DOI: https://doi.org/10.46842/ipn.cien.v24n1a03

Científica, vol. 24, núm. 1, pp. 23-32, enero-junio 2020.
ISSN 2594-2921, Instituto Politécnico Nacional México

Análisis del impacto del intervalo de tiempo de transmisión sobre la latencia en la red de acceso radio de sistemas 5G

Andrés **Castro-Delgado**
Víctor **Quintero-Flórez**

Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
COLOMBIA

correos electrónicos (emails):
sbox381@gmail.com
vflorez@unicauca.edu.co

Recibido 25-07-2019, aceptado 02-10-2019.

Resumen

Los Sistemas de Comunicaciones Móviles de Quinta Generación (5G, Fifth Generation) soportarán servicios de Comunicación Ultra Confiable de Baja Latencia (URLLC, Ultra-Reliable Low-Latency Communication), que imponen requerimientos exigentes en términos de latencia y confiabilidad. La Unión Internacional de Telecomunicaciones (ITU, International Telecommunications Union) ha considerado reestructurar la trama de datos de nivel físico en Sistemas 5G para adaptar el Tiempo de Intervalo de Transmisión (TTI, Time Transmission Interval) y cumplir con los requerimientos definidos para los servicios URRLC. El presente artículo analiza el impacto del TTI sobre la latencia de la Red de Acceso Radio (RAN, Radio Access Network) de Sistemas 5G. Para esto, se estudia la estructura de trama de nivel físico de Sistemas 5G y se diseñan escenarios de simulación teniendo en cuenta la latencia en el nivel físico. Estos escenarios son implementados en la herramienta software ns3, utilizando el módulo mmWave para sistemas 5G. Los resultados muestran que el TTI influye significativamente sobre la latencia en la RAN, sin embargo, es necesario considerar las características de la red y los servicios a desplegar.

Palabras clave: 5G, latencia, TTI, numerologías, nueva radio.

Abstract
(Analysis of the Transmission Time Interval Impact on the 5G Radio Access Network Latency)

5G Communication Systems are expected to support URLLC services, which impose strict latency and reliability requirements. The International Telecommunication Union (ITU) has consider to restructure the 5G physical frame to adapt Time Transmission Interval (TTI) according to services requirements. This paper analyze the impact of the TTI on the 5G Radio Access Network (RAN) Latency. To achieve this, the 5G physical frame is studied and scenarios with different parameters are designed according to the physical level latency. These scenarios are implemented in the ns3 simulation tool, using the mmWave module for 5G systems. Results show that TTI influences significantly on the RAN latency, however it is necessary to consider network characteristics and services.

Index terms: 5G, latency, TTI, numerology, new radio.

Lista de acrónimos:

3GPP, *3rd Generation Partnership Project*, Proyecto de Cooperación para Sistemas de Tercera Generación.
5G, *Fifth Generation*, Quinta Generación.
64 QAM 64, *Quadrature Amplitude Modulation*, Modulación de Amplitud y Cuadratura de 64 estados.
BS, *Base Station*, Estación Base.
CN, *Core Network*, Núcleo de Red.
CP, *Cyclic Prefix*, Prefijo Cíclico.
CP, *Control Plane*, Plano de Control.
CP-OFDM, *Cyclic Prefix Orthogonal Frequency-Division Multiplexing*, Multiplexación por División de Frecuencias Ortogonales con Prefijo Cíclico.
DL, *Downlink*, Enlace de Bajada.
E2E, *End to End*, Extremo a Extremo.
gNB, *Next Generation Node B*, Nodo B de próxima generación.
GPL, *General Public License*, Licencia Pública General.
HARQ, *Hybrid Automatic Repeat Request*, Solicitud de Retransmisión Automática Híbrida.
IMT-2020, *International Mobile Telecommunications for 2020 and beyond*, Telecomunicaciones Móviles Internacionales para el 2020 y futuro.
ITU, *International Telecommunications Unio*n, Unión Internacional de Telecomunicaciones.

ITU-R, *ITU Radio Section*, Sección Radiocomunicaciones de la ITU.
LDPC, *Low Density Parity Check*, Verificación de Paridad de Baja Densidad.
MAC, *Medium Access Control*, Control de Acceso al Medio
MME, *Mobility Management Entity*, Entidad de Gestión de Movilidad.
NR, *New Radio*, Nueva Radio.
OFDM, *Orthogonal Frequency-Division Multiplexing*, Multiplexación por División de Frecuencias Ortogonales.
OSI, *Open System Interconnection*, Interconexión de Sistemas Abiertos.
PDCP, *Packet Data Convergence Protocol*, Protocolo de Convergencia de Paquetes de Datos.
PDU, *Protocol Data Unit*, Unidad de Datos de Protocolo.
PGW, *Packet Data Network Gateway*, Pasarela de Red de Paquetes de Datos
PHY, *Physical Layer*, Nivel Físico.
QoS, *Quality of Service*, Calidad del Servicio.
R15, *Release 15*, Especificación 15.
R16, *Release 16*, Especificación 16.
RAN, *Radio Access Network*, Red de Acceso Radio.
RLC, *Radio Link Control*, Control del Enlace Radio.
SDAP, *Service Data Adaptation Protocol*, Protocolo de Adaptación de Servicio de Datos.
SGW, *Serving Gateway*, Pasarela de Servicio.
SR, *Scheduling Request*, Solicitud de Asignación de Recursos.
TDD, *Time Division Duplexing*, Duplexación por División de Tiempo.
TIC, Tecnologías de Información y Comunicación.
TTI, *Time Transmission Interval*, Intervalo de Tiempo de Transmisión.
UE, *User Equipment*, Terminal de Usuario.
UG, *Uplink Grant*, Mensaje de Concesión.
UL, *Uplink*, Enlace de Subida.
UP, *User Plane*, Plano de Usuario.
URLLC, *Ultra-Reliable Low-Latency Communication*, Comunicación Ultra Confiable de Baja Latencia.

