



# LATENCIA EN SISTEMAS ENTRAMADOS aerDOCSIS

Este documento trata de explicar el origen de los distintos valores de latencia que se obtienen en sistemas entramados y en sistemas a ráfagas.

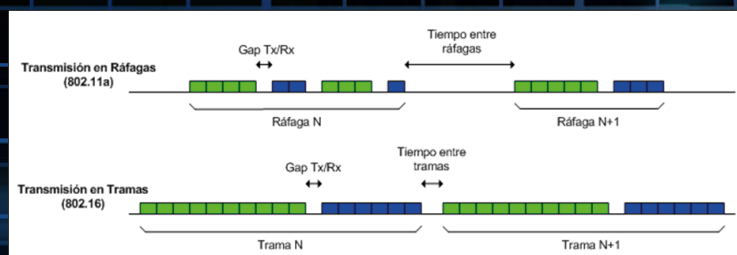
Se pretende explicar la relación que tienen la latencia round-trip en la tecnología aerDOCSIS con la duración de trama y el número de usuarios de la celda, para que el lector pueda configurar un sistema basado en esta tecnología de la mejor forma posible según cada escenario.

# LATENCIA EN SISTEMAS ENTRAMADOS

## SISTEMAS A RÁFAGAS VS ENTRAMADOS

En tecnologías como IEEE 802.11 (Wi-Fi) que tienen un Acceso al Medio aleatorio, las transmisiones se realizan paquete a paquete (a ráfagas). No existe ningún elemento que regule las transmisiones con lo que cada usuario transmite “cuando quiere”. Esto provoca que en determinados escenarios (distancias cortas, un único usuario conectado, poco tráfico total en el aire) esta tecnología puede conseguir latencias RTT realmente pequeñas (5ms o inferiores) y con las que un sistema entramado no puede competir. El problema viene cuando aumenta el número de usuarios o el tráfico neto de la red, ya que empiezan a aparecer colisiones y retransmisiones. En este escenario las latencias se disparan y son variables e impredecibles.

La tecnología aerDOCSIS, al igual que otras tecnologías inalámbricas modernas de última generación, como 5G, es un sistema entramado y por tanto transmite trama a trama. En la banda libre de 5GHz, aerDOCSIS usa el modelo TDD (Time Duplexion Division) en el que la trama aerDOCSIS se divide en el tiempo en 2 subtramas, una para el tráfico descendente (Downlink) y otra para el ascendente (Uplink). El acceso al medio se realiza sin contienda y la BS ejerce de árbitro asignando time slots a los diferentes usuarios conectados (SSs) y controlando por tanto todas las transmisiones al medio radio. Al ser un sistema determinista, la latencia mínima es algo superior que en sistemas a ráfagas, pero es constante y se puede predecir.



Las ventajas de sistemas Entramados frente a sistemas basados en Ráfagas son numerosas en redes inalámbricas Punto-Multipunto (PtMP), ya que permiten la máxima utilización del medio físico debido a su alta eficiencia espectral, aunque suelen estar caracterizados por una mayor latencia mínima que sistemas basados en ráfagas. Esto es debido a que a la hora de medir la latencia “round-trip” debe emplearse el canal en ambos sentidos (de BS a SS y de SS a BS), y esto es imposible que ocurra en una misma trama. Además hay que tener en cuenta otros factores como posibles búferes o que la BS permita la transmisión en la trama actual o en la siguiente. Lo que es evidente es que la latencia “round-trip” total siempre será un múltiplo de la longitud de trama.

# ¿CÓMO OPTIMIZAR LA LATENCIA EN aerDOCSIS?

Por la naturaleza del propio sistema, la única forma de reducir la latencia en sistemas entramados es mediante el empleo de tramas de mínima duración. A menor tiempo de trama, menor latencia. Este mecanismo de reducción de latencia tiene serias implicaciones:

- Dado que la información de control de trama y otro tipo de overhead suele ser constante independientemente de la longitud de trama, su peso relativo aumenta al reducir el tiempo de trama. Es decir, el throughput neto agregado del sistema se reduce al usar tramas más cortas.

Las tramas de corta duración permiten el acceso a un menor número de usuarios por trama. En tramas largas es posible asignar slots de transmisión a un gran número de usuarios en una misma trama, pero si

- la duración de la trama se reduce, el número de usuarios que pueden participar en una única trama desciende. Si esto ocurre, es posible que algunos usuarios tengan que esperar a la siguiente trama para transmitir, por lo que su latencia aumenta. Es decir, tramas cortas permiten baja latencia, pero a un número bajo de usuarios. Si el número de usuarios aumenta, la latencia también aumenta.

Este aumento de la latencia ocurre para cualquier longitud de trama. Si la trama es de larga duración, la latencia es alta, y si crece el número de usuarios es posible que no haya slots suficientes en una única trama para todos, por lo que la latencia crece. Sin embargo, este punto de inflexión ocurre para un mayor número de usuarios que en el caso de tramas cortas.

## EJEMPLO PRÁCTICO (I)

Vamos a ver este efecto con un ejemplo. Supongamos un sistema entramado en el que los slots de transmisión duran  $25\mu\text{s}$  (independiente de la duración de trama) y que el mínimo número de slots que se pueden asignar a un usuario en una trama concreta es dos (uno para control y otro para carga útil).

