
XIV Jornadas de Ingeniería Telemática.
JITEL 2019.
Universidad de Zaragoza.

Actas de las XIV Jornadas de Ingeniería Telemática (JITEL 2019), Zaragoza (España), 22-24 de octubre de 2019.

ISBN: X-XXXX-XX-XXX-X
DOI: XXXXXX

Optimización conjunta del nivel *split* y *scheduling* en redes 5G

Luis Diez, Víctor González y Ramón Agüero
Departamento de Ingeniería de Comunicaciones,
Universidad de Cantabria
Plaza de la Ciencia s.n. 39005, Santander, Cantabria. España
ldiez@tlmat.unican.es, victor.gonzalezcar@alumnos.unican.es, ramon@tlmat.unican.es

***Resumen*—Las técnicas de virtualización serán una pieza fundamental en las futuras tecnologías 5G. Sin embargo, las soluciones totalmente centralizadas, tales como *Cloud Radio Access Network* (C-RAN), podrían no ser factibles, debido a las requisitos adicionales impuestos al *fronthaul* de la red. En este sentido, las técnicas *flexible functional split* permiten definir niveles de centralización de manera flexible, proporcionando, así, un compromiso entre el rendimiento y la aplicabilidad práctica. A pesar del creciente interés en estas técnicas, no se ha prestado mucha atención a su interacción con la planificación a la hora de acometer la transmisión de tramas (*scheduling*). Es por ello que en este trabajo se analiza la gestión combinada del nivel de centralización y el envío de tramas, a fin de minimizar el retardo en la red de acceso. En concreto, se compara, en diferentes escenarios, la solución óptima global con aquellas aportadas por optimizaciones parciales, cuya implementación puede resultar más asequible desde el punto de vista de complejidad computacional. A la vista de los resultados obtenidos, se puede concluir que las políticas de planificación fija para el envío de tramas presenta un comportamiento similar al óptimo global en escenarios con tecnologías de acceso heterogéneas y tráfico homogéneo. Además, se ha comprobado que, para los escenarios analizados, es preferible mantener un política de envío fija y optimizar únicamente el nivel de *split*.**

***Palabras Clave*—functional split, scheduling, 5G, cloud RAN, retardo, NFV**

I. INTRODUCCIÓN

Entre los cambios en la arquitectura de red que van a caracterizar los despliegues 5G, uno de los que se consideran más determinantes es, sin duda, la virtualización de funciones de red *Network Function Virtualization* (NFV), utilizando técnicas *Software Defined Networking* (SDN). Mientras que en el pasado las arquitecturas de red celular, tales como 4G, han evolucionado hacia topologías descentralizadas, los cada vez más exigentes requisitos de las redes 5G requieren una mayor coordinación de los elementos de acceso, lo que, a su vez, solo puede lograrse mediante arquitecturas centralizadas.

Esto se consigue mediante la separación de las funcionalidades pertenecientes a la red de acceso, de modo que un conjunto de ellas se virtualizan, centralizándose en centros de procesado de datos, mientras que el resto permanece en los elementos de acceso. Con la aparición de la virtualización de funciones de red, se propusieron inicialmente soluciones totalmente centralizadas (C-RAN). En estos casos, se virtualizan todas las funciones de red, quedando en los elementos de acceso o *Remote Radio Head* (RRH), las funciones básicas de la capa física. Sin embargo, este tipo de soluciones requieren capacidades de comunicación muy altas entre la RRH y la unidad de banda base o *Base-Band Unit* (BBU) virtualizada, lo que puede ser difícil de satisfacer en ciertos escenarios. Por esta razón, en los últimos años han surgido propuestas que permitan niveles de centralización flexibles (*flexible functional split*) [1], [2].

Este cambio de paradigma en la arquitectura de red, permitirá implementar soluciones que den respuesta a los requisitos de los despliegues 5G, a la vez que habilitará una reducción de costes, en comparación con las soluciones totalmente centralizadas. Por otro lado, la posibilidad de definir diferentes niveles de centralización conlleva responder a varias cuestiones. Por un lado, se ha de decidir el nivel de centralización (*functional split*) de los diferentes elementos de acceso, de acuerdo al tipo de coordinación necesaria. Por otro lado, es necesario planificar la transmisión de tramas en las BBU hacia las RRH que gestiona. Esta planificación, o *scheduling*, deberá hacerse de forma que se minimice el retardo, ya que se trata de uno de los principales parámetros a optimizar en la tecnología 5G.

En este trabajo se analiza, sobre una arquitectura de *flexible split*, la configuración conjunta del nivel de centralización y planificación de envío de tramas. Para ello, se toma como base el trabajo realizado por Koutsopoulos [3], en el que se describe, de forma teórica, el problema de optimización subyacente. El autor caracteriza

la complejidad del problema conjunto, proporcionando pautas para solucionar los casos en los que uno de los dos parámetros (nivel de centralización o planificación del envío) se conoce. Sin embargo, no se proporcionan resultados prácticos de ninguno de los problemas.