1. Introducción

Las tecnologías que harán parte de los sistemas de comunicación móvil 5G, han sido ampliamente investigadas desde hace más de una década para proveer soluciones a los requerimientos y necesidades de usuarios y empresas, en la prestación de servicios de telecomunicaciones a partir del año 2020 [1]. Aplicaciones emergentes tales como Internet táctil, reproducción de video de alta definición, telemedicina, telecirugía, transporte inteligente y conducción autónoma impondrán requerimientos exigentes de latencia que las redes móviles actuales no pueden soportar [2]. La calidad de servicio de dichas aplicaciones dependerá de conexiones inalámbricas que garanticen una latencia consistente no mayor a 1ms, por lo que es necesario realizar cambios importantes en la arquitectura de las redes móviles actuales [3].

La ITU está trabajando en el desarrollo de la especificación de las tecnologías para sistemas 5G bajo la norma de Telecomunicaciones Móviles Internacionales para el año 2020 y futuro (IMT-2020, International Mobile Telecommunications for 2020 and beyond). El marco de trabajo y visión de dicho proyecto están plasmados en la recomendación ITU-R M.2083-0, la cual establece que uno de los objetivos principales es proveer comunicaciones con una latencia cercana a cero [4]. Para lograr esto, la ITU ha considerado modificar la arquitectura de red e implementar tecnologías de vanguardia que permitan la prestación de nuevos servicios.

En el nivel físico, la ITU ha considerado reestructurar la trama de datos para soportar diferentes escenarios de aplicación. Este cambio permitirá adaptar el TTI según los requerimientos de los servicios y el estado general de la red, haciendo un mejor uso de los recursos radio y por consecuencia reduciendo la latencia. El presente artículo estudia los diferentes parámetros de la estructura de trama de datos que definen el TTI y su impacto sobre la latencia en la RAN 5G.

2. Marco teórico

El Proyecto de Cooperación para Sistemas de Tercera Generación (3GPP, 3rd Generation Partnership Project) está desarrollando un estándar global para la tecnología de acceso radio 5G denominado Nueva Radio (NR, New Radio), el cual operará desde la frecuencia de 1 GHz hasta 100 GHz. La primera versión del estándar NR fue completada a mediados del 2018 y se definió bajo el nombre de Especificación 15 (R15, Release 15). La segunda versión del estándar, definida como Especificación 16 (R16, Release 16), se encuentra en desarrollo y será completada a finales del 2019 [5].

2.1. Pila de protocolos NR

La pila de protocolos definida en R15 se compone de los siguientes niveles y subniveles: Nivel Físico (PHY, *Physical Layer*), Subnivel de Control de Acceso al Medio (MAC, *Medium Access Control*), Subnivel de Control del Enlace Radio (RLC, *Radio Link Control*), Subnivel de Protocolo de Convergencia de Paquetes de Datos (PDCP, *Packet Data Convergence Protocol*) y Subnivel del Protocolo de Adaptación de Datos de Servicio (SDAP, *Service Data Adaptation Protocol*), como se muestra en la Fig. 1 [6].

En este artículo se denomina trama de nivel físico a la Unidad de Protocolo de Datos (PDU, Protocol Data Unit) de dicho nivel y

Científica, vol. 24, núm. 1, pp. 23-32, enero-junio 2020, ISSN 2594-2921, Instituto Politécnico Nacional, México.
DOI: https://doi.org/10.46842/ipn.cien.v24n1a03

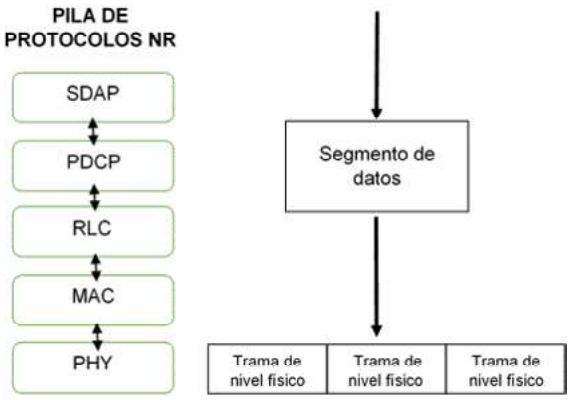

Fig. 1. Pila de Protocolos NR en el UP.

segmento de datos a la información transportada por niveles superiores que se entrega al nivel físico para su posterior transmisión por el canal radio, tal como se muestra en la Fig. 1.

2.2. Forma de onda

NR emplea Multiplexación por División de Frecuencias Ortogonales con Prefijo Cíclico (CP-OFDM, Cyclic Prefix Orthogonal Frequency-Division Multiplexing) tanto en el enlace de bajada (DL, Downlink) como de subida (UL, Uplink). En esta forma de onda, la separación entre subportadoras (Δf) asegura su ortogonalidad (véase Fig. 2), evitando la interferencia entre subportadoras y la necesidad de bandas de guarda o filtros pasabanda complejos. Así mismo, se inserta un prefijo cíclico en la señal de transmisión para evitar los efectos de la interferencia intersimbólica [6].

2.3. Numerologías

NR define diferentes valores para la duración del símbolo y el espaciamiento entre subportadoras en OFDM. Esta definición se conoce como numerología. En la primera versión de NR, 3GPP define cinco valores para el espaciamiento entre subportadoras: 15 KHz, 30 KHz, 60 KHz, 120 KHz y 240 KHz. Esta característica permitirá el soporte de diversos servicios con diferentes requerimientos [6].

2.4. Trama de nivel físico

En el dominio del tiempo, la transmisión está organizada en tramas con una duración de 10 ms, cada trama es dividida en 10 subtramas de 1 ms de duración cada una. Una subtrama es dividida a su vez en TTI compuestos por 14 símbolos OFDM. La duración de cada TTI depende de la numerología implementada como se muestra en la Tabla 1.

