protocol. The HARQ Control IE in the DL MAP or UL MAP is used to signal these parameters for a specific DL or UL HARQ burst.

Subpacket transmission and acknowledgment generation

When HARQ is enabled, the MAC delivers the constructed MPDUs to the HARQ PHY module responsible for concatenation, encoding, and subpacket generation. Since HARQ operates at the FEC level, the FEC encoder is responsible for generating the subpackets. Once generated, the subpacket with SPID 0b00 is always transmitted first. Subsequently, any one of the

subpackets with SPID 0b00, 0b01, 0b10, or 0b11 may be transmitted based on the positive or negative acknowledgment feedback from the receiver. A subpacket may be transmitted more than once, and any subpacket may be skipped, with the exception of the subpacket with SPID 0b00.

Upon receiving a subpacket, the receiver's FEC decoder attempts to decode the original encoder packet. If the original is successfully decoded, a positive acknowledgment is sent to the transmitter; otherwise, a negative acknowledgment is sent. The transmitter stops sending additional subpackets from the same encoder packet if a positive acknowledgment is received. The transmitter may transmit one of the other subpackets if a negative acknowledgment is received, and this process may continue until the receiver successfully decodes the original encoder packet. A 1-bit HARQ AISN, toggled between 0 and 1, is used to indicate a new encoder packet transmission on the same HARQ channel. Therefore, upon successful decoding of an encoder packet, the transmitter toggles this bit to indicate this fact to the receiver of subpackets belonging to the next encoder packet.

As described in Chapter 6, the synchronous HARQ positive and negative acknowledgment signaling makes use of a dedicated subchannel in the UL and a compact DL MAP with acknowledgment bitmap in the DL. However, the BS has full control over both UL and DL and may decide not to send an acknowledgment bitmap at a fixed frame offset. If no acknowledgment bitmap is present in the expected frame, the SS retains the HARQ burst and the corresponding subpackets. The BS may explicitly request a retransmission later, using the HARQ Control IE in the Compact MAP IE.

The maximum number of HARQ retransmissions in the UL and DL is advertised in UCD and DCD messages, respectively. The default maximum number of HARQ retransmissions in both DL and UL is set to four.

Performance and QoS implications

As described in Chapter 6, the stop-and-wait HARQ protocol uses synchronous acknowledgments[5] that are sent after a fixed delay specified in

[5] The BS may optionally decide not to send a synchronous acknowledgment and request a retransmission using an HARQ Control IE.

frame offset in both UL and DL. However, the number of subpacket transmissions required for successful decoding may vary, based on channel conditions. Moreover, the transmission/retransmission of the subpackets is not synchronous. The ability to support multiple HARQ channels per terminal does offer some additional performance, as multiple encoder packets may be pending at any instant. The ACID, carried in the MAPs, uniquely identifies the channels.

While HARQ may offer some performance improvements due to SNR gain and coding gain by incremental redundancy, the lack of control from the MAC and the inherent limitations of stop-and-wait ARQ may affect the type of QoS that can be guaranteed for an HARQ–enabled SS.

Chapter 8 Security

PKM protocol and cryptographic methods

The IEEE 802.16 standard defines security mechanisms affording privacy to subscribers and protecting operators from theft of service. These objectives are achieved by encrypting the data between the BS and SS, by utilizing digital certificate-based authentication of the SS, and by employing an authenticated client-server key management protocol called *privacy key management* (PKM). The PKM protocol allows the BS, as server, to control the distribution of keying material to the SSs. The security protocols of IEEE Std 802.16-2004 [B20] do not offer authentication of the BS. This leaves open the possibility of an SS associating with a malicious BS. However, for a fixed wireless system operating in licensed bands, this should not be a grave concern since the BS equipment may not be readily available; furthermore, operation of a malicious BS can be easily detected by a legitimate operator. Nevertheless, BS authentication has been specified in IEEE Std 802.16e.

SECURITY ASSOCIATIONS (SAs) AND CRYPTOGRAPHIC SUITES

The packet data encryption function in IEEE 802.16 resides in the security sublayer between the MAC CPS and PHY, as shown in Figure 4–1 (in Chapter 4). In the transmitter, the MPDUs are mapped to an SA, which dictates the processing to be applied to a given plaintext MPDU. The receiver uses the CID to determine the correct SA and applies the corresponding processing to the received MPDU.

The SA is a set of security-related information that is shared between the transmitter and receiver, with help from the PKM protocol. Each SA contains information on the cryptographic suite used for that SA. The SA may also contain keys, such as the traffic encryption keys (TEKs), along with their lifetimes and other associated state information. The exact content of the SA depends on the algorithm for encrypting and authenticating the MPDUs and

on the algorithm used for TEK exchange specified in the cryptographic suite. IEEE Std 802.16 defines two main cryptographic suites, as described in more detail below.

An SS entering a network will set up at least one SA, called the *primary SA*, with the BS. The BS may also set up a number of static and dynamic SAs. Static SAs are provisioned at the BS, while dynamic SAs are set up and torn down along with opening and closing service flows. In contrast to the primary SA, the static and dynamic SAs can be shared by the BS and multiple SSs.

Encrypted MPDUs

Only the payload part of an MPDU (which, as defined in the IEEE 802.16 standard, includes the MAC subheaders) is encrypted. The generic MAC header is always sent in cleartext. The cleartext header contains the two EKS bits indicating the LSBs of the TEK sequence number of the key that was used to encrypt the PDU. This allows smooth key transitions in both directions.

Data encryption with DES in CBC mode

One of the two main cryptographic suites defined in the standard specifies data encryption with DES in CBC mode. The MPDU data authenticity is not checked when using this method. The exchange of the 64-bit TEK is performed using 3DES in code book mode. During the TEK exchange, the BS delivers parameter CBC-IV, which acts as the basis for calculating the IV required for encryption processing. The encryption process involves the following steps:

1. The IV is derived by calculating the XOR of the CBC-IV and the frame number in which the MPDU transmission takes place. For UL transmissions, the XOR is taken between the frame number in which the UL MAP allocating the transmission was received and the CBC-IV.
2. The IV is XORed with the first 8 bytes of plaintext MPDU payload, and the result is then DES-encrypted to get the first 8 bytes of ciphered payload.

3. The ciphertext from the previous step is XORed with the next 8 bytes of plaintext payload and DES-encrypted to get the next 8 bytes of ciphered payload.
4. Step 3 above is repeated until less than 64 bits of payload remain unciphered. To cipher the n residual bits, the previously attained ciphertext is DES-encrypted once more. The XOR between the n MSBs of this encrypted result and the n plaintext bits is then computed. The n last bits of ciphered payload are made up of the result of this calculation.

In the receiver, the plaintext is recovered by performing a DES decryption of the ciphertext block and XORing it with the previous ciphertext block (the IV in case of the first 8 bytes). To recover the residual bits, the last full ciphertext block is encrypted once more, and the relevant part of result (the n MSBs) is XORed with the ciphered residual to get the plaintext.

The level of security provided by this suite is not very strong. DES uses keys that are considered by security experts to be too short to be secure. Also, there are some issues with the predictable way in which the cipher chain is initialized. Furthermore, the lack of MPDU authentication can be considered a weakness of the protocol. However, this suite has the benefit of not introducing any data expansion. If the weaker security of this mode is a concern, then the AES-CCM suite described below should be employed instead.

AES-CCM

The advanced encryption standard (AES) is the replacement for DES and considered to be state of the art. The AES is a block cipher with a block length of 128 bits. The keys can be 128 or 256 bits long. IEEE Std 802.16 defines a cryptographic suite that utilizes the AES algorithm in the counter with CBC-MAC (CCM) (see [B40]) mode with 128-bit keys. In addition to data privacy, the suite provides integrity protection for the MPDUs. The encryption method is the same as the one used in IEEE Std 802.11i[6]. The TEKs are exchanged

[6] IEEE Std 802.11 and IEEE Std 802.16 construct data fields differently. Also, IEEE Std 802.11i defines its own protocol for key exchange.

using AES in code book mode. The confidentiality and data integrity protection provided by the AES-CCM is considered very strong.

The AES-CCM processing adds two fields to the MPDU payload: a 4-byte packet number (PN) and an 8-byte message authentication code. The MPDU payload before and after processing is shown in Figure 8–1.

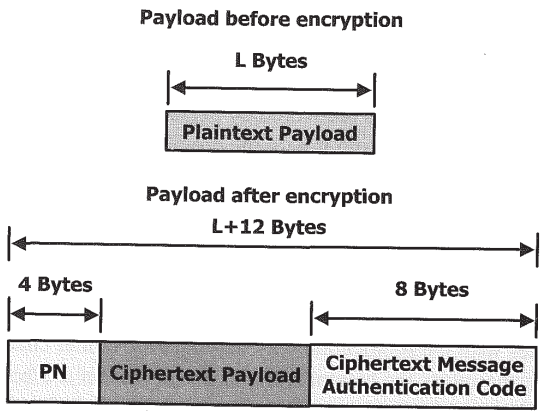


Figure 8–1: Encrypted payload format in AES-CCM mode

The PN space is split in two—the range 0–7FFFFFFF is used for DL and the remainder for UL—because only one MPDU payload should be encrypted with a given key and PN.

The message authentication code is calculated using AES in CBC mode, with counter mode encryption. The steps of the AES encryption process are described below:

1. A set of blocks $B_0,...,B_r$ are formed from the MPDU.
2. A temporary CBC checksum T is calculated over these blocks.
3. Using the nonce (see Figure 8–2) and a counter incremented by one for each block, a set of counter blocks (see Figure 8–3) $S_0,...,S_m$ is formed.
4. The first counter block S_0 is encrypted and stored.

5. Block S_1 is encrypted, and the XOR with the first 128 bits of the MPDU payload is calculated to gain the first 128 bits of encrypted MAC payload. The same is done for the subsequent blocks, if any. If the length MPDU payload is not a multiple of 16 bytes, the last block is shortened.
6. The message authentication code is derived by calculating T XOR S_0 and keeping only the eight most significant bytes.
7. The PN, the encrypted payload, and the message authentication code are concatenated to form the encrypted MPDU payload.

The initial block B_0 used in the CBC calculation is shown in Figure 8–2. By including the generic MAC header (save the redundant HCS) in the nonce, the header also will be afforded integrity protection. The remaining blocks are formed by splitting the plaintext MPDU payload in 16-byte blocks (padding with zeros to the next 16-byte boundary, if necessary).

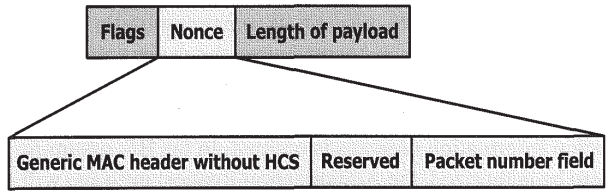


Figure 8–2: Initial CBC block and nonce

The counter blocks are shown in Figure 8–3, where the nonce is the same as above.

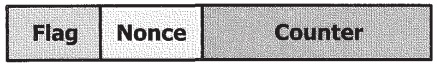


Figure 8–3: Counter blocks

The processing in the receiver is similar to that in the transmitter, but the order of operations is different. The receiver starts by generating the encrypted counter blocks and decrypts the payload. After decryption, it calculates the

CBC checksum and compares this against the message authentication code. If the check fails, the receiver discards the MPDU.

KEY MANAGEMENT

Key management is handled by the authorization and TEK state machines in the SS. The authorization state machine, initialized on SS power up, handles the authorization state and AK management. In the Authorization Reply message, the SS receives the AK and the list of SAs from the BS. For each SA, the SS starts an instance of the TEK state machine that handles the key management of that SA. A TEK state machine is also instantiated upon reception of the SA Add message that creates a dynamic SA.

AK management

The AK is always accompanied by its associated lifetime. Upon reception of an AK, the SS starts a timer to measure its remaining lifetime. At the instant named "AK grace time," which occurs before the expiration of the AK lifetime, the SS initiates a reauthorization procedure to acquire a new AK. The reauthorization is almost identical to the initial authorization, but the Authorization Information message is omitted. In the BS, assuming the SS continues to be authorized, the Authorization Request message will result in a new AK being activated with a lifetime that extends beyond the lifetime of the AK that is about to expire. In this transition period, the BS will have two active AKs for an SS. The transition period and key usage are illustrated in Figure 8–4. The AK is used to derive keys for MAC management message authentication and TEK exchange. As soon as the new AK is activated, the BS starts using the keys derived from the new AK.

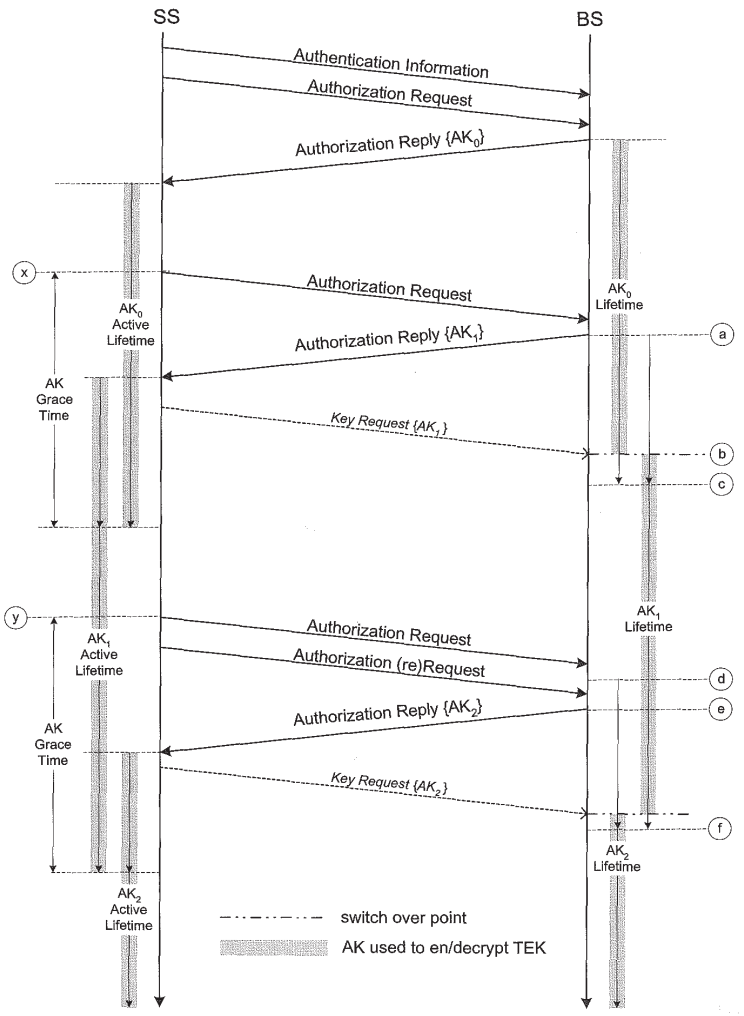


Figure 8–4: AK management in BS and SS

TEK management

As mentioned above, the establishment of a new SA either during initial authorization or after reception of the SA Add message will instantiate a new TEK state machine. The TEK state machine will trigger the SS to send a Key Request message to the BS to obtain the keying material for the cryptographic suite used for the SA. Two TEKs with overlapping lifetimes are maintained to prevent interruptions in data transfer due to key expiration. The scheme is illustrated in Figure 8–5. There are certain rules that the BS and SS have to follow for this scheme to work. These rules are summarized below:

- The SS is capable of decrypting DL transmissions using either TEK.
- The SS always uses the newer TEK for UL transmissions. Once it sees the SS use the newer TEK, the BS uses this information as an implicit acknowledgment that the SS has the new keying information.
- The BS uses the older TEK to encrypt data in the DL and starts using the new key only when the old one expires.
- The BS is capable of decrypting UL transmissions using either TEK.