En concreto, este trabajo persigue los siguientes objetivos:

- Implementación de las soluciones descritas en [3] para minimizar el retardo.
- Evaluación de dichas soluciones sobre diferentes escenarios, usando configuraciones realistas.
- Análisis de la pérdida de rendimiento cuando se usan optimizaciones parciales (uno de los parámetros conocidos), con respecto a la solución óptima global.

El resto del documento sigue la estructura que se indica a continuación. En la Sección II se analizan los trabajos relacionados existentes en la literatura, indicando las principales diferencias con el estudio que aquí se presenta. Posteriormente, en la Sección III se describe el modelo del sistema, así como las variantes del problema de optimización para obtener la solución óptima de selección de *split* y *scheduling*. A continuación, en la Sección IV se evalúa el rendimiento de estos problemas sobre diferentes escenarios. Finalmente, el artículo concluye en la sección V, donde se resumen las contribuciones y se enumeran líneas futuras de investigación.

II. TRABAJOS PREVIOS

Como se ha mencionado anteriormente, las arquitecturas C-RAN [4], [5] se consideran una de las soluciones clave para satisfacer los requisitos impuestos por la tecnología 5G. La principal idea que subyace en este tipo de soluciones es la de trasladar funciones de red, típicamente localizadas en las estaciones base, a un controlador central. Sin embargo, las soluciones en las que se centralizan todas las funciones de red pueden no resultar prácticas, debido a las altas capacidades (tanto de transmisión como de retardo) que se impondrían sobre los enlaces del *fronthaul*. En este sentido, las soluciones totalmente centralizadas requerirían la implementación de un *fronthaul* basado únicamente en fibra óptica [6], [7], lo que a su vez implica costes de despliegue muy altos. Por ello, han aparecido varias iniciativas que proponen un cambio de diseño del *fronthaul* [8], de modo que se puedan definir diferentes niveles de centralización. El lector puede encontrar una revisión detallada de los niveles de *split* que se están definiendo en [9].

Existen además trabajos que proponen no solo la selección de niveles de *split*, sino que el proceso de selección se lleve a cabo de forma dinámica, dando lugar al concepto de *flexible functional split*. En este caso el nivel de centralización se puede adaptar en función de los requisitos de retardo, así como del estado de los enlaces del *fronthaul*. Este tipo de arquitectura basada en centralización flexible para 5G se ha descrito en [10] y validado, en entorno de laboratorio, en [11]. Por otro lado, en [12], [13] se describen las principales características de este tipo de soluciones.

A partir del concepto de *flexible functional split* algunos trabajos han propuesto técnicas de mejora de la eficiencia energética [14], [15]. Otros se han centrado en su combinación con redes de transporte ópticas [16], [17]. Por otro lado, algunos trabajos han estudiado la interacción de la selección de *split* con la gestión de recursos radioeléctricos. Sin embargo, más allá del estudio realizado por Koutsopoulos [3], citado anteriormente, no se ha prestado atención a la optimización conjunta de la selección de *split* y la planificación de envío de tramas desde la BBU a los puntos de acceso gestionados por este, que es precisamente donde se sitúa la principal contribución de este trabajo.

III. MODELADO DEL SISTEMA

Como se ha mencionado en secciones anteriores, en este trabajo se adopta el modelo propuesto en [3], que se reproduce a continuación de forma resumida.

Se considera una arquitectura de red celular, en la que se realiza *functional split*, de manera que un conjunto de RRH, $\mathcal{R}$, se conectan a una unidad central en la nube, capaz de virtualizar varios BBU mediante un conjunto de enlaces $\mathcal{L}$. Se asume que la unidad central dispone de un solo procesador con capacidad computacional C^B, de forma que solo se puede procesar de forma concurrente una trama, perteneciente a una RRH. De forma similar, se usa C_i para denotar la capacidad computacional de la RRH i, mientras que L_i indica la capacidad de transferencia de datos del enlace entre la RRH i y la unidad en la nube. Cabe mencionar que en este trabajo únicamente se considerará el enlace descendente, aunque el modelo puede ser igualmente aplicado al ascendente.

Se asume un escenario en el que el tiempo está ranurado, de modo que se transmite una nueva trama para cada RRH al inicio de cada ranura. A fin de gestionar el envío de tramas, existe un controlador en la unidad central que, de manera global, decide el nivel de *split* aplicado a cada trama, así como el orden en que estas se envían.