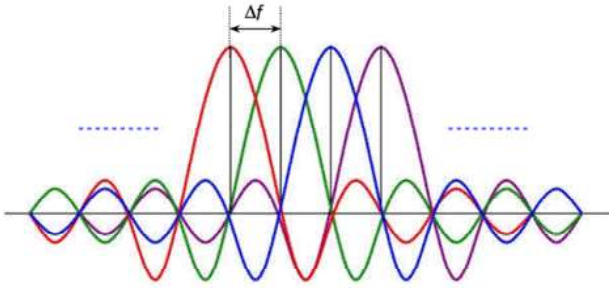

Fig. 2. Señal OFDM.

En NR es posible usar TTI con un número menor de 14 símbolos, dichas estructuras se definen como mini-TTI. En la primera versión de NR se definen mini-TTI con duración de 2, 4 y 7 símbolos OFDM (véase Fig. 3) [7].

2.5. Latencia en redes celulares

La latencia en redes celulares puede ser clasificada en dos grupos: latencia en el Plano de Usuario (UP, User Plane), de-

Tabla 1. Duración del TTI, símbolo y CP de acuerdo a diferentes numerologías.

Numerología	0	1	2	3	4
Espaciamiento entre subportadoras (KHz)	15	30	60	120	240
TTI (ms)	1	0.5	0.25	0.1	0.06
Duración del símbolo OFDM (μs)	66.67	33.33	16.67	8.3	4.17
Duración del CP (μs)	4.69	2.34	1.17	0.5	0.29
Duración Símbolo OFDM y CP (μs)	71.35	35.68	17.84	8.9	4.46

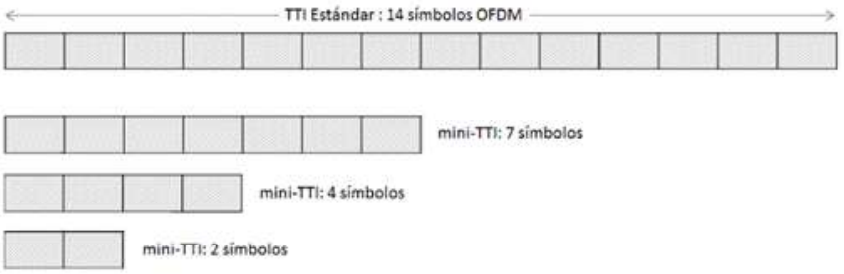

Fig. 3. Mini-TTI de 7, 4 y 2 símbolos OFDM.

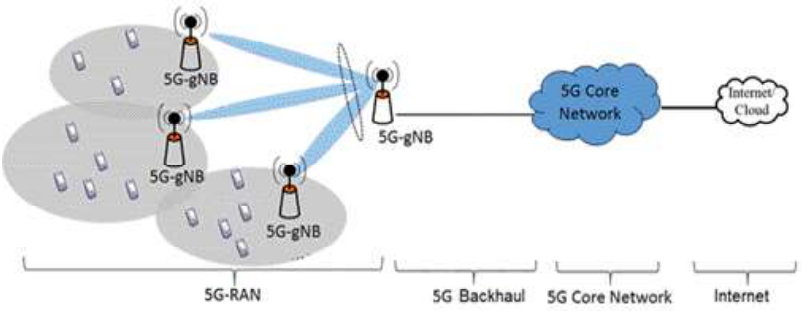

Fig. 4. Arquitectura de red de un sistema de comunicaciones móviles.

finida como el tiempo de transmisión entre la Estación Base (BS, Base Station) y el Terminal de Usuario (UE, User Equipment); y latencia en el Plano de Control (CP, Control Plane), definida como el tiempo de transición entre el estado inactivo y activo de un terminal [8].

La latencia en redes celulares puede originarse en los distintos componentes de la arquitectura del sistema: la RAN, Red de Transporte (Backhaul), el Núcleo de Red (CN, Core Network) e Internet [9]. Dichos componentes son ilustrados en la Fig. 4. En NR, las estaciones base se denominan Nodos B de Próxima Generación (gNB, Next Generation Node B) que proporcionan conectividad a los UE.

La latencia en la RAN consiste en el tiempo de transmisión de un segmento entre la BS y el UE o viceversa. Es originado principalmente por los procesos del nivel físico. Está compuesto por el tiempo de transmisión, tiempo de procesamiento de UE y BS y tiempo de propagación.

2.6. Latencia en el Nivel Físico

Inicialmente, los niveles superiores entregan segmentos de datos al nivel físico, el cual los almacena en un búffer de transmisión. Enseguida, el nivel físico lleva a cabo procesamientos en la información recibida y espera al siguiente TTI para comenzar la transmisión. La información es puesta entonces en la trama de nivel físico, la cual viaja hasta el nivel físico del receptor. El nivel físico del receptor realiza procesamientos sobre la trama recibida y entregará la información a los niveles superiores. Este proceso se muestra en la Fig. 5.

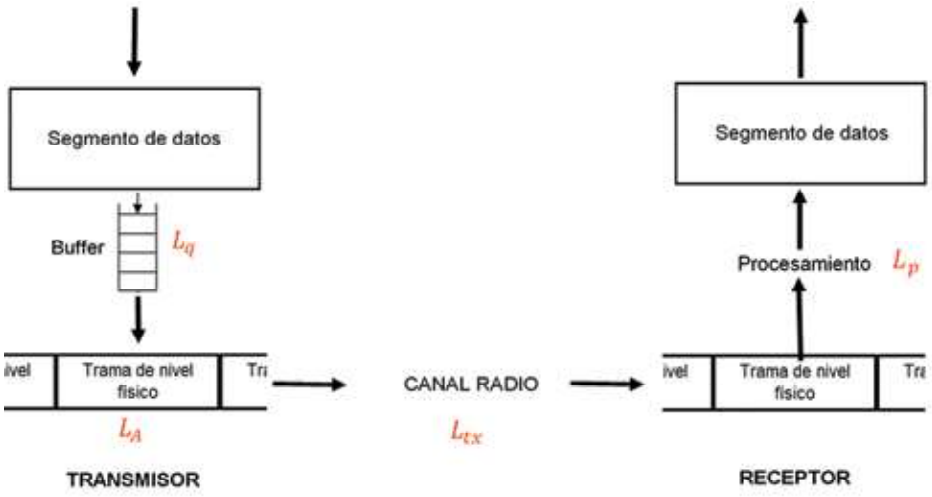

Fig. 5. Proceso de transmisión en el nivel físico.