As previously mentioned, the two LSBs of the sequence number of the TEK are carried in the generic MAC header to aid the receiver in the TEK selection for data received on a given MAC connection using a given SA.

In IEEE Std 802.16, the TEKs are usually exchanged using the same block cipher used for the data encryption; in the case of DES, the cipher is run three times with different keys. The symmetric methods are favored over key exchange using RSA, as they consume less computational resources than the RSA method, which is also specified as an option in the IEEE 802.16 standard. The symmetric methods use the key encryption key (KEK) derived from the AK using the Secure Hash Algorithm 1 (SHA-1).

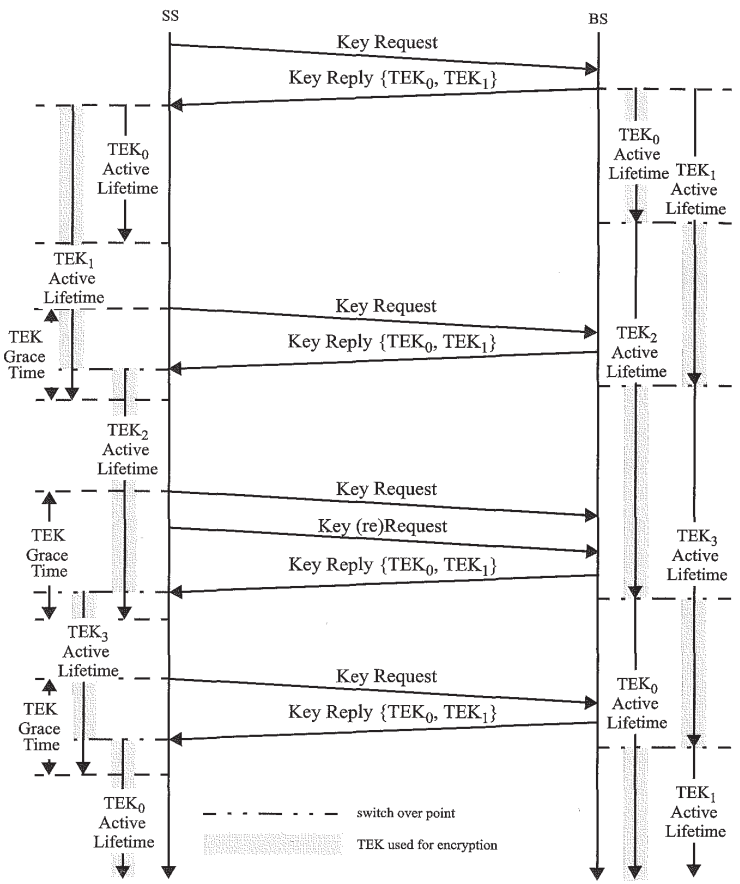


Figure 8–5: TEK management in BS and SS

Chapter 9 Mesh

MAC and PHY extensions for mesh

INTRODUCTION

The concept of mesh radio systems, as an alternative to PMP, has for the last decade been a hot topic for commercial communications, evolving from earlier military work done by the Defense Advanced Research Projects Agency (DARPA) in the United States. Beyond the fixed mesh specification already available in IEEE Std 802.16-2004 [B20], the IEEE 802.16 Working Group proposed in January 2005 the development of an amendment to facilitate mobile multihop relays for fixed and mobile infrastructure. The difference between mesh and infrastructure relay is that a mesh network allows connectivity between any set of end-user and infrastructure devices without manual intervention, where the number and location of end-user devices will fluctuate. An infrastructure relay, on the other hand, is topologically limited to a tree structure with a BS at its root and infrastructure relays as branches.

Mesh is of interest not only to the IEEE 802.16 community but also to the IEEE 802.11 and IEEE 802.15 communities. In particular, the IEEE 802.11s Task Group is currently developing a mesh amendment for WLANs, and the IEEE 802.15 Task Group 5 is currently developing a mesh amendment for WPANs.

The typical reasons for designing or deploying mesh include the following (see [B52]):

- ***Communication within a local community:*** This is where mesh performs best, but it is not a very typical scenario.
- ***Capacity increase:*** When the ingress and egress of traffic is distributed throughout a mesh network, the capacity increase is easy to comprehend as traffic can be sent simultaneously over different links due to the spatially

distributed nature of a mesh network. However, when most of the traffic within the mesh is being aggregated to a very limited number of egress points (such as the core network connection), then increased capacity can be difficult to achieve. Compared to the PMP, the capacity of a mesh network represents a balance of increased overhead to schedule the mesh traffic and forwarding of data over multiple links, both of which reduce capacity, vs the substantially increased capacity of each individual link with the mesh network resulting from shorter links and the ability to route around links severely degraded by obstacles in the transmission path.

- ***Alternative paths:*** Mesh systems that can provide alternative paths add resilience against link failures in the network.
- ***Cost-effective range extension:*** Because the nodes of a mesh are typically at least an order of magnitude less expensive than a BS and comparable in cost to PMP SS devices, mesh can be a cost-effective method of range extension of a traditional PMP network when the user population is too sparse to justify the extra BSs. Keep in mind that, if the node density decreases beyond a certain point, the mesh becomes ineffective. This is shown in Figure 9–1 and Figure 9–2. The typical mesh deployment hence requires so-called "seeding" to ensure an initial minimum coverage of the area.
- ***Cost-effective deployment:*** In fixed deployments, coverage is typically inadequate to allow a PMP SS to be installed in an arbitrary location. In some cases, a directional antenna must be installed and aligned, at substantial cost. For logical mesh networks, coverage is omnidirectional, and user self-installation may be enabled.

These tradeoffs can determine whether PMP, a true mesh, or something like a multihop relay is most appropriate.

Two general categories of mesh systems, *logical* and *physical,* are considered here. Logical mesh uses broad-beam or omnidirectional antennas to form logical links to neighbors. These links are considered logical because the hardware configuration does not change for different links. Physical mesh, by contrast, uses substantially directional antennas to create physical links. Links are considered physical because the hardware configuration (via beam forming or via switching, for example) changes to form different links with different neighbors. Both varieties have challenges and benefits.

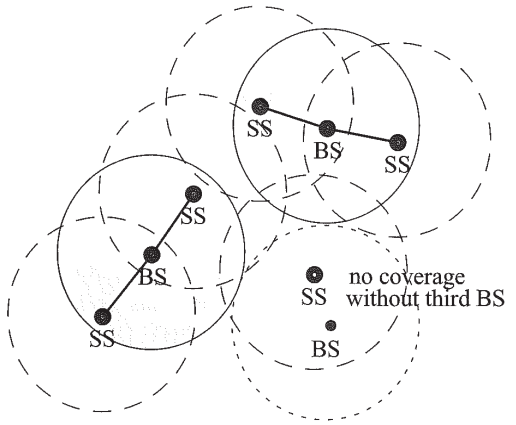


Figure 9–1: Three BSs required for full coverage

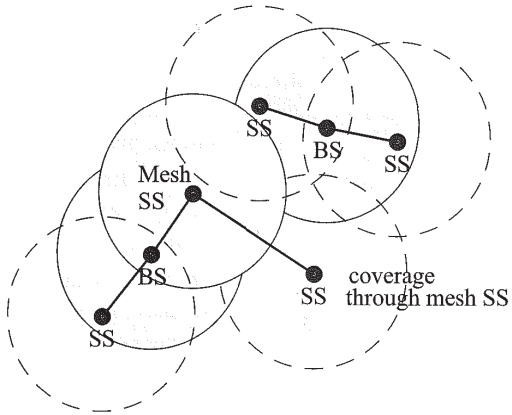


Figure 9–2: Two BSs required with mesh SS

IEEE Std 802.16 includes both concepts to some extent. The logical mesh is completely specified, whereas hooks are in place for the physical mesh, known in the standard as *directed mesh*. Although it was never fully specified in the standard, it is worth spending further discussion here on the topic, especially since some of the concepts may be reused in the proposed mobile multihop relay amendment.

LOGICAL MESH

Since the logical mesh was incorporated into the IEEE 802.16 standard primarily by the proponents of the OFDM PHY, this mode was defined to function only with the WirelessMAN-OFDM PHY. Despite the fact that the mesh MAC uses many of the same messages used by the PMP MAC (there are only five mesh-specific MAC message formats), specifically in the upper MAC, the two architectures are not interoperable.

The logical mesh architecture is based on the concept of neighborhood, where distinction is made among the following:

- ***Node:*** A device participating in the mesh network. Although some devices may have a connection to the core network and some may be assigned special functions such as centralized schedule controller (Mesh BS) or master clock, all devices are equal, and all connectivity is peer to peer at the most fundamental connectivity level of the network.
- ***Immediate neighborhood:*** The collection of nodes with which the station of reference is capable of establishing direct physical connection. These nodes are also referred to as *one-hop neighbors*.
- ***Extended neighborhood:*** The collection of nodes with which the nodes in the immediate neighborhood can establish direct physical connection. Stations in the extended neighborhood that are not immediate neighbors are referred to as *two-hop neighbors*.
- ***Cell:*** The collection of nodes that are associated with the same coordinated (distributed and/or centralized) scheduler. Between cells, only uncoordinated distributed scheduling can take place. A node that has an interface to the core network or controls the centralized scheduler is

termed *Mesh BS*. Typically these nodes include both functionalities. All other nodes are also termed SSs for the sake of simplicity.

- ***Network:*** A collection of cells where nodes may arbitrarily associate with the cell of choice and where synchronization takes place between the cells.

This concept is graphically depicted in Figure 9–3.

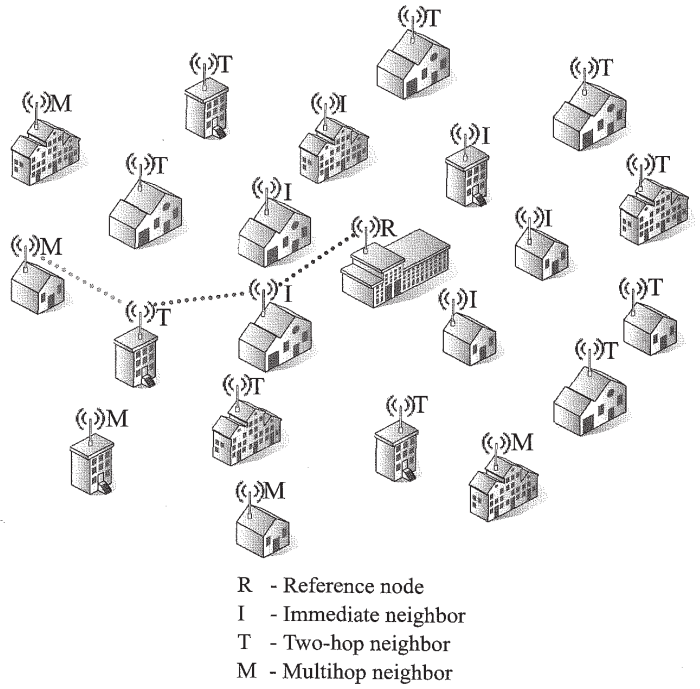


Figure 9–3: Neighborhood definitions

The definition of the mesh architecture consists of essentially four aspects that are different from PMP: PHY frame format, network entry, network configuration/maintenance, and scheduling of data exchanges. The first aspect is dealt with in Chapter 11; the others are explored below.

Mesh connections and addressing

Beyond these aspects, the single most substantial difference between logical mesh and PMP is the usage of the concept of CIDs. In PMP, each 16-bit CID represents a unique, logically numbered connection to an SS with a defined set of QoS parameters. In a logical mesh network, the CID is used to convey a Link ID, unique only in the local neighborhood, and a set of broadcast/unicast service parameters. To uniquely identify the destination node, a 16-bit Node ID, unique within the network, is appended to the MAC header. The mesh system is hence not connection-oriented, but applies priority on a packet-by-packet basis. As a packet moves through the network, the Link ID in the CID will change with each transmission, whereas the service parameters typically do not. The mesh CID and mesh Broadcast CID are shown in Figure 9–4.

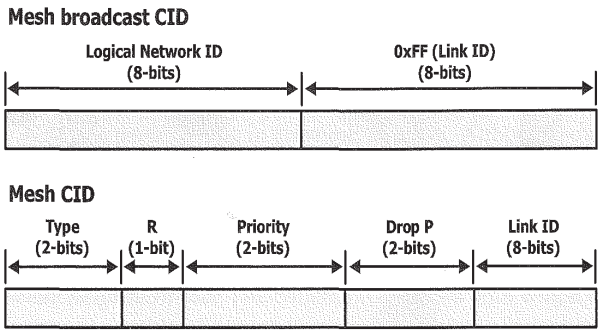


Figure 9–4: Mesh CIDs

Network configuration

Network configuration is achieved by means of broadcasting the MSH-NCFG (mesh network configuration) message. It contains both physical information of the device itself, such as the transmit power as well as network time-synchronization information, nearness to known Mesh BSs, a list of all known neighbors with which connections exist, and information about the next transmission of the MSH-NCFG from the node. MSH-NCFG is also used to assist with network entry, which is explained later.

The time-synchronization information consists of frame number, the control slot number, and the synchronization hop count. Because control slots have a fixed length, nodes are able to determine from this number when the frame must have started. Fine time synchronization is not done with this method; for this, the PHY is responsible. The synchronization hop count provides the number of hops by which the node is removed from the master synchronization node, which is typically a Mesh BS. Nodes will defer in their synchronization timing to nodes with a lower synchronization hop count. In a system that consists of multiple cells, it is possible to have only one master synchronization node so that all cells are synchronized.

Nearness to a Mesh BS is computed with two metrics. The first is a straightforward hop-count, which for any node is the lowest value reported by any of its physical neighbors, plus one. The second metric is the required energy per bit across all hops to the BS, which is computed by all nodes as the minimum of the value reported by its physical neighbors plus the ratio of the transmit power and achievable physical data rate to that neighbor. This second metric allows a better capacity tradeoff as it balances the number of hops vs the data rates over those hops.

Connections are split into two categories: *physical* and *logical*. A physical connection exists merely by means of being able to decode the MSH-NCFG message of the other nodes. Neighbors with which only a physical connection exists cannot be addressed for data transmissions. Logical connections require an underlying physical connection and are addressable for data exchanges.

When listing its physical neighbors with the MSH-NCFG message, a node indicates, for each neighbor, whether a logical link is present. It can also flag that, for this link, a logical link is desired, or, when a neighbor put up such a request flag, that the request is accepted. The physical connection information also contains schedule information for the next transmission opportunity as reported by the neighbor. For logical connections, the list also contains information such as link quality, used burst profile, and transmit power.

Transmissions of MSH-NCFG are scheduled in a distributed fashion. A substantial computational burden would result if, in a single control slot, all nodes compete to transmit a MSH-NCFG. Therefore, a method is provided to limit the control slots in which a node can compete as well as the number of

nodes that can compete in any particular slot. This method depends on the broadcasting by each node of two values, *Xmt Holdoff Exponent* (XHE) and *Next Xmt Mx* (NXM), and on reporting these values to each neighbor within the physical neighborhood so that the entire two-hop neighborhood of a node has access to the values for this node.