Por simplicidad, se asume que todas las RRH pueden acomodar el mismo conjunto de niveles de *split*, $\mathcal{F}$. De esta manera, en cada ranura el controlador selecciona un vector de niveles de *split* $\boldsymbol{s} := \{s_1, ...s_{|R|}\} \in \mathcal{S}$, donde $s_i \in \mathcal{F}$ se corresponde con la decisión tomada para la RRH $i \in \mathcal{R}$ y $\mathcal{S}$ representa el conjunto de todas las decisiones posibles. Cabe indicar que el número de posibles configuraciones crece exponencialmente con el número de opciones de centralización, de modo que $|\mathcal{S}| = |\mathcal{F}|^{|\mathcal{R}|}$.

A partir de la decisión de *split* de cada RRH i, s_i, se definen las variables ω_{i,s_i} y $\hat{\omega}_{i,s_i}$ para indicar la carga computacional necesaria, respectivamente, en la BBU y RRH. De forma similar se define la función d_{i,s_i} para representar la cantidad de datos que tienen que transmitirse a través del enlace L_i, dependiendo de la decisión concreta de *split*.

De manera similar al modelado de la centralización, Π representa el conjunto de posibles políticas de transmisión (*scheduling*). Una política concreta se representa mediante un vector, $\boldsymbol{\pi} = \{\pi_1, ..., \pi_{|R|}\}$, de números enteros positivos, donde π_i indica el orden en que se envía la trama

perteneciente a la RRH i. Por ejemplo, si se asume que se tiene un escenario con 4 RRH, la política de envío $\boldsymbol{\pi} = \{3, 4, 2, 1\}$ indicaría que en primer lugar se envía la trama de la RRH 4 y a continuación las de las estaciones 3, 1 y 2. Cabe destacar que el espacio de las políticas de envío es $|\Pi| = |\mathcal{R}|\,!$.

Teniendo en cuenta la notación presentada, se puede calcular el retardo de procesado en la BBU de una trama i como:

$$\delta^B_{i,s_i} = \frac{\omega_{i,s_i}}{C^B} \tag{1}$$

De forma similar, considerando la cantidad de datos que se deben transmitir como consecuencia de la decisión de *split* de cada trama i, el retardo de transmisión se define como:

$$\delta^L_{i,s_i} = \frac{d_{i,s_i}}{L_i} \tag{2}$$

Finalmente, el retardo asociado al procesado en la estación viene dado por la siguiente expresión:

$$\delta^R_{i,s_i} = \frac{\hat{\omega}_{i,s_i}}{C_i} \tag{3}$$

De este modo, se puede calcular el retardo total de una trama i como:

$$d_i(\boldsymbol{\pi}, \boldsymbol{s}) = \delta^B_{i,s_i} + \delta^L_{i,s_i} + \delta^R_{i,s_i} + \sum_{j:\pi_j<\pi_i} \delta^B_{j,s_j} \tag{4}$$

donde $\sum_{j:\pi_j<\pi_i} \delta^B_{j,s_j}$ indica el tiempo en que la trama espera en la BBU para ser enviada. Igualmente, se puede definir el retardo total en el sistema, D, como la suma de retardos de cada trama:

$$\begin{aligned} D(\boldsymbol{s}, \boldsymbol{\pi}) &= \sum_{i\in\mathcal{R}} d_i(\boldsymbol{s}, \boldsymbol{\pi}) = \\ &= \sum_{i\in\mathcal{R}} \left(\delta^B_{i,s_i} + \delta^L_{i,s_i} + \delta^R_{i,s_i} + \sum_{j:\pi_j<\pi_i} \delta^B_{j,s_j} \right) \end{aligned} \tag{5}$$

El retardo global definido en la Ecuación 5 se puede reformular para expresarlo en función de la contribución realizada por cada trama. Para ello, se agrupan los retardos asociados al procesado en la BBU de la siguiente manera:

$$\begin{aligned} &\sum_{i\in\mathcal{R}} \left(\delta^B_{i,s_i} + \sum_{j:\pi_j<\pi_i} \delta^B_{j,s_j} \right) = \sum_{i\in\mathcal{R}} \sum_{j:\pi_j\leq\pi_i} \delta^B_{j,s_j} = \\ &= \delta^B_{1,s_1} + (\delta^B_{1,s_1} + \delta^B_{2,s_2}) + ... + (\delta^B_{1,s_1} + ... + \delta^B_{|\mathcal{R}|,s_{|\mathcal{R}|}}) = \\ &= \sum_{i\in\mathcal{R}} \delta^B_{i,s_i} \cdot (|\mathcal{R}| - \pi_i + 1) \end{aligned} \tag{6}$$

De este modo, se puede definir el retardo global como la suma de los generados por cada trama i, $g^i_{s_i,\pi_i}$, en lugar del experimentado por ellas:

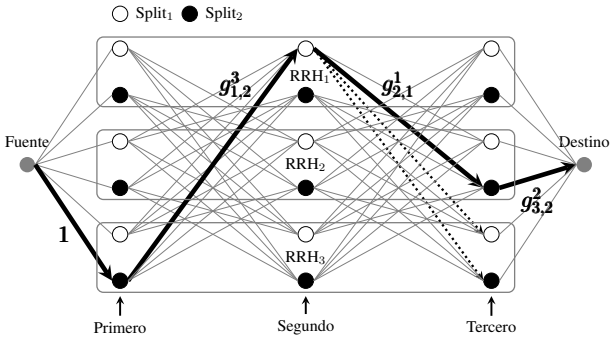


Fig. 1: Instancia del problema con 3 RRH y 2 posibles niveles de *split*. La línea continua resalta indica la solución seleccionada, mientras que las líneas discontinuas denotan soluciones no admisibles.

$$\begin{aligned} D(\boldsymbol{s}, \boldsymbol{\pi}) &= \sum_{i\in\mathcal{R}} g^i_{s_i,\pi_i} = \\ &= \sum_{i\in\mathcal{R}} \left(\delta^L_{i,s_i} + \delta^R_{i,s_i} + \delta^B_{i,s_i}(|\mathcal{R}| - \pi_i + 1) \right) \end{aligned} \tag{7}$$

A. Formulación del problema

Aunque pueden existir varios parámetros a optimizar, tales como la minimización del retardo máximo o de las diferencias de retardo entre tramas, este trabajo se centra en la reducción del retardo total del sistema.

De acuerdo con el modelo presentado en [3], el problema de optimización global del retardo se puede plantear definiendo el sistema como un grafo dirigido. Los nodos en el grafo forman una trama regular con $|\mathcal{R}|$ columnas y $|\mathcal{R}| \times |\mathcal{F}|$ filas, donde cada nodo se corresponde con una decisión conjunta de *split* y política de envío para una trama determinada, mientras que los arcos que conectan los nodos tienen pesos iguales al retardo asociado a dicha decisión. De este modo, las columnas en el grafo representan el orden de envío, mientras que las filas indican la decisión de *split* y la trama seleccionada. Finalmente, añadiendo dos nodos virtuales (*Fuente* y *Destino*), el problema se reduce a la búsqueda del camino más corto entre dichos nodos.

A fin de ilustrar el planteamiento del problema, la Figura 1 muestra el grafo resultante de un sistema compuesto por 3 RRH y 2 niveles de centralización. Como se puede ver, todas las posibles soluciones de *split* y orden de envío de cada RRH se agrupan en dos filas. En la figura se resalta una posible solución completa con líneas continuas, en las que el coste de cada enlace se corresponde con la decisión asociada el nodo fuente de dicho enlace. En el ejemplo la trama 3 se envía en primer lugar con nivel de *split* 2, a continuación la trama 1 con nivel 1 y, finalmente, la trama 2 usando nuevamente el nivel de *split* 2. Como se puede observar, tras seleccionar la trama 3 es necesario añadir restricciones adicionales, que garanticen que dicha trama no se selecciona de nuevo, lo que se indica en la figura con líneas discontinuas.

Se denota el grafo del sistema como $G(\mathcal{V}, \mathcal{A})$, donde $\mathcal{A}$ es el conjunto de enlaces y $\mathcal{V}$ el conjunto de vértices, de modo que v_0 y $v_{|\mathcal{R}|+1}$ se corresponden, respectivamente, con los nodos virtuales *Fuente* y *Destino* ($|\mathcal{V}| = |\mathcal{R}| + 2$). A continuación se define el subconjunto de nodos correspondientes a una RRH i como $\mathcal{V}_i \subseteq \mathcal{V}$.

La selección de los enlaces se realiza usando una variable binaria $x_{i,j}$, que toma valor 1 si se selecciona el enlace entre los nodos i y j, y 0 en caso contrario. Por simplicidad, se usará la variable w_{ij} para indicar el coste del enlace que conecta cada par de nodos (i, j). Finalmente, el problema global de reducción del retardo se puede expresar como:

Problem 1 (Selección conjunta de *split* y política de envío)**.**

$$\textbf{\textit{min.}} \quad \sum_{i,j} x_{ij} \cdot w_{ij} \tag{8}$$

$$\textbf{\textit{s.t.}} \quad \sum_{i \in \mathcal{V}/k} x_{ik} + \sum_{i \in \mathcal{B}/k} x_{ki} = T_k \qquad \forall k \in \mathcal{V} \tag{9}$$

$$\sum_{i \in \mathcal{V}/\mathcal{V}_i} x_{ik} = 1 \qquad \forall k \in \mathcal{V}_i \tag{10}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i, j \in \mathcal{V} \tag{11}$$

donde la Ecuación 9 se corresponde con la restricción de conservación de flujo; la constante T_k indica el flujo entrante y saliente de cada nodo. Esta constante toma valor 1 y -1 respectivamente para los nodos *Fuente* y *Destino* ($T_0 = 1; T_{|\mathcal{R}|+1} = -1$) y 0 para el resto. Mediante la Ecuación 10 se asegura que únicamente se toma una decisión por RRH, tal y como se mostró en la Figura 1 mediante las líneas discontinuas.