Científica, vol. 24, núm. 1, pp. 23-32, enero-junio 2020, ISSN 2594-2921, Instituto Politécnico Nacional, México.
DOI: https://doi.org/10.46842/ipn.cien.v24n1a03

La latencia total en el nivel físico puede ser descrita como [10]:

$$L_{PHY} = L_q + L_A + L_{tx} + L_p \quad (1)$$

donde,

L_q : el retardo de cola consiste en el tiempo que tardan los segmentos de datos en llegar al búffer de transmisión y ser transmitidos. El nivel de carga del sistema, la tasa de segmentos y el tamaño de los segmentos influyen directamente sobre este tipo de latencia.

L_A: el retardo de alineación de trama comprende el tiempo de espera de los segmentos de datos al siguiente TTI. La reducción del TTI es una solución obvia para disminuir este tipo de retardo.

L_{tx}: el retardo de transmisión consiste en el tiempo que tarda el nivel físico en colocar una trama de nivel físico en el canal radio.

L_p: el retardo de procesamiento en el receptor consiste en el tiempo que tarda el nivel físico del UE o gNB en procesar la trama de nivel físico recibida. Este retardo depende de la capacidad de procesamiento de los equipos radio.

3. Metodología

Se utiliza la metodología Proceso Racional Unificado (RUP, *Rational Unified Process*) para el desarrollo de las simulaciones. Esta metodología es un proceso de desarrollo incremental en cuatro etapas: análisis de requerimientos, diseño, implementación y pruebas.

3.1. Análisis de requerimientos

Los requerimientos de las simulaciones se describen a continuación:

- Configurar parámetros generales del sistema 5G (ancho de banda, frecuencia de operación, densidad de gNB y UE).
- Configurar parámetros asociados al nivel físico de la RAN (estructura de trama, separación de subportadoras y cantidad de símbolos OFDM por TTI).
- Configurar parámetros asociados a los servicios a desplegar en el sistema 5G (tamaño y tasa de segmentos de datos).
- Obtener la latencia de los segmentos de datos en la RAN.
- Graficar de manera intuitiva los resultados obtenidos en los diferentes escenarios para facilitar el análisis de los resultados obtenidos.

3.2. Diseño

Considerando los componentes de latencia definidos en (1), se diseña el modelo de sistema de tal forma que permita modificar, en primer lugar, la carga del sistema. Para esto, se modela la RAN compuesta por un gNB sirviendo a un número determinado de UE, siendo el número de UE el que permita variar la carga del sistema. Así mismo, se modificará el valor del TTI en la trama de nivel físico y el retardo de procesamiento del gNB y los UE. Los parámetros de tasa y tamaño de segmentos de datos caracterizarán los flujos de datos que simulan la información generada por diferentes servicios. Estos segmentos son transportados por niveles superiores mediante diferentes protocolos. Dado que el gNB no posee la capacidad de implementar dichos protocolos, es necesario incluir dos componentes adicionales en el modelo de simulación del sistema. Por lo tanto, se incluye un servidor remoto que será utilizado para establecer una comunicación a nivel de transporte con los UE de la RAN y así simular los parámetros de tasa y tamaño de segmentos de datos que caracterizarán los flujos de datos entre el servidor y UE. Dado que dicho nodo no puede conectarse directamente a la RAN, se incluye en el modelo de sistema el CN cuyo único objetivo será proporcionar conectividad entre UE y el servidor remoto.

El modelo de sistema se muestra en la Fig. 6. Los parámetros del modelo del sistema se muestran en la Tabla 2.

Tabla 2. Parámetros del modelo de sistema.

Parámetro	Valor
Frecuencia central	28 GHz
Ancho de banda	100 MHz
Número de bloques de recursos radio	1
Subportadoras por bloques de recursos radio	12
Esquema de duplexación	TDD
Modulación	64 QAM
Codificación de Canal	LDPC
Potencia de Transmisión	23 dBm
Número de gNB	1
Número de UEs	20, 40, 60
Tasa de segmentos	400, 900, 1400 segmentos/s
Tamaño de segmentos	500, 1500 Bytes
Tiempo de decodificación	0 ms, 0.1 ms, 0.5 ms, 2TTIs

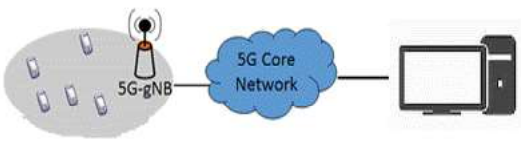

Fig. 6. Modelo de sistema.

3.3. Implementación

ns-3 es un simulador de redes de eventos discretos desarrollado con fines educativos e investigativos. ns-3 es software de código abierto bajo Licencia Pública General (GPL, General Public License). Dicha herramienta software incluye el módulo de ondas milimétricas (mmWave module) desarrollado por la Universidad de Nueva York. Este módulo cuenta con diversos parámetros configurables para modelar diferentes escenarios y permite simular redes 5G completas [11]. Para la implementación del modelo de sistema en ns-3, es necesario definir los nodos, el canal, dispositivos de red y aplicaciones que conforman el modelo, de la siguiente forma:

- **Nodos**: para simular la RAN, se crea un nodo que representa a la estación base y un conjunto de nodos que representan a los UE. El CN es representado por defecto en ns3 mediante 2 nodos, los cuales representan la Entidad de Gestión de Movilidad (MME, Mobility Management Entity), la Pasarela de Servicio (SGW, Serving Gateway) y la Pasarela de Red de Paquetes de Datos (PGW, Packet Data Network Gateway). Finalmente, se crea un nodo adicional para simular el servidor remoto.
- **Canal**: el modelo de canal implementado por defecto en el módulo mmWave de ns3 es el definido en el reporte técnico 38.901 del 3GPP [12], el cual es el apropiado para sistemas 5G. Este componente se configura de acuerdo a los valores presentados en la Tabla 2.
- **Dispositivos de red**: se utilizan los dispositivos de red correspondientes a 5G implementados en el módulo mmWave y los dispositivos de red IP.
- **Aplicación**: se utiliza un generador de tráfico de Protocolo de Datos de Usuario (UDP, User Datagram Protocol) que toma como parámetros la tasa y el tamaño de segmentos.