Xmt Holdoff Exponent is used to compute the number of control slots following the next MSH-NCFG transmission during which the node will not be eligible to transmit another MSH-NCFG as being 2^{XHE+4}. Both *Xmt Holdoff Exponent* and *Next Xmt Mx* are used to compute the eligibility interval for the next MSH-NCFG as

$$2^{XHE}(NXM) < \text{Next Xmt Time} \leq 2^{XHE}(NXM + 1)$$

For any control slot that falls within an eligibility interval, a node computes the list of nodes in its two-hop neighborhood whose eligibility intervals also contain this control slot. Each node then applies an algorithm to see whether it wins the election among this list of nodes. This guarantees that nodes transmitting MSH-NCFG simultaneously are at least three hops apart, which is generally enough to avoid mutual interference.

Network entry

Network entry for mesh devices consists of five steps:

- Scanning for active networks and coarse synchronization.
- Obtaining network parameters.
- Opening sponsor channel.
- Negotiating basic capabilities.
- Performing node authorization, registration, IP connectivity establishment, time-of-day establishment, and operational parameters transfer.

The first three are PHY and lower MAC functions that are substantially different from PMP behavior and are described below. The capabilities negotiation is identical to that performed under the PMP architecture, with the new node sending the SBC-REQ to the sponsor node (the node assisting the new node with network entrance). This capability exchange will be performed

with any node with which a new logical connection is established (contrary to the entire network entry process, which is performed only once). The last step is similar to behavior under the PMP architecture, with the main difference being that the new node tunnels the message exchanges over UDP between itself and the mesh BS through the sponsor node. Because this process is transparent to the sponsor node, the sponsor node cannot determine when the upper MAC initialization is successful; therefore, it is the new node that, as shown in Figure 9-5, indicates completion of this phase to its sponsor node.

Scanning for active networks and coarse synchronization

Scanning is a lower PHY functionality and hence architecture independent. A node will listen to the medium cycling through all available channels until it detects energy and a strong correlation between the known preamble sequence and the received signal. This correlation allows the PHY to synchronize itself to the timing of the start of that burst. The new node will then continue listening to subsequent bursts until its MAC detects MSH-NCFG messages. This synchronization process is termed *coarse synchronization* and should not be confused with the coarse and fine timing synchronizing the lower PHY performs when synchronizing the PHY to the burst itself.

Obtaining network parameters

A new node is required to continue monitoring the channel for MSH-NCFG messages until it hears the same node at least twice, ensuring that the new node has collected a reasonable subset of potential neighbors. It also will have to wait until it receives at least one MSH-NCFG with a Network Descriptor subfield, which is one of the five defined subfields to the MSH-NCFG. The network descriptor contains such parameters as the number of control slots, the scheduling methods used, information on used channels, and criteria for the use of burst profiles as well as the Operator ID; and it is only occasionally attached to the MSH-NCFG by each node. The entire process of obtaining network parameters may take seconds, but in a fixed system, it is required only when the system is turned on (the neighbor list and network parameters can be maintained across losses of synchronization that make reentry necessary).

Opening sponsor channel

Once the collection of network parameters is completed, the node starts its request to join the network, as outlined in Figure 9–5.

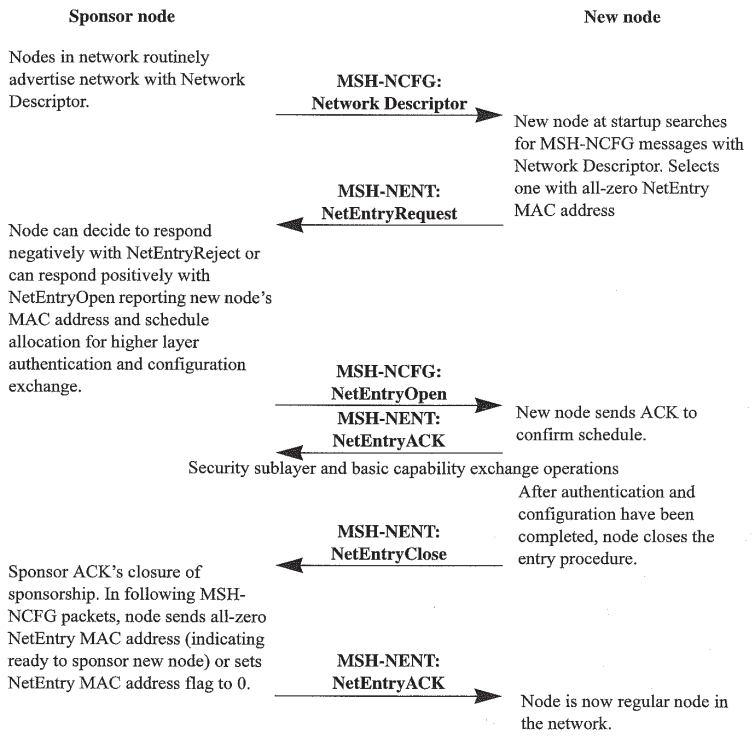


Figure 9–5: Mesh network entry: Opening a sponsor channel

For the purpose of joining the network, the node selects any of its neighboring nodes that are not currently assisting another node in network entry and transmits a MSH-NENT (mesh network entry) message with the NetEntryRequest subframe in the control slot reserved for network entry (see Chapter 11). The solicited node indicates its acceptance of the request by

setting the new node's MAC address in the next MSH-NCFG and publishing an arranged schedule for data transfer in the data portion of the frame to this new node; the solicited node then takes on the role of sponsor. (A solicited node that is currently assisting another candidate node in network entry will signal this fact by setting that candidate's MAC address in its MSH-NCFG messages until its role as sponsor is terminated.) Upon reception of the MSH-NCFG and schedule, the new node confirms the schedule with a MSH-NENT message with the NetEntryAck subfield and proceeds with the fifth step of network entry by use of the facilitated schedule. Upon completion of this fifth step, an exchange of MSH-NENT messages with NetEntryAck, initiated by the new node, terminates the sponsor's role and adds the new node into the network. Due to the tunneling, this last exchange is initiated by the new node.

Distributed scheduling

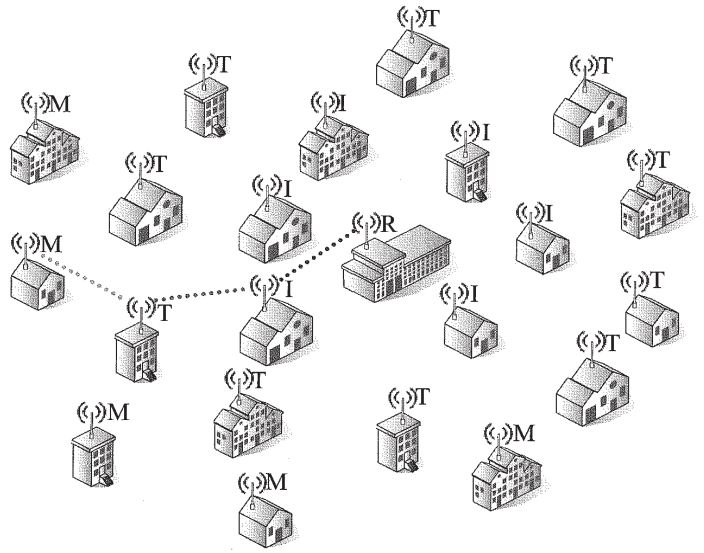
Distributed scheduling is also based on the assumption that if a node is not part of the extended neighborhood, the probability of interference is sufficiently low to be irrelevant. The distributed scheduling mechanism, therefore, ensures that the entire two-hop neighborhood is aware of the pending data transmission to ensure a near-collision-free transmission. The distributed scheduling protocol uses only the MSH-DSCH (mesh distributed scheduling) message format to execute a three-way handshake for exchanging: requests (in which data slots in which the medium is free to fill those requests are indicated), grants, and grant confirmations. This handshake occurs in the control portion of the frame, which is collision-free using the distributed algorithm also used for MSH-NCFG, with the two parameters (as for MSH-NCFG) sent with every coordinated MSH-DSCH message. The transmitting party sends out a request, consisting of a demand level and a persistence ranging from 1 to 128 frames, and the status of the medium of which it is aware. The status consists of "busy," "free for transmission and reception," "free for transmission," or "free for reception." This effectively notifies its immediate neighbors that the medium will be busy during the time indicated as free for the indicated direction of the transmission and during the time indicated as busy due to other known scheduled transmissions among the one-hop neighbors. The receiving party responds with a grant in some or all of

the requester's indicated free time, effectively notifying its immediate neighbors that the medium is reserved for the particular transmission during that period, which those neighbors will then mark as busy in their own schedules. The requester then confirms the indicated grant. The free medium time is computed as the period that none of the immediate neighbors has indicated as busy.

The only difference between coordinated and uncoordinated distributed scheduling is that for uncoordinated distributed scheduling, the handshake takes place in the data portion of the frame, making it susceptible to collisions due to attempts by other stations to set up an uncoordinated schedule.

The method is depicted in Figure 9–6, where node 1 is the node requesting the schedule, and node 2 is the node at which the request is directed. Node 3 is an immediate neighbor of node 1, but not of node 2; node 4 is an immediate neighbor of node 2, but not of node 1. Node 5 is a two-hop neighbor (i.e., within the extended neighborhood) of node 1.

When node 1 issues the request including an indication of potential free slots for the allocation to meet the request, it is received both by node 2 and node 3. Until it receives the grant confirmation sent by node 1, node 3 will assume the request will be granted in the explicitly indicated free slots and mark those temporarily as busy in its schedule. Node 2 will reply to node 1 with a grant in some or all of the indicated free slots, which also notifies node 4 of the allocation, and node 4 will mark those free slots as busy in its own schedule. The grant confirmation sent by node 1 finalizes the handshake between node 1 and node 2 and also serves to notify node 3 of the exact slots being used, as a result of which node 3 may mark the unused slots as free again. Because node 3 will advertise its availability before node 1 sends its grant confirmation (because of the cyclic way the control schedule is designed), node 5 (and any other node in the two-hop neighborhood) will effectively avoid any transmissions within the two-hop neighborhood during the same slots because they know from the free slots advertised by the one-hop nodes that those slots are being used by one of their two-hop neighbors.



I - Immediate/one-hop neighbors of either transmit or receive node
T - Two-hop neighbor
M - Multihop neighbor

Figure 9–6: Distributed scheduling

Centralized scheduling

The concept of centralized mesh is based on the notion that most traffic passes the gateway into the core network and that the node immediately connected to this core network, the Mesh BS, determines the flow of traffic throughout the entire cell. To achieve this, the BS dynamically splits the mesh-connected nodes up into a tree structure, which, by means of a schedule decided by the BS, is then used to arrange upstream and downstream traffic allocations. In a pure OSI model, knowledge of the routing tree would exist only at the routing layer. However, the IEEE 802.16 mesh specification defines knowledge of the tree at the MAC, which enables higher MAC efficiency.

To be able to decide how to update the tree structure (which at network initiation consists only of the BS itself), the current tree structure is used for a fixed schedule to pass down the tree MSH-CSCF (mesh centralized scheduling configuration) messages, which contain the number of nodes attached to the BS and for each of those nodes its ID and the IDs of all its child nodes. An example of a MSH-CSCF list is shown on the right-hand side in Figure 9–7, where the ID of each node is indicated with a single character, and the burst profile indicators in both directions are omitted. The order in which the nodes are listed in the MSH-CSCF is irrelevant, but it is used to indicate the child-parent relationship from which each node can reconstruct the entire tree.

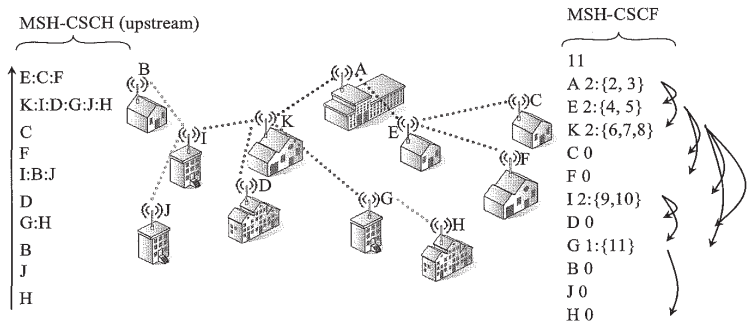


Figure 9–7: Centralized scheduling

The transmission rule for MSH-CSCF is that the BS transmits it first, followed by all its immediate children in list order, followed by the children of the first listed child in list order, etc. In the example in Figure 9–7, the transmission order is, therefore, identical to the list order; note that nodes without children do not rebroadcast the schedule. Due to its overhead burden, MSH-CSCF is used only occasionally, depending on the volatility of the network, whereas incremental changes can be distributed through the frequently sent MSH-CSCH (mesh centralized scheduling) message.

The MSH-CSCH messages serves three purposes:

- To distribute BW requests and grants up and down the tree, respectively. The requested and granted bandwidth is indicated as a function of instantaneous bits per second. The BS and all other nodes use their knowledge of the burst profile on each branch of the tree to translate this into an allocation of time. To facilitate the requests, each node uses a table the size of its branch in the MSH-CSCF list and fills, in MSH-CSCF list order, the BW requests for itself and all nodes down its branch. For example, in Figure 9–7, node K transmits a table with six entries, containing the BW requests of itself and the five nodes (I, D, G, J, and H) in its branch. The list order within the table is dictated by the list order in the MSH-CSCF, shown also in Figure 9–7, which all nodes including node K received from the BS. The transmission order for the requests, shown on the left side of Figure 9–7 starting with node H, is the reverse of the MSH-CSCF message, resulting in the table growing in size until the BS finally combines the pieces of the table into the full table. In their own request (for example, the entry for node K in the request table sent by node K), nodes indicate only the bandwidth needed for traffic that terminates in the node itself, not the cumulative traffic from the nodes downstream. In Figure 9–7, the upstream order of transmissions for MSH-CSCH with the request content of each transmission is shown on the left-hand side.
- To allow a node to request the BS to facilitate bandwidth to add a new node to the cell. The node initiating the request is termed a *sponsor node*.
- To enable the BS to notify the nodes of changes to the tree structure or changes in burst profile, which may be necessitated by changes in the propagation conditions. In this case, nodes and their new parent (or old parent, if only the burst profile changes) are identified with their respective list number in the last MSH-CSCF, which are 1-byte IDs.

Note that MSH-CSCH currently does not allow a node to send up a request for a downstream flow starting in the node, as the indication for this is limited to grants (which are issued by the Mesh BS). Hence, this requires that all downstream traffic be routed through the Mesh BS. As a result, the Mesh BS would need to inspect traffic to predict the required downstream flow grants,

and, although this need does not break the specification, it is impractical. This should, therefore, be considered an error in the specification, which is correctable by the removal of a single "if" statement in the MSH-CSCH frame format.

Note also that there is no facility for nodes to pass up changes to the tree except for the facilitation to add new nodes. This is done because deciding the tree structure is primarily a routing layer function, and hence the indication of the parameter set used by the Mesh BS to decide the tree is left to this layer. In other words, in practice, if a node loses the physical connection with its parent node in the tree, it would have to use the distributed scheduling mechanism using MSH-DSCH to achieve an update to the tree through another neighbor.