Como se puede observar, el problema resultante es binario y lineal, *Binary Linear Program* (BLP), que es conocido por ser *np-hard* y, por tanto, difícil de resolver. Además, el tamaño crece exponencialmente con el número de RRH y posibles niveles de *split*, siendo el espacio de posibles soluciones $|\mathcal{F}|^{|\mathcal{R}|} \times |\mathcal{R}|!$ y el número de variables $2|\mathcal{R}||\mathcal{F}| + (|\mathcal{R}| - 1) \times |\mathcal{F}||\mathcal{R}| \times (|\mathcal{F}||\mathcal{R}| - |\mathcal{F}|)$[1].

En las siguientes secciones se describen brevemente las modificaciones del problema original cuando bien el nivel de *split* o la política de envío son conocidas.

B. *Selección de* split *conocida*

Si se fija el nivel de *split* para cada RRH, se pueden conocer los valores de retardo, de modo que el problema se reduce a minimizar el producto del retardo en la BBU por el orden de transmisión. Se puede observar que la optimización global en este caso se obtiene usando una política que envía en primer lugar las tramas con menor retardo en la BBU (*shortest-job-first*).

[1]El primer término, $2|\mathcal{R}||\mathcal{F}|$, se corresponde con los enlaces salientes y entrantes a los nodos virtuales. En el segundo término se multiplica el número de columnas ($|\mathcal{R}| - 1$) por el número de filas $|\mathcal{R}||\mathcal{F}|$ y enlaces salientes de cada nodo ($|\mathcal{F}||\mathcal{R}| - |\mathcal{F}|$).

C. *Política de envío conocida*

En el caso en que se fije la política de envío, la complejidad del problema se reduce. En concreto se trataría de seleccionar el nivel de *split* que minimiza la expresión $\delta^L_{i,s_i} + \delta^R_{i,s_i} + \delta^B_{i,s_i}(|\mathcal{R}| - \pi_i + 1)$, donde π_i es conocido para cada RRH. Si se tiene en cuenta que, en la práctica, el posible número de *splits* es bajo, esta tarea se puede realizar de manera eficiente mediante simples algoritmos de búsqueda.

IV. EVALUACIÓN DE RENDIMIENTO

Como se ha visto anteriormente, la complejidad asociada al planteamiento general del problema podría no ser de utilidad práctica en casos reales, especialmente en escenarios con un número elevado de elementos de acceso. Por esta razón, parece razonable explorar alternativas subóptimas y analizar su rendimiento en diferentes escenarios.

En esta sección se analiza el comportamiento del problema general definido en Problema 1 y, a continuación, se compara con el resultado obtenido por las soluciones subóptimas que se han presentado anteriormente. Para realizar esta comparación se usarán 3 escenarios diferentes, en los que se varían los dos parámetros con mayor impacto en el problema: la relación de las capacidades computacionales entre la RRH y BBU y la longitud de las tramas, que es directamente proporcional a los retardos. A fin de obtener unos resultados más claros y fáciles de comparar, se usará la capacidad computacional de la BBU como referencia. De esta manera, los escenarios se definen en función de la relación de capacidades de procesado entre la RRH y BBU , $r^R_i = C_i/C^B$, y del retardo de procesado de esta última δ^B_{i,s_i} para una longitud de trama fija. Respecto al retardo de transmisión, se asume que se dispone de enlaces con alta capacidad [18], por lo que se considera despreciable en comparación con los otros.

En general, los escenarios están compuestos por una BBU y 10 RRH, y se analiza el comportamiento estadístico de cada algoritmo, realizando 1000 experimentos independientes para cada configuración. Las capacidades de cómputo de los elementos de la red están obtenidos de los modelos reales descritos en [19], [20] y para la longitud de las tramas se usan como referencia los valores mencionados en [21], [22]. Finalmente, cabe indicar que el nivel de *split* se define como la variación de carga de procesado de las tramas entre la BBU y RRH. Por lo tanto, el nivel de centralización de una RRH i para un *split* s viene dado por la expresión $w_{i,s_i}/(w_{i,s_i} + \hat{w}_{i,s_i})$, como se detalla en [23], [24]. El problema conjunto se soluciona usando la herramienta de optimización GLPK [25] y para cada uno de los escenarios se compara el comportamiento obtenido con:

- La solución usando una política de envío conocida, según la cual las tramas se envían en orden ascendente en relación a su tamaño.
- La solución con la selección de *split* conocida, aplicando niveles de centralización del 0 (red tradicional), 50, y 100% (C-RAN).