3.4. Pruebas

Teniendo en cuenta los componentes de la latencia de nivel físico definidos en (1), es posible determinar que el TTI tiene un impacto directo sobre el retardo de alineación de trama y el retardo de transmisión. Se diseñan escenarios de simulación que tendrán en cuenta los parámetros de carga del sistema, tasa y tamaño de segmentos y capacidad de procesamiento. Dado que cada escenario impone diferentes requerimientos, es posible analizar cuál es el valor del TTI adecuado para dichos escenarios y su impacto en la latencia. El retardo de procesamiento es analizado en un escenario específico por lo que no es tenido en cuenta en los demás escenarios. Las variables de entrada y salida se muestran en la Fig. 7 y se definen a continuación:

- Número de UE: esta variable tiene como objetivo variar la carga del sistema. Dado que los recursos radio son limitados, se espera que la variable afecte directamente la latencia en la RAN. Se escogen los valores de 20, 40, 60 UE para simular carga baja, media y alta, respectivamente.
- Tasa de segmentos: esta variable define cuán rápido los segmentos de datos son enviados al nivel físico. Se escogen los valores de 400, 900 y 1400 segmentos/s para simular tasa baja, media y alta de segmentos, respectivamente.
- Tamaño de segmentos: se escoge esta variable dado que el nivel físico debe segmentar los segmentos y adaptarlos a la trama de nivel físico. El tiempo de procesamiento en dicha tarea se verá reflejada en la latencia. Se escogen los valores de 500 y 1500 B para simular segmentos de datos pequeños y grandes, respectivamente
- Tiempo de procesamiento: esta variable simula la capacidad de procesamiento de los equipos radio. Consiste en el tiempo que toma el gNB o UE en procesar la trama de nivel físico recibida. Se escogen los valores fijos de 0 ms, 0.1 ms, 0.5 ms para simular el caso ideal, procesamiento rápido y lento, respectivamente. También se utiliza un valor que depende del TTI, en este caso 2 TTI.
- TTI: el TTI toma los valores de las diferentes numerologías presentadas en la Tabla 2. De esta forma se cuentan con cinco valores de TTI que son implementados en cada escenario. Así mismo, se utilizan mini-TTI con el objetivo de variar el número de símbolos por TTI.

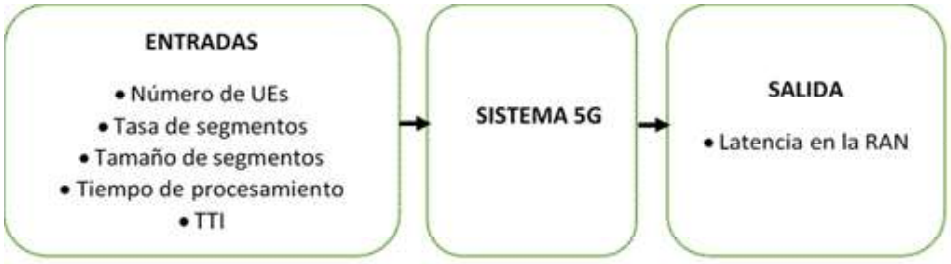

Fig. 7. Variables de entrada y salida.

Científica, vol. 24, núm. 1, pp. 23-32, enero-junio 2020, ISSN 2594-2921, Instituto Politécnico Nacional, México.
DOI: https://doi.org/10.46842/ipn.cien.v24n1a03

- Latencia en la RAN: consiste en el promedio de los valores de latencia de los segmentos cuando son transmitidos desde el gNB a los UE o viceversa.

4. Análisis de resultados

En esta sección se presentan los resultados y el análisis de los escenarios diseñados anteriormente.

4.1. Análisis de resultados

Se simula un sistema 5G con diferentes valores de carga. La carga del sistema depende del número de UEs asociados al gNB. Se simulan 20, 40, 60 UE que corresponden a carga baja, media y alta, respectivamente. Así mismo, se implementan las numerologías presentadas en la Tabla 1 para cada caso. La latencia en la RAN para el DL se muestra en la Fig. 8.

Se observa que la latencia en la RAN se incrementa al aumentar la carga del sistema. Es posible explicar esta tendencia analizando el proceso de distribución de TTI por parte del gNB. Los segmentos de datos son asignados en TTI para su transmisión. A cada usuario se le asigna un TTI determinado, dependiendo de la carga del sistema. Si existiese un único usuario, se le asignarían TTI consecutivos teniendo una transmisión continua e ininterrumpida. Si existen N usuarios, cada usuario tendrá que esperar N-1 TTI para que se le sea asignado un TTI nuevamente y así continuar con la transmisión de su información. De esta forma, el incremento en la carga aumenta el intervalo de espera de cada usuario y por consiguiente la latencia en la RAN se incrementa.

Los segmentos que llegan al buffer del gNB deben esperar un tiempo adicional para ser transmitidos hacia los UE debido al proceso anteriormente descrito. Cada segmento puede ser transmitido hasta que todos los segmentos previos a él sean transmitidos, causando que la latencia en la RAN se incremente aún más. De esta manera, el retardo de cola constituye el componente predominante en la latencia de la RAN para los casos de carga media y alta.

Al implementar numerologías superiores, es decir al disminuir el valor del TTI, se aprecia una reducción significativa en la latencia de la RAN. Teniendo en cuenta (1), es posible analizar este efecto de la siguiente manera:

- Al disminuir el valor del TTI, el tiempo que cada usuario debe esperar para que se le asigne un TTI disminuye también, reduciendo la latencia en la RAN.
- El retardo de alineación disminuye proporcionalmente con el valor del TTI. Este retardo es cuantificado como la mitad del valor del TTI, de esta forma cada segmento debe esperar cada vez menos a la siguiente oportunidad de transmisión.
- El retardo de transmisión es igual al valor del TTI, por lo que disminuye en la misma cantidad.