DIRECTED MESH AND POINT-TO-POINT (PTP)

Directed mesh in IEEE Std 802.16 can be thought of to some extent as transient PtP links from devices that could act as repeaters for the BS. The hooks for both directed mesh and PtP have been added for systems that use the WirelessMAN-SC PHY, but neither has been completely developed in the standard.

While a BS talking to a single SS in PMP mode is in principle a PtP link, it carries a significant amount of MAC overhead that is essential in the PMP case, but simply adds inefficiency to PtP operation. Furthermore, PtP range can be quite long through the use of highly directional antennas. This concept is shown in Figure 9–8.

The basic PtP link illustrated is a single-path link between the BS and an SS. With directed mesh, the concept is expanded; the SS has some ability to repeat all or some of the BSs transmission on another beam, in a different direction, and possibly on a different frequency. Such a system might appear as shown in Figure 9–9.

Physical or directed mesh can have many benefits. The reduced beam width can significantly improve the carrier to interference ratio (C/I) and improve frequency reuse. This gives high spectral efficiency. Conversely, network entry is greatly complicated, and the antenna systems are more expensive than with a standard PtP SS.

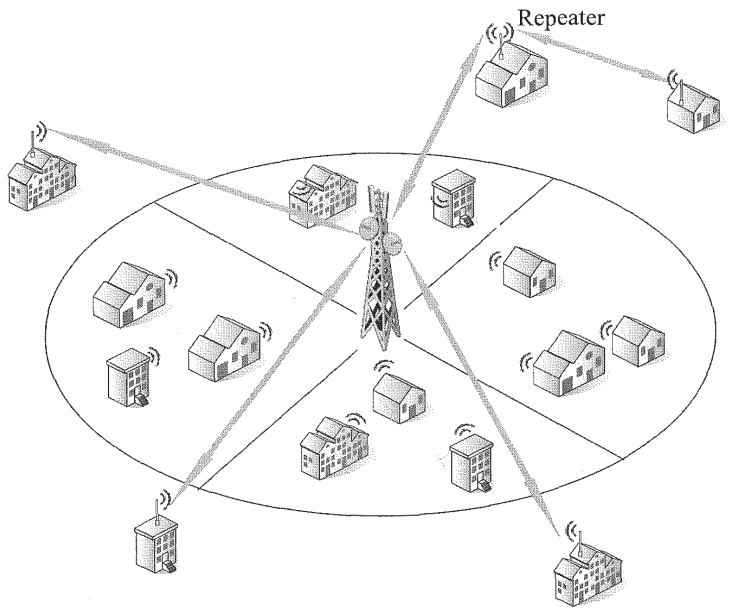


Figure 9–8: Addition of PtP

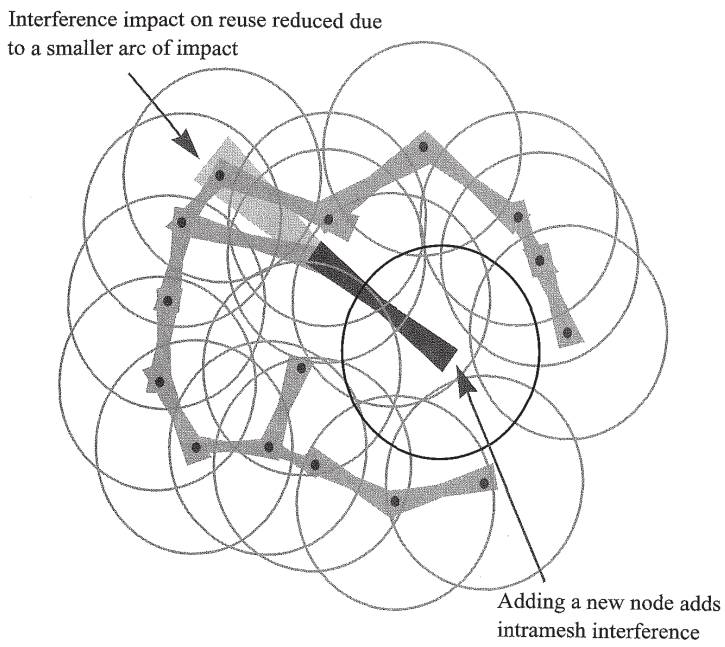


Figure 9–9: Physical or directed mesh

Chapter 10 PHY: WirelessMAN-SC

Single-carrier PHY for 10–66 GHz

The WirelessMAN-SC PHY specification is intended for operation in the 10–66 GHz frequency range. At these frequencies, LOS operation is a practical necessity due to the propagation characteristics. This makes mobility infeasible. While radios operating at these frequencies are typically expensive, highly directional antennas of practical size are available; therefore, given a LOS, multipath propagation can be negligible. This allows for a low-complexity, single-carrier PHY with good performance, even over a broadband channel. This is compatible with the large spectrum allocations in these frequency bands and the user data rates necessary to justify the radio costs with today's technologies.

As the original IEEE 802.16 PHY specification, WirelessMAN-SC set some precedents. For instance, both TDD and FDD configurations are supported in order to allow operation in worldwide spectrum allocations. Care was taken to make the FDD and TDD systems as similar as possible to minimize the impact of implementing both variants. Both have the same basic frame structure supporting adaptive burst profiling. The main difference is whether the DL subframe and UL subframe are sequential or simultaneous in time. The FDD case looks very much like the TDD case, but with the DL and UL subframes "stretched" toward each other so that they overlap. The FDD configuration also has special features to support half-duplex SSs.

The allocation of channel time (generally known in the IEEE 802.16 standard as *bandwidth allocation*) at the PHY is based on a combination of time division multiple access (TDMA) and demand-assigned multiple access (DAMA). In the UL, the channel is divided into a number of time slots (i.e., PSs) that are assigned to various purposes on a frame-by-frame basis by the BS MAC, based on demand. Time slots are available for ranging, contention-based BW requests, and user traffic. In the DL, the data bound for each SS are

multiplexed in a TDM fashion into a single stream of data and available for reception by all SSs within the same sector. This stream may change its PHY burst profile on QAM symbol boundaries within the frame to allow a tradeoff of robustness for efficiency. To support H-FDD SSs, provisions are made for resynchronization to the DL in midframe.

FRAME STRUCTURE

The WirelessMAN-SC PHY, like the other IEEE 802.16 PHYs, operates in a framed format. The frame structure eases the task of PHY synchronization. Within each frame, in the TDD case, are a DL subframe and an UL subframe. The DL subframe begins with information necessary for frame synchronization and control. While the DL data PDUs are addressed to individual SSs or to a broadcast address, the DL MAPs describing the DL merely indicate the occurrences of DL PHY burst profile changes; they do not indicate to which SS a series of time slots is directed. So each SS attempts to receive all portions of the DL except for the bursts whose profiles either are not implemented by the SS or are less robust than its current operational DL burst profile. The SS filters the data by CID, retaining only the data intended for it. Encryption ensures it cannot make use of data destined for other SSs anyway.

The WirelessMAN-SC PHY allows three different frame durations: 0.5 ms, 1 ms, and 2 ms. While not strictly required, the different frame durations are typically associated with different channel bandwidths, allowing a tradeoff of overhead vs protocol latency, with 1 ms frames being the most typical. Frame durations of 0.5 ms might typically be used for large channel bandwidths of 50 MHz or 56 MHz, while 2 ms frame durations might be preferred for channel bandwidth less than 20 MHz.

To keep the MAPs compact and to ensure that allocations are on byte boundaries, the building blocks for allocation are PSs of four QAM symbols each, yielding 1, 2, or 3 bytes, respectively, for QPSK, 16-QAM, and 64-QAM modulations. In the DL, the allocations are at this PS granularity. Because the number of QAM symbols in a 2 ms frame on a 56 MHz wide channel is greater than can be represented by the fields in the UL MAP (a 2 ms frame on a 56 MHz channel is allowed, but not recommended), the

granularity of the UL allocations are 2^m PSs (m ranging from 0 to 7). The typical value for m is 0 (1 PS granularity).

The basic structure for the DL subframe is taken from the TDD configuration of the system and is shown in Figure 10–1. The DL subframe starts with a frame start preamble used by the PHY for synchronization and equalization. This is followed by the frame control section, consisting of the DL MAP and UL MAP, broadcast to all SSs indicating the PSs at which the various bursts begin. The frame control section is followed by a TDM section carrying data and organized into burst profiles ordered by decreasing burst robustness. For example, assuming the FEC parameters do not change, data transmission begins with QPSK modulation, followed by 16-QAM, followed by 64-QAM. In the case of TDD, a TTG follows the DL subframe, allowing time for the BS to switch from transmit to receive and for the first SS of the UL to switch from receive to transmit mode.

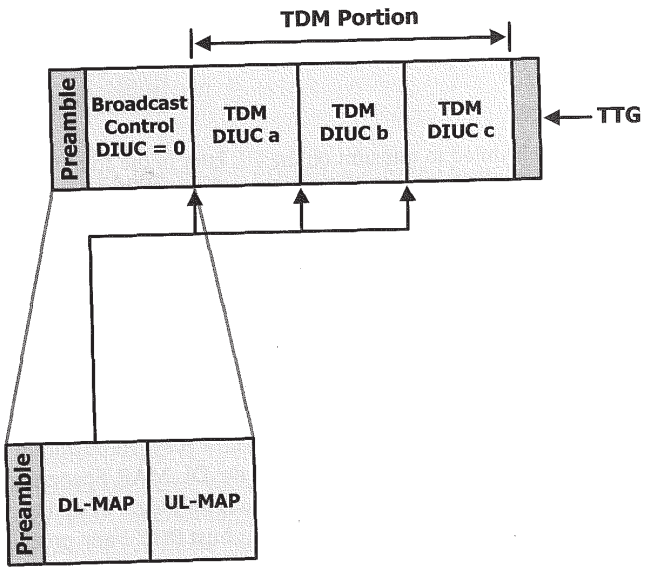


Figure 10–1: TDD DL subframe structure

The structure of the DL subframe for the FDD case can be quite similar to that of TDD operation, but becomes more complex with the introduction of H-FDD SSs. In the FDD case, the BS is full-duplex, but both full-duplex and half-duplex SSs are allowed. Although a full-duplex SS is capable of functioning as a half-duplex device, the IEEE 802.16 standard assumes that H-FDD SSs differ in being incapable of receiving while transmitting. This assumption allows for simplicity and cost savings in designing the H-FDD SS. When only full-duplex SSs are used, the FDD DL subframe is identical in structure to the TDD DL subframe except for the lack of a TTG. H-FDD SSs complicate matters, however, because whenever they are directed to transmit in the UL, they potentially lose synchronization (in particular, phase synchronization) with the DL. This makes it difficult or impossible for them to receive more DL data until resynchronized. As discussed below, a special preamble is specified for this purpose.

The FDD DL subframe structure is shown in Figure 10–2. As can be seen, the initial TDM portion is identical in structure to that of the TDD DL subframe. During this portion of the DL subframe, the BS can transmit data to the following:

- Full-duplex SSs.
- H-FDD SSs scheduled to transmit on the UL later in the frame than when they receive in the DL (the SSs that do not lose synchronization).
- H-FDD SSs not scheduled to transmit on the UL in this frame.

Unfortunately, in a system that has primarily H-FDD SSs, use of only the TDM portion could waste upwards of half the channel bandwidth, due to SS loss of synchronization.

To allow H-FDD SSs to transmit earlier than they receive in a frame, a TDMA section was introduced that differs from the TDM section in two important respects. First, extra preambles are inserted to allow resynchronization with the DL. The positions of these are described in the DL MAP. Second, given the extra preambles for resynchronization, there is no longer a need to smoothly transition through progressively less robust PHY burst profiles; therefore, the burst profiles used may occur in any order convenient to the BS's scheduling algorithm.

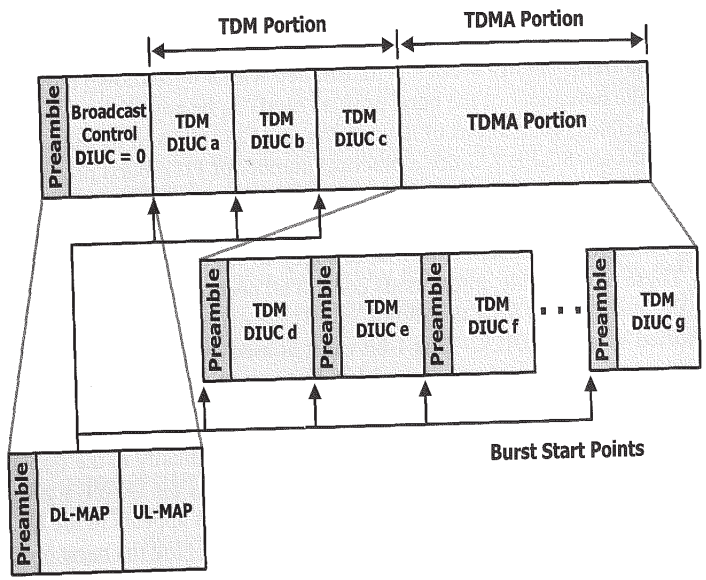


Figure 10–2: FDD DL subframe structure

Since the H-FDD SSs are nearly synchronized with the BS, the resynchronization preambles are shorter than the frame start preamble. The frame start preamble is 32 bits, created from the repetition of a 16-bit constant amplitude zero autocorrelation (CAZAC) sequence. The resynchronization preambles are 16 bits, created from the repetition of an 8-bit CAZAC sequence.

For WirelessMAN-SC, the frame control section of the DL always contains a DL MAP and a UL MAP, encoded with an especially robust well-known modulation/FEC combination (using QPSK). The DL MAP and UL MAP are structured so that the frame control section is always at least two FEC blocks in duration, but the first DL PHY transition will always be identified in the first FEC block of the frame, even if the DL MAP message does not terminate

in that block. This ensures that the SS will be warned of the upcoming PHY burst change at least one FEC block in advance.

While the frame control section is transmitted with a well-known burst profile, the remainder of the burst profiles used can be any of the over 4000 possible combinations of modulation and FEC allowed by the WirelessMAN-SC PHY. Since a practical number of PHY burst profiles is closer to four than to 4000, and also to avoid ballooning the DL MAP with numerous PHY parameters, the frame control section periodically contains DCD messages that describe the four or so (up to 13) burst profiles used in the DL. These burst profile descriptions contain all necessary information, such as FEC, modulation, SNR thresholds, etc. These parameter sets are then identified by a 4-bit DIUC. This 4-bit value is used in the DL MAP as a shorthand for the numerous bytes of data that would otherwise be required to describe each PHY burst.

Similarly, for the UL, the UCD message appears periodically in the frame control section of the DL. It describes the parameters of the UL PHY burst profiles in use and identifies them by a 4-bit UIUC that appears in the UL MAP as a shorthand for the actual parameter sets describing the burst profiles.

DL CHANNEL ENCODING

The DL data sections are used to transmit data and control messages to specific SSs. The data are always FEC coded and are transmitted at the current DL operating modulation of each SS. In the TDM portion of the DL subframe, the data are transmitted in decreasing order of burst profile robustness. Within a particular burst, the data transmitted at a given burst profile may contain data for multiple SSs multiplexed together. It is the responsibility of the SS to extract its data from data destined for other SSs sharing a common operational DL burst profile. In the TDMA section, the data are grouped in separate burst profiles that need not be in order of robustness. The BS may, but is not required to, segregate the data by SS as well as by burst profile. However, the DL MAP generally does not indicate the target SS; therefore, SSs listen to all bursts for which they are capable.