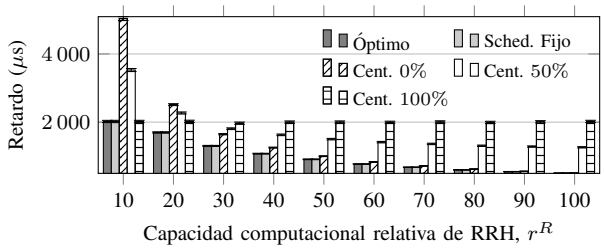


Fig. 2: Retardo medio por trama con longitud de trama heterogénea y diferentes valores de capacidad computacional de las RRHs

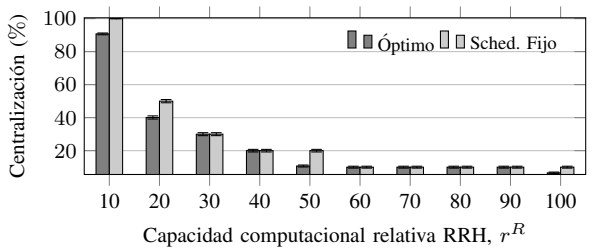


Fig. 3: Nivel medio de centralización con longitud de trama heterogénea y diferentes valores de capacidad computacional de las RRHs

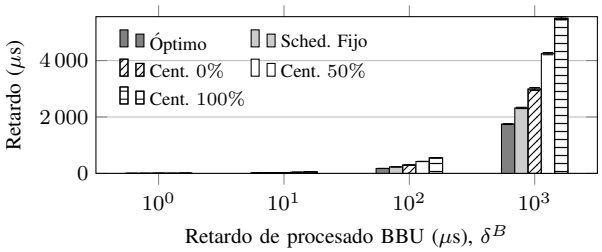


Fig. 4: Retardo medio por trama con capacidad de cómputo de las RRHs heterogénea y diferentes valores de longitud de trama

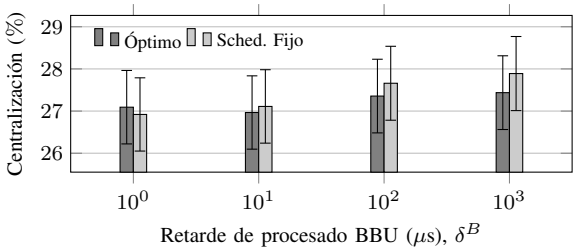


Fig. 5: Nivel medio de centralización con capacidad de cómputo de las RRHs heterogénea y diferentes valores de longitud de trama

A. RRHs homogéneos y tráfico heterogéneo

En el primer escenario se considera un despliegue de red homogéneo, con elementos de acceso iguales, de forma que todos tienen la misma capacidad de cómputo ($r_i^R = r^R \quad \forall i \in \mathcal{R}$). Por otro lado, el retardo de procesado en la BBU, δ^B, se distribuye uniformemente en el intervalo $[1, 1000]$ μs, en cada experimento independiente.

La Figura 2 muestra, para diferentes valores de la carga computacional en los elementos de acceso, el retardo medio por trama, así como su intervalo de confianza del 95%. Como se puede observar, la solución con política de envío fija (primero las tramas más cortas) muestra un comportamiento similar al óptimo global para todas las configuraciones. Por otro lado, el esquema con el nivel de *split* fijo manifiesta un comportamiento menos previsible. En este sentido, para configuraciones en las que los elementos de acceso tienen menor capacidad el mejor resultado es aquel en que se tiene un mayor nivel de centralización (C-RAN), como cabría esperar. En el otro extremo, cuando la capacidad de las RRH es similar a la de la BBU la mejor solución sería un esquema distribuido. En general, cuando se usa el esquema que fija un nivel de *split*, es necesario adaptar esa configuración al escenario concreto sobre el que se aplica.

En lo que se refiere a la selección del *split*, en la Figura 3 se muestra el nivel de centralización para las soluciones óptima y de política de envío fija. Como se puede ver, de nuevo, ambos esquemas tienen un comportamiento muy similar, de forma que son capaces de seleccionar el *split* de acuerdo a las características del escenario.

B. RRHs heterogéneos y tráfico homogéneo

En el segundo escenario se fija la longitud de las tramas, lo que da lugar a una configuración de tráfico homogéneo, que se tiene en cuenta fijando el valor del retardo de cómputo en la BBU, δ^B. Por otro lado, la capacidad relativa de las RRH, r_i^R, se selecciona de manera aleatoria en el intervalo $[0.1, 1]$.