Enseguida, se varía el número de símbolos por TTI sin modificar el espaciamiento entre subportadoras. Se utiliza la numerología 0, es decir un espaciamiento fijo de 15 KHz. Se consideran los valores de 2, 4 y 7 símbolos por TTI que son comparados con el valor estándar de 14 símbolos por TTI. Se utilizan los valores de 20, 40 y 60 UE para simular carga baja, media y alta, respectivamente. La latencia en la RAN para el DL se muestra en la Fig. 9.

Se observa que la reducción del TTI en este caso también impacta la latencia de la RAN, sin embargo, no posee el mismo efecto que en el caso anterior. De la Fig. 9 se puede observar que:

- La carga del sistema afecta negativamente la latencia como en el primer caso de estudio. Las razones que explican este efecto fueron descritas anteriormente.

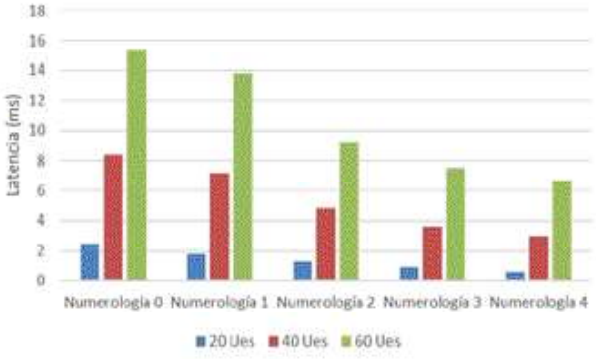

Fig. 8. Latencia en la RAN en el DL para distintas numerologías y diferentes niveles de carga.

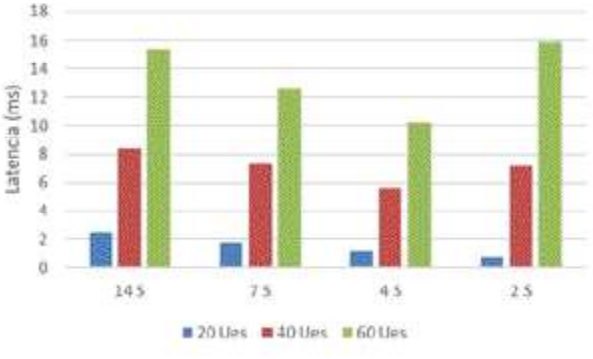

Fig. 9. Latencia en la RAN en el DL para diferentes longitudes de TTI.

- Considerando la carga baja en el sistema, la reducción del TTI disminuye la latencia en la RAN, siendo 2 símbolos por TTI el valor que proporciona menor latencia.
- Considerando carga media y alta en el sistema, la reducción del TTI no siempre implica disminución de la latencia en la RAN. Es posible observar que los valores de 7 y 4 símbolos por TTI disminuyen la latencia en comparación con el valor de 14 símbolos por TTI, sin embargo, al utilizar el valor de 2 símbolos por TTI la latencia en la RAN se deteriora significativamente. Para estos niveles de carga, el valor de 4 símbolos por TTI proporciona menor latencia.

Esta tendencia puede ser explicada al tener en cuenta el *overhead* (hace referencia a la información de señalización y control que es necesaria para llevar a cabo el proceso de comunicación en un sistema de telecomunicaciones) introducido por las señales de referencia que el Sistema 5G necesita para operar correctamente. Dichas señales proporcionan información sobre el estado del canal radio y la red, y son transmitidas tanto en el Enlace de Bajada (UL, Uplink) como en el DL. La señal de Referencia para Demodulación (DMRS, Demodulation Reference Signal), por ejemplo, ocupa dos símbolos del TTI en tres subportadoras de cada bloque de recursos radio.

Para el caso de 7 símbolos por TTI, el overhead introducido por la señal DMRS es del 7.14 %. Para el TTI formado por 2 símbolos, el overhead corresponde al 25%. Al reducir el número de símbolos por TTI, el overhead se incrementa, por consiguiente, los segmentos deben esperar TTI adicionales para su transmisión, incrementando el retardo de cola.

4.2. Escenario 2

En este escenario se varía el flujo de segmentos, para esto, se consideran los valores de 400, 900 y 1400 segmentos/s que representan una tasa baja, media y alta de segmentos, respectivamente. Se implementan las numerologías definidas en la Tabla 1 para cada caso. La latencia en la RAN para el UL y DL se muestra en la Fig. 10 y Fig. 11, respectivamente.

De las Fig. 10 y Fig. 11 es posible observar que: en el DL, el incremento en la tasa de segmentos afecta negativamente la latencia en la RAN. Esta tendencia puede ser explicada debido a que, al aumentar la tasa de segmentos, el búffer del gNB se carga más rápidamente, incrementando el retardo de cola. Los segmentos que llegan al búffer de transmisión tienen que esperar un tiempo mayor para ser transmitidos. Al reducir el TTI, los segmentos que esperan en el búffer de transmisión son transmitidos más rápidamente, reduciendo el retardo de cola y por lo tanto la latencia en la RAN.

En el UL, el incremento en la tasa de segmentos no posee el mismo efecto que en el DL. Es posible observar que no es el valor de 400 segmentos/s el que proporciona menor latencia. Este resultado puede ser explicado al analizar el proceso de transmisión de datos en el UL.