While a new DL burst always starts on a PS boundary, the choice of FEC parameters and modulation may cause FEC blocks to end in the middle of a PS or even in the middle of a QAM symbol. However, they always end on bit boundaries. The next FEC block starts on the next available bit, without padding to the next symbol or PS boundary. The last data for a burst may not be sufficient to fully fill an FEC block. In this case, the last FEC block of the burst may be shortened. Unused bits in an unshortened FEC block or the bits used to allow realignment of the next DL burst to a PS boundary are padded with all ones. Figure 10–3 illustrates the allocation of data to a DL burst. If the burst contains a preamble, as is the case for the frame control section and TDMA bursts, the preamble is considered part of the burst and precedes the first FEC block of the burst.

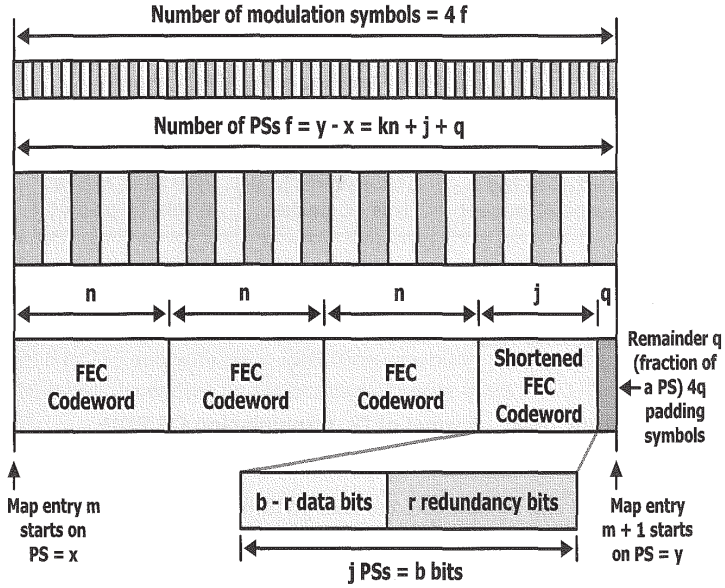


Figure 10–3: DL MAP usage with shortened FEC blocks—TDM case

Within each FEC block, additional information is added to help resynchronize the MAC to the next MPDU within the bit stream in the case of an unrecoverable error in an FEC block. This capability is handled by the transmission CS. Since MPDUs are variable in length, while the FEC blocks of a burst are of fixed length, it is common for MPDUs to begin or end in the middle of an FEC block. The MAC header contains the length of the MPDU, which is required for extracting the MPDUs from the received bit stream. Unfortunately, if an FEC block has unrecoverable errors, this length information may be lost or erroneous, causing difficulty in finding the start of the next MPDU. To avoid the risk of the SS losing all data until the end of the PHY burst, each FEC block uses the first byte of its payload not for data but for a field indicating where, if anywhere, within the FEC block the next MPDU starts. This allows the SS to quickly resynchronize to the next MPDU rather than losing the remainder of the burst. This FEC payload structure is shown in Figure 10–4. The same mechanism is employed in the UL.

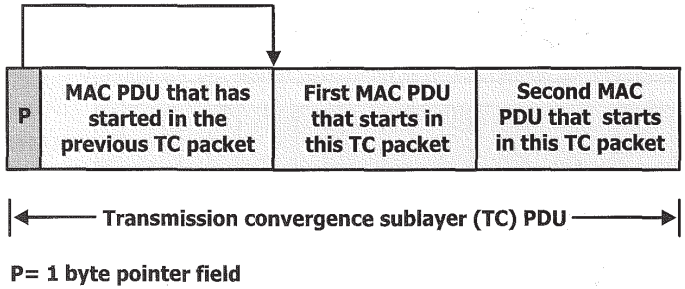


Figure 10–4: Format of DL transmission CS PDU

In addition to FEC and modulation, the data go through a randomization process and pulse shaping. The processing chains for transmit and receive are shown in Figure 10–5.

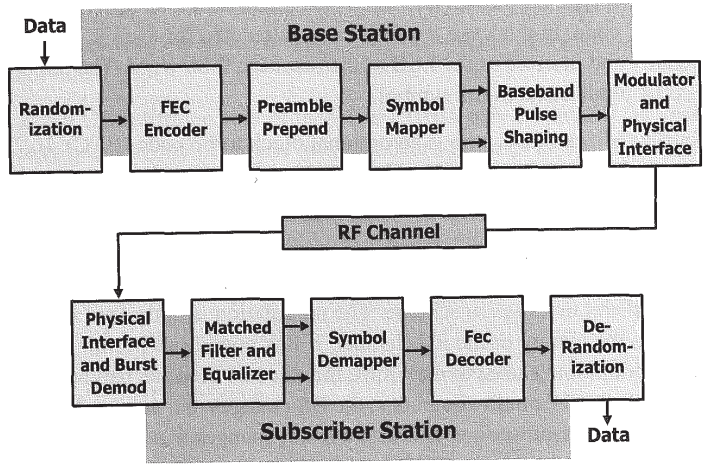


Figure 10–5: Conceptual block diagram of DL PHY

There are four basic types of FEC used in WirelessMAN-SC. These can be further parameterized to give many possible combinations.

The first type, which is specified as mandatory, is a Reed-Solomon (RS) outer code over Galois field (GF) 256 with no inner code. It is generally used for user data except when, even with QPSK modulation, a stronger coding is desired. It can be parameterized with the information block length K variable from 6 to 255 and with error correction capability to correct errors from $T = 0$ to 16 bytes, depending on the number of parity bytes $R = 2T$.

The second type of FEC, also mandatory, is used for the frame control section and for transmissions that need to be very robust. The RS is used with an inner (24, 16) block convolutional code (BCC) derived from a four-state, nonsystematic punctured (7, 5) convolutional code. In this case, the number of information bytes K for the outer RS code must be an even number so that the total size $(K + R)$ is also even, since the inner BCC operates on pairs of bytes. In the frame control section, the parameters are set to $K = 26$ and $R = 20$.

The third FEC type is optional and combines the RS outer code with a (9, 8) parity check code. In this case, a parity bit is calculated for each RS symbol individually and inserted as the LSB of the resulting 9-bit word. The parity bit is the XOR of all 8 bits in the symbol.

The fourth and final FEC type is an optional block turbo code (BTC). If implemented, both (39, 32) × (39, 32) and (53, 46) × (51, 44) codes are required. Shortening can apply to the last FEC block of a BTC-encoded burst, as in the RS case.

To maximize the spectral efficiency of the airlink, the WirelessMAN-SC PHY uses a multilevel modulation scheme. The modulation is optimized for each SS based on the quality of the RF channel. If link conditions permit, a more efficient modulation is used to maximize the tradeoff between bandwidth and robustness. If the RF quality degrades, the BS changes to a more robust modulation for that subscriber. If it improves, it changes to a more efficient modulation.

The WirelessMAN-SC PHY supports QPSK, 16-QAM, and 64-QAM providing 2 bits, 4 bits, and 6 bits per QAM symbol, respectively. Two power adjustment rules can be used when transitioning from one modulation to another: *constant peak power* and *constant average power*. In the constant peak power scheme, corner points are transmitted at equal power levels regardless of modulation. In the constant average power scheme, the signal is transmitted at constant average power regardless of modulation. The power adjustment rule is communicated via the DCD message. If a burst at a specific modulation does not end exactly on a PS boundary, the remainder of the burst is filled with zero bits.

Prior to modulation, the I and Q signals are filtered by square-root raised cosine filters with the excess bandwidth factor α set to 0.25.

UL CHANNEL ENCODING

In the WirelessMAN-SC PHY, three types of bursts may be transmitted in the UL subframe:

- Contention-based initial ranging bursts.

- Contention-based multicast or broadcast BW requests.
- Dedicated data bursts in response to unicast bandwidth grants allocated to individual SSs.

In a given UL subframe, any of these burst types may be present (or absent) and may occur in any order and quantity. The scheduler at the BS indicates which bursts are present via the UL MAP message. The bursts are separated by subscriber station transition gaps (SSTGs), allowing for the ramping down of the previous burst and the ramping up and preamble of the next burst. The preamble allows the BS to fine-tune its synchronization to the transmission of the individual SS. As in the DL, the preambles may be 16- or 32-bit repetitions of 8- or 16-bit CAZAC sequences. The UCD message communicates the gap and preamble lengths. Figure 10–6 illustrates an example of the UL subframe.

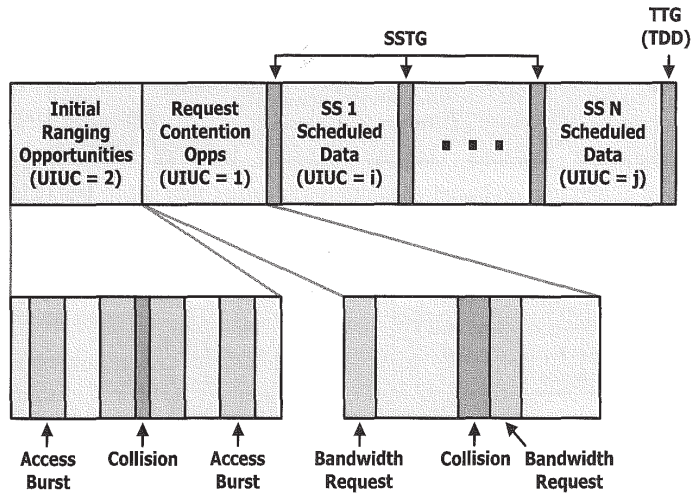


Figure 10–6: UL subframe structure

Request bursts always use UIUC = 1, which always uses QPSK modulation and FEC type 1 or type 2 as defined for the DL. Request bursts are always at

least as robust as the most robust unicast UL data bursts. Initial ranging bursts always use UIUC = 2, which specifies burst characteristics identical to those used in the DL in the frame control section. The UL MAP entries of WirelessMAN-SC are directed to the Basic CID (or multicast group or Broadcast CID) of a specific SS and are relative to the previous MAP entry. They specify the start of the burst, but the next MAP entry implies the end of the burst. The last burst of a UL subframe has an end symbol indicated by an end-of-MAP entry allocated to CID = 0 with UIUC = 10 (decimal).

The UL uses the same FEC and the same transmission convergence subframe as the DL. Also, like the DL, in addition to FEC and modulation, the UL data go through a randomization process and pulse shaping. The processing chains for transmit and receive are shown in Figure 10–7.

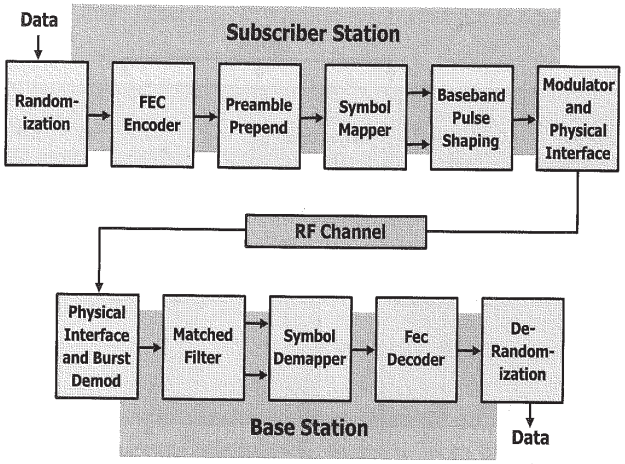


Figure 10–7: Conceptual block diagram of UL PHY

As in the DL, the I and Q signals, prior to modulation, are filtered by square-root raised cosine filters with the excess bandwidth factor α set to 0.25.

This 0.25 roll-off factor defines the symbol rate, which, for a 25 MHz channel, would be 20 Mbd. This would give bit rates for QPSK, 16-QAM, and 64-QAM of 40 Mbit/s, 80 Mbit/s, and 120 Mbit/s, respectively. Other channel sizes scale linearly. For example, a 28 MHz channel has a 22.4 Mbd symbol rate.

CONTROL MECHANISMS

The DL demodulator must provide an output reference clock derived from the DL symbol clock. This reference is then used by the SS to provide timing-critical interfaces. One would expect that the BS reference is locked to an accurate (stratum 1) reference such as GPS. This highly accurate timing is used to provide frequency tuning to within $\pm 10 * 10^{-6}$.

The WirelessMAN-SC PHY requires accurate UL time-slot synchronization and UL transmit time advance to ensure that multiple users do not interfere with each other's transmissions. The ranging procedure outlined in Chapter 7 allows the fine-tuning of timing required for robust operation with efficient burst profiles.

For WirelessMAN-SC, the predominant dynamic impairment is rain fade. The typical extent of the fade varies geographically based on rain region and is affected by both carrier frequency and range. For a desired link availability and given set of transmission parameters, the rain region dictates the maximum cell radius. Additionally, uncorrelated rain fade between an interfering signal and the desired signal may also affect cell radius.

Power control is applied statically in the DL to effectively set cell size and dynamically in the UL to account for fading due to rain and foliage. This dynamic UL power control must handle fades of up to 20 dB/s with dynamic range of at least 40 dB. The specific algorithm is left for vendor differentiation. The BS's EIRP spectral density is capped by +28.5 dBmi/MHz or a lower limit if established by local regulations. The SS is limited to an EIRP spectral density of +39.5 dBmi/MHz or local regulation.

Received signal strength indicator (RSSI) and CINR signal quality measurements and associated statistics are used in burst profile selection. RSSI measurements do not necessarily require receiver demodulation lock

and are, therefore, used for measurement of low signal levels. CINR requires receiver lock, but provides better information on the actual operating condition of the receiver, including interference and noise levels as well as signal strength.

WirelessMAN-SCa

This chapter concludes with a brief description of the WirelessMAN-SCa PHY specification, which is targeted at operation below 11 GHz. In many respects, it is very similar to the WirelessMAN-SC PHY. It was originally added to the standard in IEEE Std 802.16a-2003, with the tacit assumption that it would be used in conjunction with frequency-domain equalization. It has not found subsequent industry support and, at this time, appears to be an orphan.

The WirelessMAN-SCa FEC specification is similar to, but slightly different from, WirelessMAN-SC. The basic FEC is a concatenated FEC, with an outer RS code derived from a systematic RS ($N = 255$, $K = 239$) code using GF(256). The block may be shortened. An inner trellis-coded modulation (TCM) code derived from a $K = 7$, rate 1/2 convolutional code is used. Optionally, an interleaver may be placed between the inner and outer codes.

In addition to QPSK, 16-QAM, and 64-QAM, the WirelessMAN-SCa PHY also requires the implementation of spread binary phase shift keying (BPSK) and BPSK modulations. 256-QAM modulation is optional. Block turbo codes (BTC) and convolutional turbo codes (CTC) are also optional. When ARQ is implemented, the FEC can be disabled (through the "no FEC" option), and the ARQ can be used as a sole method for error correction.

Like WirelessMAN-SC, the WirelessMAN-SCa PHY is defined for both FDD and TDD operation. The WirelessMAN-SCa PHY includes explicit support for beam forming.

