FullMAX Latency

Version 1.0 - February 25th, 2016

By Menashe Shahar, CTO, Full Spectrum Inc.

## General

1. The FullMAX end to end latency in the uplink and in the downlink directions is determined by the system configuration and by the traffic characteristics:
	1. The main system configuration parameters effecting the latency are the frame duration and the QOS configuration used.
	2. The traffic characteristics effecting the latency include:
		1. The load vs capacity in the common channel used in the sector
		2. Packet size distribution
		3. TCP vs UDP

## Frame Duration and Scheduling Type

1. For a given traffic scenario and QOS configuration, the downlink and uplink end to end latency is determined by the Time Division Duplex (TDD) frame duration. The following scenario may serve as a latency reference model:
	1. The channel is lightly loaded
	2. The Service Data Units (SDUs) received over the Ethernet interface are encapsulated un-fragmented within a single 802.16e Protocol Data Unit (PDU) which is then transmitted within a single downlink/uplink sub-frame. The maximum size of the SDU received over the Ethernet interface is 1518 bytes. This translates into an 802.16e PDU size of 1528 bytes. [[1]](#footnote-1)
2. The downlink end to end latency in the above scenario equals two TDD frames. This includes the latency of one frame carrying the SDU plus a worst case latency of a second frame, depending on the time of arrival of the SDU at the Base Station relative to the time at which the scheduling of the next frame was completed.
3. The uplink end to end latency depends on the scheduling type (this is configured per Remote and per service flow) as follows:
	1. **Unsolicited Grant Service (UGS) scheduling** (i.e., the respective Remote gets an unsolicited bandwidth allocation at every uplink sub-frame):

The end to end uplink latency equals two frames. This includes the latency of one frame carrying the SDU plus a worst case latency of a second frame, depending on the time of arrival of the SDU at the Remote relative to the time at which the allocation for the next frame was received.

* 1. **Enhanced Real Time Polling Service (ertPs) scheduling** (i.e., the respective remote gets an unsolicited bandwidth allocation at every frame while the connection is active. This is similar to UGS except the connection is started and stopped dynamically by triggering signals, e.g., on hook/off hook in the case of VoIP):

The end to end uplink latency is the same as for UGS while the connection is active.

* 1. **Real Time Polling Service (rtPs) scheduling (i.e.,** the Base Station sends each Remote a unicast opportunity to transmit a bandwidth request at every frame):

The end to end uplink latency is 4 frames. In this case, one additional frame is used by the Remote to send a bandwidth request to the Base Station and a second additional frame is used by the Base Station to send a bandwidth allocation to the respective Remote.

* 1. **Non Real Time Service Scheduling** (i.e., the Base Station send each Remote, unicast opportunities to transmit bandwidth requests but the period at which these opportunities are allocated is configurable and not necessarily at every frame):

The end to end latency in this case equals the latency for rtPs plus the duration of unicast bandwidth request allocation.

* 1. **Best Effort (BE) Scheduling** (i.e., the Base Station send each Remote, broadcast opportunities to transmit bandwidth requests. The period at which these opportunities are allocated is configurable and not necessarily at every frame):

The end to end latency in this case equals the latency for rtPs plus the duration of the configurable broadcast bandwidth request allocation period.

1. FullMAX Software Definable Radio (SDR) architecture offers extensive flexibility in configuring the TDD frame structure and the frame duration. The frame durations currently offered are 5 ms, 10 ms, 12.5 ms, 20 ms, 25 ms, 40 ms and 50 ms. For example, if a 12.5 ms frame duration is used, the end to end latency for a lightly loaded network is 25 ms in downlink and 25 ms in uplink if UGS scheduling is used. The uplink latency increases to 50 ms if rtPs scheduling is used.
2. A congestion scenario is created when the offered traffic load is above the capacity of the sector in the downlink or in the uplink direction or both. SDUs received in the congested direction are buffered and the latency increases relative to the lightly loaded scenario. QOS parameters are used to classify the incoming traffic into multiple service flows and prioritize certain flows over other flows depending on their latency tolerance. Prioritization is done by using service classes which are assigned to each service flow. They include a list of QOS parameters including scheduling type and priority within the same scheduling type:
	1. Prioritization through scheduling type: Scheduling type are prioritized according to the following order:
		1. UGS
		2. rtPs
		3. nrtPs
		4. BE
	2. Prioritization within the same scheduling type: FullMAX offers 7 levels of priorities. The higher the value, the higher the priority.
3. TCP flow control impact on latency: The TCP Ack based flow control mechanism may introduce extra delay in one direction due to congestion in the reverse direction. This is addressed by assigning high priority to the TCP Ack packets which according to our experience, eliminates the extra flow control delay.

## Frame Duration Optimization

1. Latency optimization is accomplished by configuring the frame duration such that the bulk of the packets do not need to be fragmented, i.e., the capacity of the downlink and uplink sub-frames is sufficient to accommodate the un- fragmented packets. The capacity of the downlink and uplink sub-frame depends on:
	1. **Symbol rate**: this depends on the channel size and on the permutation used. For example, FullMAX configuration for 1 MHz wide channel is 2 X 3 band AMC (i.e., 2 bins X 3 OFDMA symbols per sub-channels) with a total of 108 out of 128 active subcarriers. The symbol rate is 1.12 MHz and the effective used bandwidth is 945 KHz.
	2. **Number of sub-channels per sector**: The higher the number, the higher the capacity. For example, if band AMC is used, up to 6 sub-channels can be configured per sector. The actual number depends on frequency reuse and link budget considerations.
	3. **Number of symbols in the downlink and uplink sub-frames**
2. Example: proposed configuration for the Upper 700 MHz A block:
	1. The Upper 700 MHz A block, offers 2 X 1 MHz wide channels, one centered at 757.5 MHz and the other centered at 787.5 MHz. Within each of the 1 MHz channels, FullMAX has 12 band AMC sub-channels in the uplink and in the downlink, i.e., a total of 12 independent sub-channels in each direction. A non-aggressive re-use factor of 4 is proposed which leads to the usage of 3 out of 6 sub-channels in each sector in both the uplink and the downlink direction.
	2. Assuming a reverse asymmetrical traffic with maximal length packets in the uplink and relatively short packets in the downlink, we can use the following TDD frame configuration:
		1. Frame duration = 12.5 ms
		2. Number of symbols in downlink sub-frame: 21
		3. Number of symbols in uplink sub-frame: 69
		4. Uplink sub-frame capacity @ FEC Code = 5: 1560 bytes
		5. Downlink sub-frame capacity @ FEC Code = 5: 360 bytes.
		6. Required minimum CINR for FEC Code 5 = 18 dB
		7. Uplink scheduling type: rtPs

In this example, the end to end latency under light traffic conditions is 25 ms in downlink and 50 ms in uplink.

1. For general purpose Internet traffic, significant percentage of the packets are 1500 byte long. Fragmentation of these packets does not reduce latency but increases overhead so it is better to configure the frame size to accommodate the maximal length SDU. If however the traffic is known to consist primarily of short frames, it makes sense to reduce the frame size to accommodate the typical frame size. The capacity of the frame depends on the channel size, on the number of sub-channels used in the sector and on the typical FEC Code used. [↑](#footnote-ref-1)