Analizaremos la latencia en dos casos: para una trama “corta” de  $2.5\text{ms}$ , y para una trama “larga” de  $10\text{ms}$ . En ambos casos se asignará el 50% de la trama a tráfico descendente y el 50% a tráfico ascendente.

Supondremos que la latencia es igual al triple del tiempo de trama, es decir,  $7.5\text{ms}$  en trama corta y  $30\text{ms}$  en trama larga. A continuación se calculará la evolución de la latencia para dos situaciones distintas: sin considerar el overhead de la trama y teniéndolo en cuenta.



# EJEMPLO PRÁCTICO (II)

## CASO A:

### Considerando overhead de trama

En este caso, la trama corta dispondría de 100 slots y la larga 400 slots, todos ellos utilizables, ya que no hay sobrecarga de trama. Con una división de trama del 50% y que cada usuario como mínimo transmite dos símbolos, una única trama corta permite transmitir a 25 usuarios, mientras que la larga 100.

¿Qué ocurre si hay 50 usuarios en el caso de la trama corta? En este caso es imposible que todos transmitan en una única trama, por lo que serán necesarias dos tramas para transmitir la información correspondiente a todos los usuarios: la latencia se duplica. ¿Y en el caso de trama larga? La latencia sigue siendo la original, ya que todos los usuarios pueden transmitir en una trama.

¿Y si el número de usuarios sube a 100? La trama corta necesita cuadruplicar su latencia, mientras que la trama larga está al límite de usuarios. Lo más interesante es que en este caso ambas tramas presentan la misma latencia. Para más de 100 usuarios, ambas tramas presentan exactamente la misma latencia.

## CASO B:

### Considerando overhead de trama

Ahora supondremos que el overhead de trama es de ocho slots por trama, independientemente de la duración de la trama. En este caso los slots disponibles por trama para transmisión de datos son 92 en la trama corta y 392 en trama larga. Teniendo en cuenta al 50% de tráfico ascendente y el mínimo de dos slots por usuario, la trama corta permite un número máximo de 23 usuarios en una misma trama, mientras que la trama larga permite 98 usuarios.

En este caso para un número elevado de usuarios, la latencia en ambos casos crece linealmente con el número de usuarios, pero debido al overhead de trama, la latencia es ligeramente mayor en el caso de trama corta.

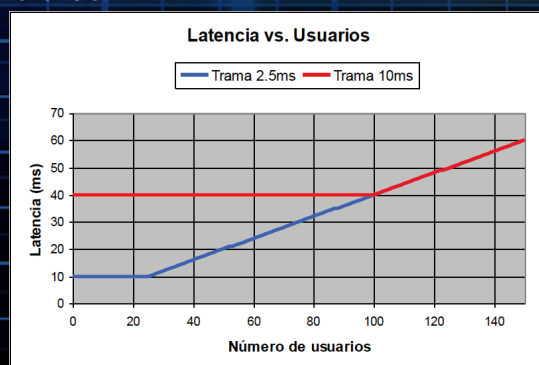


Figura 1 - Relación entre latencia y numero de usuarios

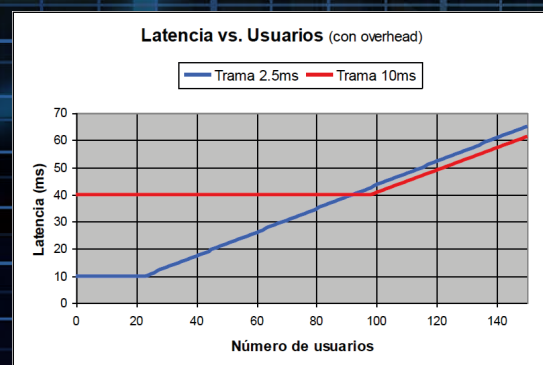


Figura 2 - Relación entre latencia y numero de usuarios, con sobrecarga

# CONCLUSIONES

En sistemas entramados la latencia es proporcional a la longitud de trama, por lo que una técnica para reducir la latencia es minimizar el tiempo de trama. Sin embargo, hay que tener en cuenta tres puntos importantes:

- El throughput total agregado disminuirá debido al mayor peso del overhead de trama
- La latencia es baja, pero sólo puede mantenerse baja para un cierto número de usuarios. Cuanto más larga es la trama, mayor es la latencia mínima, pero puede mantenerse para un mayor número de usuarios
- Si el número de usuarios es alto, la latencia puede ser mayor en el caso de trama corta debido al overhead de trama

En base a estas conclusiones se concluye que a la hora de establecer el objetivo de latencia de la red es necesario tener en cuenta el número de usuarios. Si el número de usuarios es bajo, la latencia puede reducirse sin riesgo, pero si el número de usuarios es alto, no se recomienda reducir demasiado la longitud de trama, ya que la latencia puede incrementarse y en cualquier caso el throughput neto se reduce notablemente. Es mejor emplear tramas largas, que ofrecen mayor throughput y mejor latencia en caso de alto número de usuarios.