The system must handle fade rates of 30 dB/s with depths of 10 dB. Both RSSI and CINR are used for measurements of signal quality. Frequency accuracy must be within $\pm 15 * 10^{-6}$.

Chapter 11 PHY: WirelessMAN-OFDM

Multicarrier PHY for frequencies below 11 GHz

WirelessMAN-OFDM is one of the three PHY specifications defined in IEEE 802.16 for applications below 11 GHz. For brevity and convenience, the PHY of a WirelessMAN-OFDM system is abbreviated as the "OFDM PHY."

Waveform Construction

Selection of the OFDM waveform

The choice of OFDM for use below 11 GHz was in general based on two considerations. The first of these was that, among the IEEE 802.16 members, a fair number had experience with OFDM systems due to the approval of the IEEE 802.11a amendment in 1999. In addition, a small number of participants were involved with companies developing proprietary OFDM-based systems for BWA. Selecting OFDM was for these members a natural choice. Aside from these external issues, however, OFDM was supported because of its favorable technical characteristics.

As noted in Chapter 3, robust BWA systems in spectrum below 11 GHz must deal with multipath without resorting to highly directional antennas, which would make deployment costly. As noted, one approach is to use redundant samples preceding transmitted symbols to ensure that any received samples affected by energy from the previous symbol can be discarded. This process is efficient only if the duration of the transmitted symbols is considerably longer than that of the expected multipath; otherwise, the redundancy would be constituting an unacceptably large relative overhead. In addition, the method used to modulate data onto these large symbols must be computationally simple. This is important because the number of bits per symbol will have to be very large to maintain a given data rate.

The data rate, in bits per second, can be expressed as [(bits/symbol) × (symbol/s)]. The symbol rate, (symbol/s), is the inverse of the symbol duration so that increasing the symbol duration requires modulating more bits per symbol to maintain the data rate (assuming a given channel bandwidth).

The preferred method of creating such symbols is the use of the IFFT, which effectively transforms a set of orthogonal narrowband transmissions into a single waveform. An additional advantage of the IFFT is that it is cyclical in the time domain so that when the redundant samples are copied from its tail, the transmitted symbol remains continuous in phase. OFDM provides for the subcarriers to be orthogonal and separable. When the subcarrier spacing is sufficiently small, the channel characteristics can be safely assumed to be constant across the subcarrier. This assumption drastically simplifies channel estimation because it can be performed independently for each subcarrier and, for each subcarrier, the channel can be represented by a single amplitude and phase.

Selection of FFT size

The FFT size is a critical parameter with implications in the time domain as well as the frequency domain. The designer must take into consideration a extraordinarily wide range of system factors, including the multipath propagation environment, the cost of precision electronics, and the nature of the network traffic. This job requires system-level thinking; it is not for narrow specialists.

The most fundamental frequency domain characteristic is the channel bandwidth. IEEE Std 802.16 is designed to accommodate a number of channel bandwidths, depending on the local regulatory environment and the operator requirements. For the purpose of this discussion, we will assume that the channel bandwidth is fixed.

In a given channel bandwidth, the subcarrier spacing is inversely proportional to the number of subcarriers and, therefore, to the FFT size. Meanwhile, the duration of the OFDM symbol is set to be the inverse of the subcarrier spacing. This setting is the necessary and sufficient condition to ensure that the subcarriers are orthogonal. As a result, the OFDM symbol duration is proportional to the FFT size.

The choice of FFT size is inextricably tied to assumptions about the physical environment in which the propagation takes place. This is because, to ensure that multipath propagation does not corrupt the received signal, a copy of the last T_g s of the symbol waveform with duration T_b is prefixed to the symbol before transmission (see Figure 11–1). This copy is termed the *cyclic prefix*. The receiver will sample only an interval T_b within this extended interval $T_b + T_g$. The prefix has no impact on the received symbol due to the cyclic nature of the FFT. Multipath components shorter than T_g s need not negatively affect the performance as long as the sampling interval is chosen correctly. How much smaller the multipath components need to be depends on the accuracy of sampling timing and the symbol windowing used because both negatively affect the protection against multipath. Multipath components in excess of the effective multipath protection duration, however, will significantly degrade the performance (see [B55]). It is, therefore, desirable to choose the cyclic prefix large enough to encompass the expected multipath of the application.

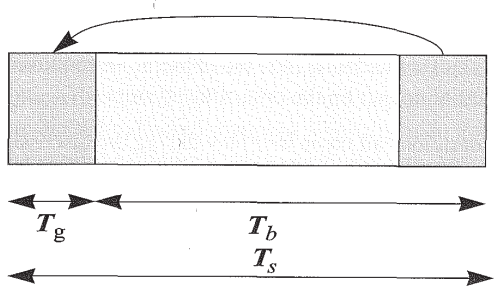


Figure 11–1: OFDM symbol structure

The addition of the cyclic prefix adds overhead due to the redundancy. It also increases the energy per symbol, resulting in an SNR drop of $10\log_{10}(1 - T_g/(T_b + T_g))$ dB. With the duration of the cyclic prefix dictated by the multipath, choosing a shorter OFDM symbol duration hence increases the proportion of time and energy lost on the cyclic prefix.

The designer must also consider the network traffic requirements. To understand why, begin with the fact that the FFT size determines the number of available subcarriers. Each subcarrier is modulated to carry data. For this discussion, assume QAM modulation, with each subcarrier delivering one QAM symbol during the duration of the OFDM symbol. The total number of bits carried in one OFDM symbol by all the subcarriers, therefore, scales with the FFT size (as does the symbol duration). This is where the network factors enter the problem. As the data capacity per symbol gets larger, the granularity may grow too coarse to efficiently carry variable-length data packets, particularly when the traffic consists of many small data packets that do not fill the symbol. For instance, if the UL is transmitting a few small packets, as typically found in VoIP and most other IP traffic, without a sufficient packet rate to take advantage of packing, then large symbols may need to be transmitted without a full load of data. In this case, a shorter symbol would be more efficient.

Finally, a longer OFDM symbol duration translates into narrower subcarriers in the frequency domain. This closer spacing makes the receiver performance more susceptible to phase noise, which corrupts the orthogonality of the subcarriers. With current technologies, the phase noise requirements have a direct and substantial impact on the cost of the system.

The above considerations all came into play during the discussions on the FFT size N_{FFT}, an issue complicated substantially by the targeted support of a large range of bandwidths (up to 28 MHz). In the early stages of development, the IEEE 802.16 Working Group considered many different FFT sizes, particularly 64 and 256 for OFDM and both 2048 and 4096 for OFDMA. Although it was realized that 64-FFT was suitable only to short links due to its low multipath protection capabilities, the opportunity to reuse IEEE 802.11a implementations was a strong consideration for many participants. Eventually, the 64-FFT option was dropped as an OFDM alternative, as performance issues outweighed issues regarding the reuse of legacy components; and OFDMA was restricted to 2048. With the IEEE 802.16e amendment, additional sizes of 128, 512, and 1024 were added for OFDMA. To provision in a more optimum fashion for deployments with short links and deployments with relatively little multipath, the IEEE 802.16 standard allows for several different cyclic prefix durations.

The OFDM waveform

The OFDM waveform as a function of time can be described as follows:

$$s(t) = \text{Re}\left\{ e^{j2\pi f_c t} \sum_{\substack{k=-N_{used}/2 \\ k\neq 0}}^{N_{used}/2} c_k \cdot e^{j2\pi \cdot k \cdot \frac{n \cdot BW}{N_{\text{FFT}}} \cdot (t - T_g)} \right\} \qquad 0 < t < T_s$$

where

c_k	is a data point out of the modulation constellation to be mapped onto the frequency offset index k
f_c	is the RF
BW	is the transmission bandwidth
n	is the oversampling factor
N_{used}	is the number of nonsuppressed subcarriers

In the early stages of the development of the OFDM and OFDMA PHYs, the working group attempted to maintain as much commonality between the two specifications as possible. This, along with the fact that IEEE Std 802.11a used 48 data subcarriers per OFDM symbol and the fact that 48 data subcarriers conveniently allowed mapping whole bytes to a single OFDM symbol for almost all modulations, led to a symbol structure containing a multiple of 48 data subcarriers. For the case of the 256-FFT, the working group settled on 192 data subcarriers. Eight additional "pilot subcarriers" are used mainly to track phase changes during a burst, making for a total of $N_{\text{used}} = 200$ subcarriers. The resulting frequency domain description is shown in Figure 11–2.

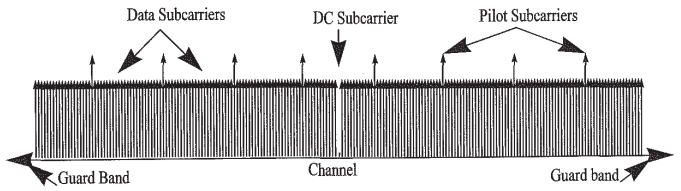


Figure 11–2: OFDM frequency description

At this point, it would have been possible to use the 200 subcarriers, along with a dc subcarrier in the center and guard subcarriers on either side (28 on one side and 27 on the other). However, the resulting spectral shape would end up unnecessarily conservative with regards to spectral masks established by relevant regulatory agencies; the result would be akin to a waste of spectrum. To take advantage of the opportunity to fully utilize the spectrum within the mask, the sampling factor n was introduced to create a sampling bandwidth somewhat larger than the actual bandwidth BW. This stretches the spectral shape of the signal to more fully fill the space with the spectral mask, shortening the symbol duration and hence leading to slightly higher throughput. The allowed values of n are as follows:

- 8/7 for BW multiples of 1.75 MHz
- 86/75 for BW multiples of 1.5 MHz
- 144/125 for BW multiples of 1.25 MHz
- 316/275 for BW multiples of 2.75 MHz
- 57/50 for BW multiples of 2 MHz
- And again 8/7 for any other bandwidth

These specific values ensure that there is an integer number of samples per frame for all likely bandwidths less than 28 MHz.

Subchannelization

The access method for the OFDM PHY is TDMA. This is, however, not an entirely complete description, as the PHY also provides an integrated optional OFDMA component. To avoid confusion between this component and the OFDMA PHY described in Chapter 12, the OFDMA component in the OFDM PHY is referred to as *subchannelization*.

In late 2002, the IEEE 802.16 Working Group was completing the IEEE 802.16a amendment, which added the new PHY specifications for frequencies below 11 GHz. The working group recognized the importance of specifying profiles as the basis of compliance testing for these alternatives, but the group was anxious to complete IEEE Std 802.16a. Therefore, it decided to develop an amendment, IEEE P802.16d, to add the missing profiles; the same project could be used to address any errors identified in

IEEE Std 802.16a. At about the same time, the working group, following formal study group activity, was initiating activities to expand IEEE Std 802.16 from a fixed-only to a fixed-and-mobile system, beginning with the opening of the IEEE P802.16e PAR in late 2002. Within the IEEE 802.16e effort, it was soon realized that the OFDM PHY as defined did not provision sufficient UL link budget to move from a fixed deployment with relatively large SS antenna gain to a mobile platform with typically very low SS antenna gain, while maintaining a reasonable cell size. Subchannelization was a known tool to extend the link budget. While it made sense for mobile terminals, the working group began to accept the idea that it would be applicable to fixed terminals also.

Another important factor in introducing subchannelization into the IEEE 802.16 OFDM PHY was the ETSI BRAN HiperMAN project. ETSI BRAN HiperMAN, insisting on the need to minimize alternatives in the standard, had at one point reached agreement on a standard that was fully compatible with WirelessMAN-OFDM. However, as IEEE Std 802.16a concluded, ETSI BRAN HiperMAN continue to progress with the introduction and development of subchannelization, even though ETSI BRAN HiperMAN was not chartered to consider mobile operation. The desire to maintain compatibility with ETSI BRAN HiperMAN was one reason that the IEEE 802.16 Working Group sought to introduce the feature. In 2003, the amendment project IEEE P802.16d was converted into a full revision project. This expanded the scope of the work and allowed for the introduction of features like subchannelization.

After the idea of adding subchannelization gained acceptance, the main contention point was how many subchannels to use. Some sought small subchannels to maximize the link-budget gain, while others sought larger subchannels due to concerns with the minimal achievable throughput per subchannel, losses due to the scrambling and encoding blocks getting very small, and the general increase in receiver complexity.

Eventually, the number of subchannels was chosen to be 16, each consisting of 12 data subcarriers. Subchannels may be combined in a predetermined fashion so that an SS can be allocated all of the data subcarriers or one-half, one-fourth, one-eighth, or one-sixteenth of them. Only eight pilot subcarriers

are available for all of the subchannels. Therefore, for allocations of one-sixteenth of the data subcarriers, phase estimation must rely solely on data-driven approaches. The subchannel structure is shown in Figure 11–3.

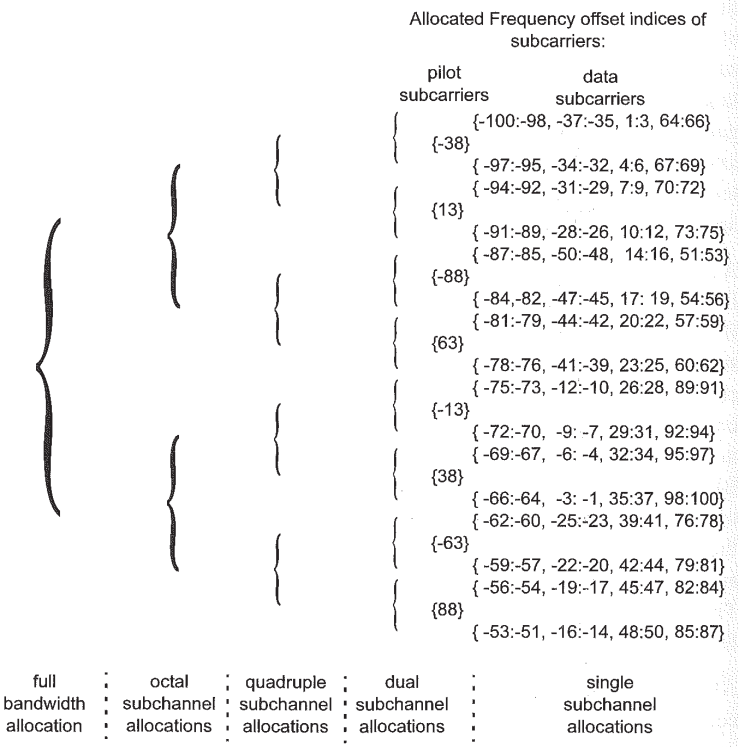


Figure 11–3: Subchannel subcarrier allocations

The OFDM PHY can be applied to make use of the increased UL link budget provisioned by subchannelization, by limiting the maximum number of subchannels that are assigned in the UL to a given SS. This increased link budget is available from subchannelization because an SS can distribute its available power to fewer subcarriers, achieving a power gain per subcarrier (and hence link budget) of 3 dB every time the number of excited data

subcarriers is halved. With a division into 16 subchannels, a total gain of 12 dB is hence possible (although implementation losses will somewhat reduce this gain). To exploit this opportunity, an SS with insufficient link budget to establish a link using full bandwidth can establish a link using less than the full bandwidth at the cost of peak throughput (but not at the cost of system capacity). To support this, special provisions allow efficient initial ranging and BW requesting for these cases. In order to make full use of this feature, however, sufficient link budget must be achieved on the DL, generally through to the use of a higher gain BS amplifier, to enable establishment of a connection. Especially in license-exempt bands, where allowed EIRP may be fairly restrictive, this may not be possible.