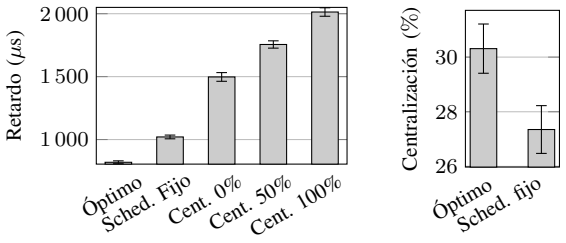
De forma similar al escenario anterior, en primer lugar se muestra en la Figura 4 el retardo medio de las tramas, representando también el intervalo de confianza. Como cabía esperar, con independencia del algoritmo seleccionado, el retardo aumenta con el tamaño de las tramas. Además, se puede ver que el esquema de política de transmisión fija siempre tiene un comportamiento mejor al mostrado por el de nivel de centralización fijo.

Seguidamente, la Figura 5 muestra el nivel de centralización seleccionado. En este caso, se puede observar que para valores de longitud de trama pequeños (menor valor de δ^B) la solución óptima selecciona niveles de centralización superiores, mientras que el comportamiento es precisamente el contrario cuando se incrementa la longitud de trama.

C. RRHs heterogéneos y tráfico heterogéneo

En el último escenario se selecciona de forma aleatoria tanto la longitud de trama como la capacidad computacional de las RRH usando los intervalos indicados anteriormente.

La Figura 6a muestra el retardo experimentado por las tramas cuando se usan los diferentes esquemas de selección. Se puede ver que, de forma similar a los casos anteriores, la política de transmisión fija siempre se comporta mejor que aquella en la que se fija el nivel



(a) Retardo medio

(b) Nivel de centralización

Fig. 6: Rendimiento de los algoritmos con longitud de trama y capacidad de cómputo de las RRHs heterogéneos

de centralización. Sin embargo, en este escenario ambas soluciones sub-óptimas están alejadas del óptimo global.

Finalmente, en la Figura 6b se puede apreciar que la solución óptima tiende a niveles de centralización superiores.

V. CONCLUSIONES

En este trabajo se han analizado diferentes algoritmos para minimizar el retardo en escenarios con *flexible functional split* en los que la selección del nivel de centralización y la gestión de envío de tramas afectan notablemente al retardo. Con esta premisa, resulta preciso tener en cuenta ambos aspectos de manera conjunta para obtener el mejor rendimiento del sistema en términos de retardo. Sin embargo, el problema de optimización subyacente resulta *np-hard*, por lo que su uso práctico podría verse comprometido. Por ello, se han analizado alternativas que, bajo ciertos supuestos, reducen de forma considerable la complejidad del problema original.

A partir del trabajo realizado por Koutsopoulos [3], se ha implementado un algoritmo que soluciona tanto el problema original como las alternativas que llevan a cabo optimizaciones parciales. Se han analizado y evaluado las diferentes soluciones en diversos escenarios, comparando los retardos y niveles de centralización de las optimizaciones parciales con aquellas proporcionadas por el problema original. A la luz de los resultados, en escenarios donde los elementos de acceso son homogéneos en términos de capacidad de procesado, el rendimiento de soluciones que consideren esquemas fijos de envío de tramas es similar al óptimo global. Además, también se ha observado que la selección del *split*, realizado por este tipo de optimización parcial, también es semejante al óptimo. Por otro lado, las soluciones con nivel de centralización fijo presentan un rendimiento notablemente peor que, a su vez, se ve afectado por las características concretas del escenario.

También se ha estudiado el comportamiento de las diferentes alternativas en escenarios donde la red de acceso es heterogénea. En este caso, el rendimiento mostrado por las optimizaciones parciales se encuentra lejos del óptimo global. Sin embargo, de acuerdo a los resultados presentados, nuevamente las soluciones con políticas de envío constante muestran un mejor comportamiento que aquellas en las que se fija en nivel de *split*.

Partiendo de este trabajo, en el futuro se pretende analizar diferentes alternativas a la optimización global, especialmente en escenarios con heterogeneidad en la red de acceso. En concreto, se analizarán técnicas de agrupamiento (*clustering*) que agruparían tramas o elementos de acceso con características similares. Así, se reduciría la complejidad del problema. Por otro lado, se analizarán escenarios más complejos en los que se considerarán procesos de llegada de tramas a la BBU menos predecibles, lo que precisará otro tipo de soluciones, tales como las basadas en teoría de colas.

AGRADECIMIENTOS

Los autores agradecen la financiación del Gobierno de España (Ministerio de Economía y Competitividad, Fondo Europeo de Desarrollo Regional, FEDER) de este trabajo a través de los proyectos ADVICE: *Dynamic provisioning of connectivity in high density 5G wireless scenarios* (TEC2015-71329-C2-1-R) y FIERCE: *Future Internet Enabled Resilient Cities* (RTI2018-093475-A-100).

REFERENCIAS

[1] I. . W. Group, "Next generation fronthaul interface." [Online]. Available: http://sites.ieee.org/sagroups-1914/

[2] "Study on new radio access technology: Radio access architecture and interfaces," 3rd Generation Partnership Project (3GPP), TR 38.801, 2017.