En el UL, una vez que el UE posee datos disponibles para su transmisión, envía una Solicitud de Asignación de Recursos (SR, Scheduling Request) al gNB. Al recibir dicho mensaje, el gNB responde con un Mensaje de Concesión (UG, Uplink Grant)

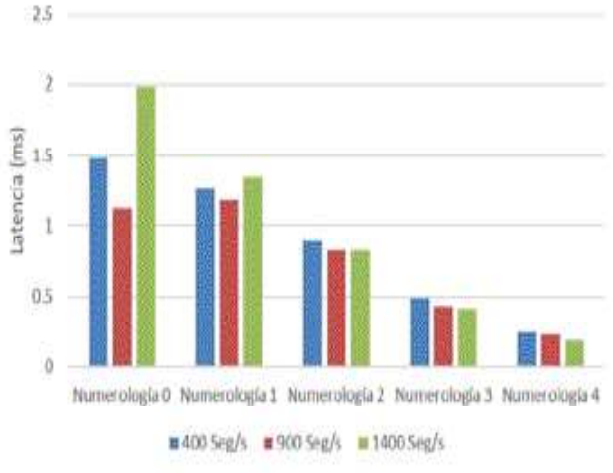

Fig. 10. Latencia en la RAN en el UL para distintas numerologías y diferentes tasas de segmentos.

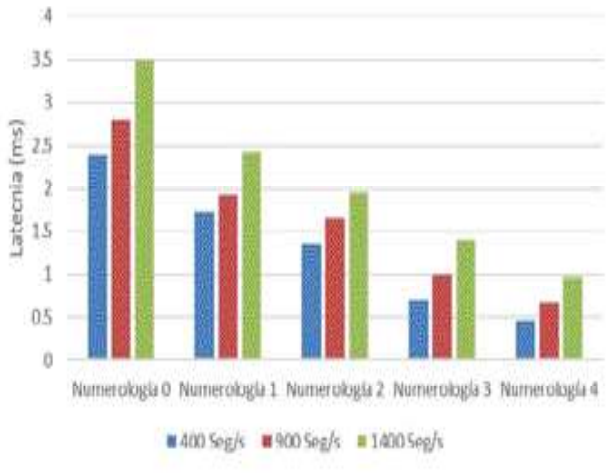

Fig. 11. Latencia en la RAN en el DL para distintas numerologías y diferentes tasas de segmentos.

Científica, vol. 24, núm. 1, pp. 23-32, enero-junio 2020, ISSN 2594-2921, Instituto Politécnico Nacional, México.
DOI: https://doi.org/10.46842/ipn.cien.v24n1a03

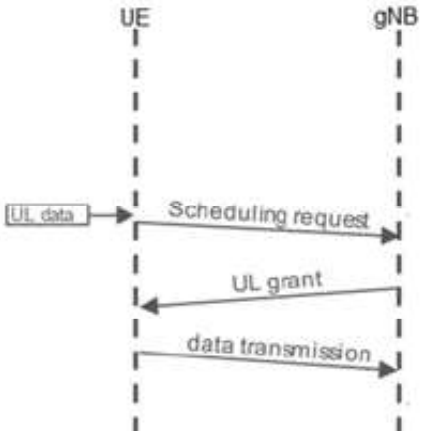

Fig. 12. Proceso de acceso a recursos en el UL.

indicando los recursos asignados al UE y la configuración para la transmisión. Dicho proceso se muestra en la Fig. 12.

El proceso de asignación de recursos en el UL toma más tiempo que en el DL, en el cual el gNB asigna los recursos correspondientes inmediatamente. En el UL, el UE debe esperar a la siguiente oportunidad de transmisión para enviar el SR, dicha transmisión durará por lo menos 1ms. Enseguida, el gNB debe procesar el SR recibido para luego transmitir el UG en 1ms. El UE finalmente procesa el UG e inicia la transmisión de los datos correspondientes. Este proceso incrementa significativamente la latencia en la RAN, si el UE necesita enviar SR frecuentemente para acceder a los recursos del sistema.

El búffer de transmisión se llena con los segmentos de datos del UE y este inicia el proceso anteriormente descrito mediante el envío de un SR para la transmisión de dichos datos. Todos los segmentos que están en el búffer pueden ser transmitidos usando el mismo SR, pero si el búffer está vacío, el UE debe enviar un nuevo SR y empezar el proceso de acceso a recursos nuevamente.

Considerando la numerología 0, la tasa media de segmentos (900 segmentos/s) implica el envío de 1 segmento cada milisegundo aproximadamente, valor que disminuye la probabilidad de que cada segmento encuentre un búffer vacío, si se compara con la tasa baja de segmentos. De esta forma, para la tasa media de segmentos, el UE envía mensajes SR con menor frecuencia, disminuyendo la latencia en la RAN. Sin embargo, al considerar la tasa alta de segmentos, el búffer de transmisión no es capaz de vaciar los segmentos con la misma velocidad con la que llegan, incrementando el retardo de cola y deteriorando la latencia en la RAN.

Para las numerologías superiores (2,3,4), la velocidad de transmisión se incrementa significativamente, esto significa que el búffer se vacía con mayor rapidez, incrementando la probabilidad de envíos de mensajes SR. Es posible notar que las tasas baja y media de segmentos proporcionan mayor latencia en este caso, debido el efecto descrito anteriormente.

4.3. Escenario 3

En este escenario se varía el tamaño de segmentos, considerando los valores de 50 bytes y 1500 bytes que representan segmentos pequeños y grandes, respectivamente. Se implementan las numerologías definidas en la Tabla 1. La latencia en la RAN para el DL se muestra en la Fig. 13.

Es posible observar que el tamaño de segmentos afecta negativamente la latencia. Dicha tendencia es explicada debido a que segmentos más grandes requieren mayor cantidad de procesamiento para su transmisión. Así mismo, se asignan un mayor número de TTI para transmitir segmentos más grandes, por lo que el búffer de transmisión se llenará con más rapidez, aumentando el retardo de cola y la latencia en la RAN.

4.4. Escenario 4

En este escenario se varía el retardo de procesamiento del UE y el gNB. Para esto, se consideran los valores de 0, 0.1 y 0.5 ms que representan el caso ideal, procesamiento rápido y lento, respectivamente. También se considera un valor que depende

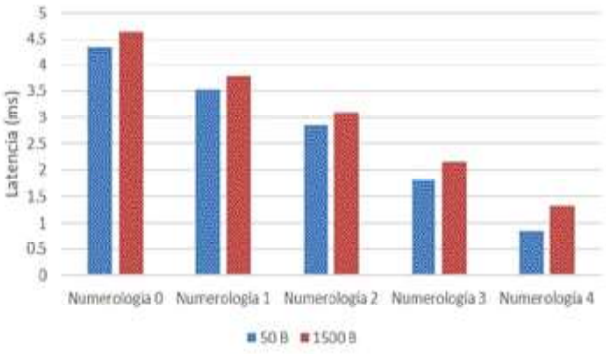

Fig. 13. Latencia en la RAN en el DL para distintas numerologías y diferentes tamaños de segmentos.