Aside from link range extension, subchannelization may additionally be used to exploit the lower overhead of the subchannelized preamble and to increase the modulation/coding efficiency. This may be easily understood from a simple example. A full bandwidth transmission of data filling x OFDM symbols requires $x + 1$ OFDM symbols (the extra OFDM symbol resulting from the preamble). The overhead ratio is hence $1/(x + 1)$. Transmitting that same data over eight subchannels (one-half the bandwidth) using the same data rate results in an overhead ratio of $1/(2x + 1)$ within those subchannels because the data now occupy eight subchannels of twice as many OFDM symbols, but the preamble only occupies eight subchannels of a single OFDM symbol. Because the remaining eight subchannels remain free for another SS, no loss of system capacity occurs (in fact, the reduced overhead slightly increases the system capacity). Because of the 3 dB gain achieved by using only half the subchannels, the SS may additionally be able to increase the data rate by increasing the modulation/coding efficiency, requiring less OFDM symbols (e.g., $1.6x$) to transmit the data. The total system resources used is then $(1 + 1.6x)/2 = 0.5 + 0.8x$, compared to the full-bandwidth $1 + x$.

The data subcarriers assigned to a single subchannel are distributed over the entire OFDM symbol, in four blocks of three subcarriers each. This spreading was introduced as a frequency diversity technique to reduce the suspectibility to partial band fading, whereas the grouping in blocks of three subcarriers results in slightly lower phase noise and frequency offset sensitivity. The allocation of blocks is such that combining subchannels still results in an evenly spread allocation of subcarriers. During a single OFDM symbol

interval, it is possible to assign different numbers of subchannels to individual SSs as long as the frequency offset indices of the subchannels do not overlap.

One of the four major changes introduced in WirelessMAN-OFDM by the IEEE 802.16e amendment is the capability for provisioning subchannelization in the DL (the other three changes are the additions of private MAPs, private initial ranging opportunities, and open-loop power control). This DL subchannelization is provided as a separate zone, similar to the zone concept of the OFDMA PHY (see Chapter 12), with its own subchannelized preamble, FCH, and MAPs so that it is completely independent of the remainder of the frame. The DL subchannelization is restricted to the use of quadruple subchannel allocations (see Figure 11–3). The DL MAP in this construct provides allocations only for future frames, not for the current one, as it is assumed that mobile devices do not have the computational power necessary to decode this MAP quickly enough to efficiently avoid receiving data not intended for the station itself; this would unnecessarily drain the battery.

The addition of DL subchannelization is motivated by the desire for a larger coverage area for mobile devices, which have comparatively low antenna gain.

Preambles

The OFDM PHY specifies two 200-element sequences from which all preambles are derived. One sequence is used for all DL preambles and full-bandwidth UL preambles; the other is for UL allocations spanning less than the full channel. These two sequences are termed *mother sequences*. Computing an actual preamble from a mother sequence involves operations no more complex than multiplication with a constant, conjugation, and puncturing. This makes it practical to store only the two mother sequences.

All preambles are one OFDM symbol long, with exception of the DL preamble, initial ranging preambles, and mesh preamble, which consist of two OFDM symbols.

The first symbol of the two-symbol preambles, shown in Figure 11–4, consists of a cyclic prefix (CP) followed by four repetitions of a frequency domain derivative of the first mother code. Each repetition is said to be

length 64, which means that it contains 64 samples and has a duration four times shorter than that of the 256-FFT OFDM symbol. This symbol primarily allows the receiver to perform large-gain adjustments of its amplifier, as well as coarse frequency and time correction. This is necessary on the DL to allow new SSs to properly track and adjust to the BS signal. On the UL, it is necessary for initial ranging, as the new SS's transmit gain and frequency/timing offsets have not yet been adjusted to within an acceptable range for the BS receiver. For other transmissions in the UL, the BS directs each SS to adjust its transmit power and frequency/timing offsets to within a desired range so that the preamble need not facilitate this functionality for these transmissions.

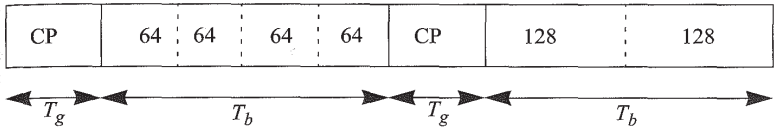


Figure 11–4: DL and network entry preamble structure

The second symbol of the long preambles as well as all of the one-symbol (short) preambles consists of a cyclic prefix followed by two repetitions of a length 128 frequency domain derivative of the mother codes, which is transmitted on the even subcarriers. Special cases exist for the DL STC-encoded region and for DL AAS transmissions. For the former, alternate subcarriers are transmitted from the two different antennas. For the latter, two identical "regular" OFDM symbols (each with cyclic prefix) are transmitted, where each beam (not each antenna) transmits on every fourth carrier. If a BS has more than four simultaneously active beams, the fifth beam transmits the same symbol as the first antenna, etc. This is possible because of the spatial separation.

For UL allocations smaller than the full bandwidth, subcarriers not part of the allocated subchannels are not excited when transmitting the preamble. If midambles are transmitted, they are identical to the short preamble or the second symbol of the long preamble that was used for the burst. Midambles are generally not necessary in fixed deployment, but the OFDM PHY

provisions them to enable extension to mobile applications as defined in IEEE Std 802.16e. The purpose of the short preamble and the second symbol of the long preamble is to allow fine frequency/time estimation and channel estimation.

In the UL, SSs can be ordered through the UL MAP to cyclicly rotate the preamble by a given factor between 0 and 255. The idea behind this is to allow better simultaneous detection of preambles by AAS-enabled BSs. This is achieved by the concept that, if preambles are sufficiently rotated, correlation of the received signal with the nonrotated preamble and with the rotated preamble will result in two better estimates due to low cross-correlation of the preambles. Because this shift must be substantially larger than the main multipath components of any of the channels of involved SSs, the number of shift possibilities is limited. For the UL, the BS has a receiving beam in the direction of the individual SSs, which suppresses signals from other simultaneously transmitting SSs with substantially different angles of arrival (AoAs). However, this feature is only marginally useful for SSs that have relatively similar AoAs. Despite the fact that the cyclic shift method is a mandatory specification, it is of little use for non-AAS devices.

FRAME STRUCTURE

The OFDM PHY supports two noncompatible frame structures: a mandatory one for PMP architectures and an optional one for mesh-based architectures. Either frame structure allows for a number of frame durations between 2.5 ms and 20 ms.

Point-to-multipoint (PMP)

For PMP architectures, the duplexing method may be either TDD or FDD (with H-FDD SSs supported), except in license-exempt bands where only TDD is provisioned to ensure better coexistence with other IEEE 802 standards.

The TDD structure for PMP architectures is shown in Figure 11–5. The FDD structure is similar, except that TTG and RTG are omitted and the DL and UL subframes exist in parallel on different frequencies.

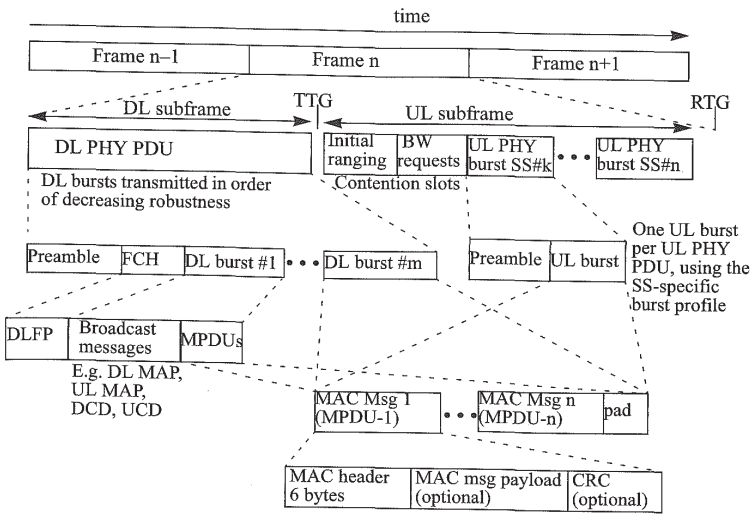


Figure 11–5: Example OFDM structure with TDD

The DL PHY PDU is initiated by the DL frame preamble, as previously described, followed by the FCH, which only contains the DLFP. The DLFP contains the lower 4 bits of the BS ID and the frame number (the full fields are present in the following DL MAP); this helps SSs with synchronization and identification of the right burst. The DLFP also contains a list of up to four DL bursts, indicating for each burst the DIUC (indicating, among other parameters, the modulation and coding methods), the existence of a preamble, and the length of each burst in OFDM symbols. For the first burst (Burst #0), the DIUC field is replaced by the Rate_ID, which indicates only the modulation and coding method. Lastly, the DLFP contains an HCS to identify errors in the FCH. It may also contain the start time of an STC-encoded region, if present, which lasts till the end of the DL frame. The FCH uses the most robust modulation method and coding (BPSK-1/2). The FCH hence allows the SS to individually demodulate each DL burst correctly.

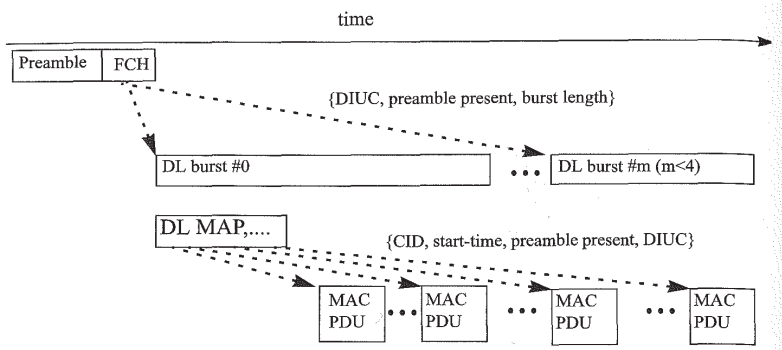


Figure 11–6: DL burst and MPDU mapping

The first DL burst subsequently contains the DL MAP, the UL MAP, the DCD, and the UCD. Transmission of each of the messages is optional (although insertion of DL MAP and DCD is mandated periodically to ensure that synchronization is maintained); but, when any of these four messages is present, the order is mandatory. The DL MAP may also contain markers to indicate that, from a certain starting point until the end of the DL subframes, STC, AAS, or DL subchannelization is enabled. These are the so-called *STC*, *AAS*, and *DL subchannelization zones*, respectively. If these zones are used to extend the cell range for devices, then all broadcast messages must be transmitted in these zones as well. Specifically, in the case of AAS, which has no broadcast PHY method, the FCH and unicast versions of the broadcast messages must be repeated on as many beams as necessary to reach all SSs. Similarly, in the UL, separate allocations must be inserted in these zones for the contention-based methods (ranging and BW requests).

The remainder of the first DL burst, as well as any subsequent DL burst within the frame, carries other MPDUs as indicated by the DL MAP.

The UL may contain contention slots for initial ranging and full and focused BW requests (explained later in this chapter). UL bursts may be received in parallel on nonoverlapping sets of subchannels (or on overlapping sets when exploiting the space division component through an antenna array). The main

difference between the allocation mapping on the DL and UL is that on the DL, the duration of a burst is implicit from the start time of the next burst, whereas on the UL, the duration is explicitly stated. This was introduced to avoid the need to derive the duration from multiple parallel subchannelized subsequent allocations (i.e, the duration of an allocation could be derived from the earliest subsequent start time of any other allocation in any of the subchannels assigned to the allocation).

Mesh

For the mesh architecture, only TDD is specified. Each frame consists of two configurable fixed-duration subframes. The first period is the control subframe, which is used for network and schedule control, whereas the second is used for data transfer.

As illustrated in Figure 11–7, network and schedule control are fully separated, in that one out every few control subframes is dedicated to network control, whereas the remaining control subframes are dedicated to scheduling. Each control subframe consists of a number of transmission slots.

The first slot of a network control subframe is dedicated to contention-based network entry. The message used for network entry (MSH-NENT) is substantially shorter than the duration of the slot. This provides robustness with respect to timing errors due to unknown propagation durations, similar to the technique discussed later in this chapter for initial ranging (and is only used in PMP architectures). The remainder of the slot in the network control subframe is used for contentionless distribution of network configuration parameters (including neighbor lists) by the devices already part of the mesh network.

Three different messages are used in the scheduling control subframe, which manage the scheduling of bursts in the data subframe. The MSH-CSCF message is used to distribute the connectivity tree for centralized scheduling, whereas the MSH-CSCH contains the actual centralized transmission schedule. MSH-CSCF and MSH-CSCH messages are not bound by slot boundaries. MSH-DSCH is used to establish distributed schedules. MSH-DSCH may also occur during the data subframe to establish contention-based schedules.

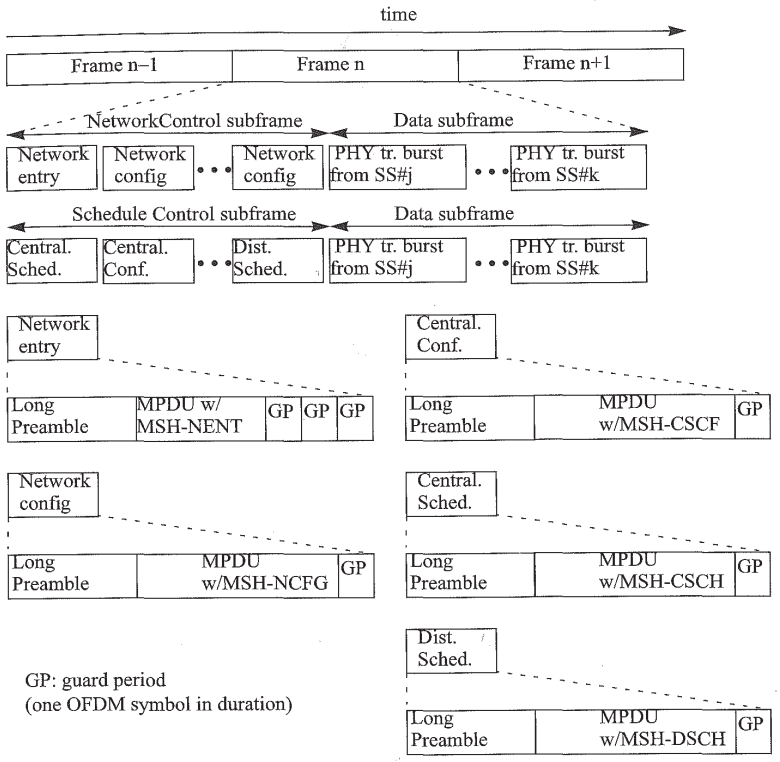


Figure 11–7: Mesh frame structure

Each control message is terminated by a guard period. This is because, although all devices are synchronized, propagation durations will still vary between devices. The additional symbol avoids the end of a burst received over a long link suffering interference by the start of a subsequent burst received over a short link. Relying only on the cyclic prefix of the OFDM symbol for this would not have been sufficient.

Each of the (broadcast) control message formats, not only the network entry message, starts with a long preamble. This is because a device may,

contention-free, conclude whether it is its time to transmit, but it cannot always determine who has the right to transmit otherwise. This is handled differently in the data subframe, where the transmission schedule is known. Although they use the long preamble by default, links can establish use of the short preamble in the data subframe.