[3] I. Koutsopoulos, "Optimal functional split selection and scheduling policies in 5g radio access networks," in *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2017, pp. 993–998.

[4] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (c-ran): a primer," *IEEE Network*, vol. 29, no. 1, pp. 35–41, Jan 2015.

[5] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud ran for mobile networks—a technology overview," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 405–426, Firstquarter 2015.

[6] G. O. Pérez, J. A. Hernández, and D. Larrabeiti, "Fronthaul network modeling and dimensioning meeting ultra-low latency requirements for 5g," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, no. 6, pp. 573–581, June 2018.

[7] A. Garcia-Saavedra, J. X. Salvat, X. Li, and X. Costa-Perez, "Wizhaul: On the centralization degree of cloud ran next generation fronthaul," *IEEE Transactions on Mobile Computing*, vol. 17, no. 10, pp. 2452–2466, Oct 2018.

[8] C. I, Y. Yuan, J. Huang, S. Ma, C. Cui, and R. Duan, "Rethink fronthaul for soft ran," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 82–88, Sep. 2015.

[9] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5g mobile crosshaul networks," *IEEE Communications Surveys Tutorials*, vol. 21, no. 1, pp. 146–172, Firstquarter 2019.

[10] P. Arnold, N. Bayer, J. Belschner, and G. Zimmermann, "5g radio access network architecture based on flexible functional control / user plane splits," in *2017 European Conference on Networks and Communications (EuCNC)*, June 2017, pp. 1–5.

[11] Y. Alfadhli, M. Xu, S. Liu, F. Lu, P. Peng, and G. Chang, "Real-time demonstration of adaptive functional split in 5g flexible mobile fronthaul networks," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, March 2018, pp. 1–3.

[12] D. Harutyunyan and R. Riggio, "Flex5g: Flexible functional split in 5g networks," *IEEE Transactions on Network and Service Management*, vol. 15, no. 3, pp. 961–975, Sep. 2018.

[13] ——, "Flexible functional split in 5g networks," in *2017 13th International Conference on Network and Service Management (CNSM)*, Nov 2017, pp. 1–9.

[14] D. A. Temesgene, M. Miozzo, and P. Dini, "Dynamic functional split selection in energy harvesting virtual small cells using temporal difference learning," in *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2018, pp. 1813–1819.

[15] L. Wang and S. Zhou, "Flexible functional split in c-ran with renewable energy powered remote radio units," in *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, May 2018, pp. 1–6.

[16] A. Marotta, D. Cassioli, K. Kondepu, C. Antonelli, and L. Valcarenghi, "Efficient management of flexible functional split through software defined 5g converged access," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6.

[17] Y. Li, J. Mårtensson, M. Fiorani, B. Skubic, Z. Ghebretensaé, Y. Zhao, J. Zhang, L. Wosinska, and P. Monti, "Flexible ran: A radio access network concept with flexible functional splits and a programmable optical transport," in *2017 European Conference on Optical Communication (ECOC)*, Sep. 2017, pp. 1–3.

[18] Q. C. Li, H. Niu, A. T. Papathanassiou, and G. Wu, "5g network capacity: Key elements and technologies," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 71–78, Mar. 2014.

[19] P. Rost, I. Berberana, A. Maeder, H. Paul, V. Suryaprakash, M. Valenti, D. Wübben, A. Dekorsy, and G. Fettweis, "Benefits and challenges of virtualization in 5g radio access networks," *IEEE Communications Magazine*, vol. 53, no. 12, pp. 75–82, Dec. 2015.

[20] K. Wang, K. Yang, H. Chen, and L. Zhang, "Computation diversity in emerging networking paradigms," *IEEE Wireless Communications*, vol. 24, no. 1, pp. 88–94, Feb. 2017.

[21] X.-L. Wu, W.-M. Li, F. Liu, and H. Yuand, "Packet size distribution of typical internet applications," in *2012 International Conference on Wavelet Active Media Technology and Information Processing (ICWAMTIP)*, Dec. 2012, pp. 276–281.

[22] Z. Sun, D. He, L. Liang, and H. Cruickshank, "Internet qos and traffic modelling," *IEE Proceedings - Software*, vol. 151, no. 5, pp. 248–255, Oct. 2004.

[23] N. Makris, P. Basaras, T. Korakis, N. Nikaein, and L. Tassiulas, "Experimental evaluation of functional splits for 5g cloud-rans," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[24] D. Wubben, P. Rost, J. S. Bartelt, M. Lalam, V. Savin, M. Gorgoglione, A. Dekorsy, and G. Fettweis, "Benefits and impact of cloud computing on 5g signal processing: Flexible centralization through cloud-ran," *IEEE Signal Processing Magazine*, vol. 31, no. 6, pp. 35–44, Nov. 2014.

[25] G. Project, "Gnu linear programming kit." [Online]. Available: https://www.gnu.org/software/glpk/