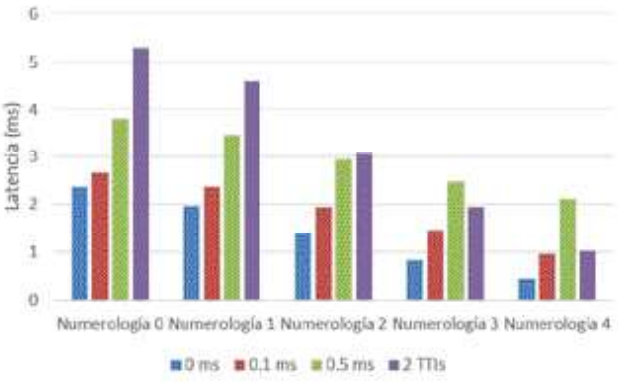

Fig. 14. Latencia en la RAN en el DL para distintas numerologías y diferentes tiempos de procesamiento.

del TTI, en este caso 2 TTI. Así mismo, se implementan las numerologías definidas en la Tabla 1. La latencia en la RAN para el DL se muestra en la Fig. 14.

De la Fig. 14 es posible observar que: la capacidad de procesamiento de los equipos radio influye significativamente sobre la latencia de la red. Si se toma como base el caso ideal (0 ms), el retardo de procesamiento duplicará la latencia de la RAN en algunos casos.

Para las numerologías 0, 1 y 2, el valor que proporciona mayor latencia es el de 2 TTI. Esto se debe a que, para estas numerologías, el TTI es relativamente grande, por lo que el valor de 2 TTI supera a los valores fijos considerados.

Para las numerologías 3 y 4, el TTI se reduce significativamente, por lo que el valor de 2 TTI constituye ahora un tiempo de procesamiento rápido. En este caso dicho valor es menor que el tiempo de procesamiento lento y aproximadamente igual al tiempo de procesamiento rápido.

5. Conclusiones

Es posible concluir que la reducción del TTI, ya sea implementando diferentes numerologías o reduciendo el número de símbolos por TTI posee un impacto significativo sobre la latencia de la RAN. La utilización de una u otra alternativa para la reducción de la latencia dependerá de los requisitos de la red. Al utilizar numerologías altas, se debe contar con un mayor ancho de banda. Dadas las frecuencias de operación altas, la cobertura del sistema se verá afectada debido a los fenómenos de propagación al implementar numerologías superiores. Así mismo es necesario considerar la carga del sistema y el posible overhead introducido por señales de referencia o protocolos de red.

Referencias

[1] N. Isachenko, "El papel de la información y las tecnologías de la información y la comunicación en la sociedad moderna", *Utopia y Praxis Latinoamericana*, vol. 23, pp. 361-367, 2018. Disponible en: 10.5281/zenodo.1512122

[2] M. Agiwal, A. Roy, N. Saxena, "Next Generation 5G Wireless Networks: A Comprehensive Survey," *IEEE Commun. Surv. Tutor.*, vol. 18, no. 3, pp. 1617-1655, thirdquarter 2016.

[3] S. Zhang, X. Xu, Y. Wu, L. Lu, "5G: Towards energy-efficient, low- latency and high-reliable communications networks," in Proc. *IEEE Int. Conf. on Commun. Syst. (ICCS)*, noviembre 2014, pp. 197-201.

[4] J. Butler, "5G Spectrum Challenges", 5G Radio Technology Seminar. *Exploring Technical Challenges in the Emerging 5G Ecosystem*, 2015. Available: 10.1049/ic.2015.0028

[5] The Mobile Broadband Standard, "Release 15 - 3GPP", 2019. [en línea]. Disponible en: https://www.3gpp.org/release-15

[6] B. Rong, J. Zhou, M. Kadoch, G. Sun, "Emerging Technologies for 5G Radio Access Network: Architecture, Physical Layer Technologies, and MAC Layer Protocols", *Wireless Communications and Mobile Computing*, vol. 2018, pp. 1-2, 2018. Disponible en: 10.1155/2018/6082161

[7] S. Lien, S. Shieh, Y. Huang, B. Su, Y. Hsu, H. Wei, "5G New Radio: Waveform, Frame Structure, Multiple Access, and Initial Access", *IEEE Communications Magazine*, vol. 55, núm. 6, pp. 64-71, 2017. Disponible en: 10.1109/mcom.2017.1601107

[8] Techmahindra.com, 2019 [en línea]. Disponible en: http://www.techmahindra.com/documents/WhitePaper/WhitePaperLatencyAnalysis.pdf

[9] N. Maskey, S. Horsmanheimo, L. Tuomimaki, "Analysis of latency for cellular networks for smart grid in suburban area", *IEEE PES Innovative Smart Grid Technologies, Europe*, 2014. Disponible en: 10.1109/isgteurope. 2014.7028750

[10] S. Ahmadi, "A Practical Systems Approach to Understanding 3GPP LTE Releases 10 and 11 Radio Access Technologies". *Elsevier Science*, LTE-Advanced, 2013.

[11] GitHub, 2018 [en línea]. Disponible en: https://github.com/nyuwireless-unipd/ns3-mmwave

[12] 3GPP, *Study on channel model for frequencies from 0.5 to 100 GHz*, Reporte Técnico 38.901, 2018.

Científica, vol. 24, núm. 1, pp. 23-32, enero-junio 2020, ISSN 2594-2921, Instituto Politécnico Nacional, México.
DOI: https://doi.org/10.46842/ipn.cien.v24n1a03