The mesh architecture cannot make use of the subchannelization functionality. Although there is principally no obstruction to its use, the complexity of the scheduler would be prohibitive with current technology. The architecture does allow the use of patch antennas or antenna arrays, either with up to eight distinguishable directions. This allows for a substantially higher spacial reuse and hence substantial system capacity increases during the data subframe, at the cost of higher scheduling complexity. Extending link range through AAS as previously defined is, however, not possible due to the reliance on broadcasting for network control.

BW requesting is performed by use of the scheduling messages. The contention-based methods as described earlier are hence not used when using the mesh architecture.

CHANNEL ENCODING

The OFDM PHY channel encoding, which precedes the IFFT, consists of four steps; randomization, FEC, interleaving, and modulation.

Randomization

A specific sequence of data modulated on an OFDM symbol results in a given peak to average power ratio (PAPR) for that symbol. With a small probability of occurrence, this PAPR can be quite large (e.g., more than 7 dB). It is generally inefficient to back the transmitter power amplifier off far enough to avoid unrecoverable nonlinear distortion at the worst possible PAPR. Practical designs instead back off the amplifier only far enough to keep the probability of unrecoverable distortion acceptably low. However, retransmitting a data sequence with a high PAPR would result in a repeated receiver error. To avoid such persistent errors, the data are passed through a randomizer at the transmitter.

Randomization is performed by passing the block of data to be transmitted through a pseudo-random binary sequence (PRBS) with polynomial $1 + x^{14} + x^{15}$. For every data burst, the initial state of the PRBS memory is reset with a combination of the BS ID, the IUC for the allocation, and the frame number, except at the start of the frame, where it is reset with a fixed value, and on the first DL burst (i.e., the burst following the FCH), where it is not reset at all. Due to the fact that the FCH is fixed in duration, the state of the PRBS will always have the same value at the start of the first DL burst. Hence not resetting the PRBS at this point is equivalent to resetting it with a fixed value.

Forward error correction (FEC)

The OFDM PHY specified three methods of channel coding: Reed-Solomon concatenated with convolutional coding (RS-CC), block turbo coding (BTC), and convolutional turbo coding (CTC). RS-CC is mandatory, whereas BTC and CTC are optional. Although BTC and CTC generally provide on the order of 2 dB to 3 dB better coding gain, their substantial complexity made them less attractive as a mandatory requirement. The performance of BTC and CTC relative to each other varies slightly as a function of the coding rate and the size of the data to be encoded.

Reed-Solomon concatenated with convolutional coding (RS-CC)

The RS-CC encoding consist of a systemic RS(255,239,8) code over GF(2^8), which can be shortened and punctured to appropriate block sizes followed by a native rate 1/2 constraint-length 7 convolutional encoder (identical to the encoder defined in IEEE Std 802.11a), which can be punctured as well.

The randomized burst of data to be encoded is first appended by a byte of zeroes, as shown in Figure 11–8, in order to facilitate zero-tail termination of the CC encoder. Given the constraint length of 7, the usage of 6 tail bits would suffice (the last 2 bits will always remain zero after encoding). However, since data are generally passed to the PHY in whole bytes, the byte appendage is simpler. During development of the IEEE 802.16 standard, the use of tail-biting, which does not require the tail-byte overhead, in combination with per-OFDM symbol encoding, was favored. However, the disadvantage of having

to process a substantial portion of the data twice was deemed to outweigh the incurred zero-tailing overhead. A separate RS block is computed for each OFDM symbol. The parity bits for each block are sent to the CC encoder before the block itself, ensuring that the added tail byte is passed last to the CC encoder. The CC encoder is employed over the entire burst. The RS encoder is bypassed entirely when less than the entire channel is allocated (i.e., when eight or fewer subchannels are allocated). This avoids performance deterioration due to the weakness of RS codes for small blocks.

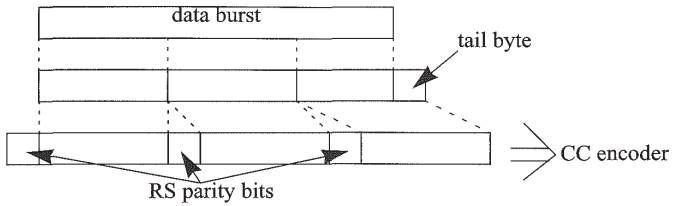


Figure 11–8: RS-CC encoding process

Block turbo codes (BTCs)

Unlike RS-CC, where the data are treated as a long string, BTC encodes the data in block form, where the block is filled row by row. Subsequently, a code polynomial is used to encode each row, with check bits added to the row. Subsequently, another code polynomial is used to encode each column, including the columns consisting of the row check bits. The polynomials are selected from a set of three extended Hamming codes and three parity check codes. The block (see Figure 11–9) may be shortened by not filling the first I_x rows, I_y columns, and B bits and zero-filling the Q bits until the next byte boundary. An encoding is hence fully described by its two constituent codes and the four parameters I_x, I_y, B, and Q. To specify blocks that fit into each of the five possible allocation sizes per OFDM symbol (see Figure 11–3), each of the four possible modulations and each of the two code rates per modulation—a total of thirty different encodings—are provided. The performance of BTC for an allocation of only a few subchannels was never thoroughly studied during the development of the subchannelization option. The performance may turn out to be problematic due to the small block sizes.

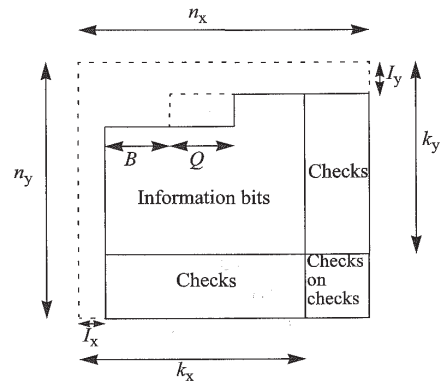


Figure 11–9: BTC encoding block

Convolutional turbo codes (CTCs)

The CTC encoder is most easily described from its depiction in Figure 11–10.

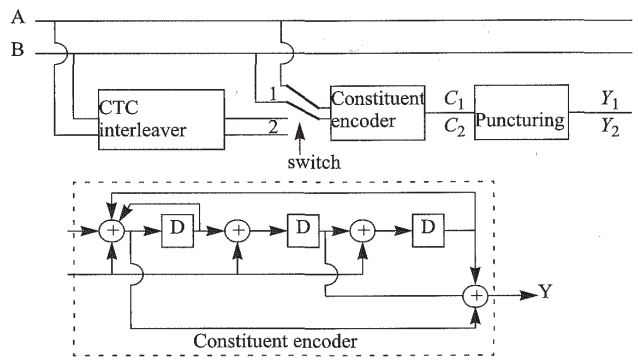


Figure 11–10: CTC encoder

Bits are alternately fed into A and B (MSB first), with the switch in position 1 as shown. The encoding process must be performed twice: one with the

memory elements initialized to 0b000 with the switch in position 1 and one with the memory elements initialized with a value computed by passing the final value of the first run through a lookup table and the switch in position 2.

The final encoded stream is achieved by the sequence of alternated bits from A and B, followed by all punctured parity bits from the first encoding (switch position 1), followed by all punctured parity bits from the second encoding (switch position 2).

Interleaving

The interleaver is defined by a two-step permutation. The first ensures that adjacent coded bits are mapped onto nonadjacent subcarriers. The second permutation ensures that adjacent coded bits are mapped alternately onto less or more significant bits of the constellation, thus avoiding long runs of low-reliability bits. The interleaver is virtually identical to that used in IEEE Std 802.11a and IEEE Std 802.11g™, except that it uses 12 as computational base, rather than 16. The reason for this is that each subchannel has 12 data subcarriers. A base-12 permutation hence ensures that adjacent bits are mapped not only to different subcarriers but to different subchannels as well (if multiple subchannels are allocated).

Block sizes for the interleaver are defined so that the number of bits in a block fits into all allocated subchannels within an OFDM symbol. There is hence exactly one interleaving block per OFDM symbol for each transmitter.

Modulation

The OFDM PHY mandates BPSK as well as Gray-mapped 4-QAM (QPSK), 16-QAM, and 64-QAM constellations. The Gray-mapped 64-QAM constellation is optional for license-exempt bands, mostly to allow the use of IEEE 802.11 RF components that would typically not meet the 64-QAM performance requirements of IEEE Std 802.16.

Pilot tones are modulated with a BPSK signal. The values to be used are derived from passing fixed initialization sequences (one each for UL and DL) through a PRBS with polynomial $x^{11} + x^9 + 1$ clocked with the OFDM symbol rate. The values of the individual pilots within a single OFDM symbol

are then defined as being either equal or the negative of the BPSK-modulated PRBS output.

CONTROL MECHANISMS

Of all control mechanisms that are available to the BS to control SS behavior, three have special provisions in the PHY: ranging, BW request, and power control.

Ranging

Ranging consists of two types: *initial* and *maintenance*. The first serves to initiate registration (or re-registration) of an SS with a BS. The second serves to periodically readjust PHY parameters of the SS and to verify that the SS is still on line.

Maintenance ranging is predominantly a MAC process, for which the PHY makes no special provisions in terms of scheduling or waveform creation. From the PHY's perspective, maintenance ranging is treated like any other data exchange.

Initial ranging is different in this respect in that it is, by necessity, a contention-based protocol with a nonzero probability of collision. The time period scheduled for initial ranging must be at least the total duration of the round-trip propagation delay and the transmission time of the initial ranging message, as a new SS is not yet able to correct its clock for propagation delays. The round-trip propagation delay is a function of the maximum cell size. An initial ranging attempt uses the long preamble. The initial ranging data fit in one OFDM data symbol (using the most robust modulation/coding) so that the total duration of an initial ranging allocation must be at least $3 + \lceil (n \cdot BW)/(256(1 + T_g/T_b)) \cdot (2r)/c \rceil$ OFDM symbols, where r is the cell radius in meters and c the speed of light. For typical deployments, this turns out to be a four-OFDM-symbol allocation. In fixed systems, contrary to mobile applications, registration is a relatively rare occurrence. It is, therefore, not necessary to facilitate initial ranging slots in every frame, and, as a result, the incurred overhead is negligible. The different initial ranging formats are shown in Figure 11–11.

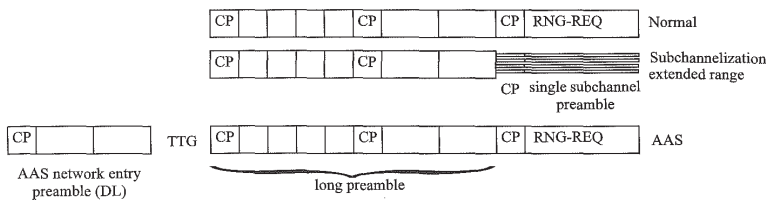


Figure 11–11: Initial ranging formats

As mentioned before, the PHY provides special provisions to allow SSs with insufficient link budget to use less than the full bandwidth. As a consequence of this design, a method is specified for these stations to start initial ranging with a slightly different method than the one described above. When an SS deduces, from estimating the required transmit power (for which the BS provides the necessary data in the UCD message), that it will not be able to close the link or when it has failed initial ranging attempts multiple times, it may replace the RNG-REQ message with the subchannelization preamble on one random subchannel so that the total initial ranging burst consists of the long preamble followed by the subchannelized preamble. The long preamble has sufficient inherent gain to be detected, whereas only an energy detect is needed to establish on which subchannel the subchannelization preamble was sent. Because the same total power is used, the spectral density of the single subchannel preamble will be roughly 12 dB higher, further improving the accuracy of the energy detect. Upon detection, the BS can issue a subchannelized UL allocation referencing the SS by the ranging slot and subchannel. For efficiency, SSs using the full bandwidth as well as those requiring a subchannelized initial ranging are using the same ranging opportunities. This avoids duplication of allocations.

An exception to this exists when the BS operates using AAS. Under the assumption that the cell range for AAS devices may be much larger than that of non-AAS devices, it cannot be assumed that AAS-enabled SSs can receive the broadcast DL MAP. They may, therefore, have no knowledge of the UL schedule, including initial ranging slots. To get around this problem, a BS must schedule an AAS initial ranging slot of eight OFDM symbols, the first symbol of which the BS uses to transmit the AAS network entry preamble,

which has sufficient gain to be detected by AAS-enabled SSs. Following a TTG, an AAS-enabled SS may then transmit the long preamble and RNG-REQ message. The sum of SSRTG and round-trip delay may hence not exceed four OFDM symbols.

One of the few changes provided by the IEEE 802.16e amendment for the OFDM PHY is, as mentioned previously, the capability of the BS to allocate an SS-specific (and hence noncollision-based) initial ranging opportunity. This feature is useful when the BS receives notice (for example, from a network controller) that a MS will be attempting a handover to the BS in question. This shortens the handover time and avoids the possibility that many collision-based initial ranging opportunities need to be allocated because of handover frequency.

BW requests

There are multiple ways in which bandwidth can be requested: through polling, through piggybacking, and through two contention-based methods. From the PHY perspective, only the latter two are of interest. These two methods, of which only the first is mandatory, are *full contention* and *focused contention*.

Full contention may be used both with full bandwidth allocations and with subchannelized allocations. The only difference is that, using full bandwidth, the request is limited to two OFDM symbols to be filled with the short preamble and one data symbol containing the BW request, whereas using subchannelization, the width (in subchannels) and length (in OFDM symbols) is announced in the UCD. The BS may provide multiple request opportunities within a single allocation. The BS can identify the SS by the Basic CID in the BW request.

Focused contention is an optional feature. To request bandwidth using focused contention, an SS selects randomly a contention channel index out of the 48 available. This maps to four unique data subcarriers. These are arranged to form three sets of contention channels within each subchannel. On these four subcarriers, the SS transmits one of eight BPSK contention codes. On the following OFDM symbol, the SS again transmits the same contention code on the same subcarriers, with the difference that, for all subcarriers with

contention element –1, the phase is inverted. The BS is now able to identify the SS by the contention channel index and contention code and can issue an UL allocation on that basis. The advantage of this scheme is that many SSs can simultaneously indicate a need for bandwidth with a low probability of collision (same contention channel) and an even lower probability of undetectable collisions (same contention channel and same contention code). A minor drawback of this scheme is that it introduces a single frame of delay, as the focused contention indication does not provide the BS with any knowledge about the size of the bandwidth needed. When a BS supports subchannelization, it may indicate through the UCD that a certain amount of the contention channel indices are reserved to indicate the SS needs a subchannelized allocation. In the case that it receives one of these codes, the BS does not know the identity of the requesting SS; therefore, it has no choice but to allocate the lowest number of subchannels required by any of its associated SSs.

Power control

Power control is needed and specified only in the UL, where it serves to ensure that the received signal power density at the BS falls within a narrow range. The SS transmit power level is adjusted under control of the BS. To allow this, each SS reports, during initialization, its current transmit power level and the maximum transmit power level for each of the modulations as part of its basic capabilities. This allows the BS to compute the feasibility of an SS changing its power level for full bandwidth allocations with the current modulation. It also allows the BS to compute the effects of a change in modulation and subchannelization. This can be used to improve bandwidth efficiency.

The BS can use both the maintenance ranging and fast power control MAC messages to force an SS to adjust its power levels. The SS must be capable of changing power at a rate of at least 30 dB/s.