JITEL 2021

LIBRO DE ACTAS

XV Jornadas de Ingeniería Telemática

A CORUÑA 2021

ISBN: 978-84-09-35131-2

Editores:
Victor Manuel Carneiro Díaz
Laura Victoria Vigoya Morales

XV Jornadas de Ingeniería Telemática – JITEL 2021


XV Jornadas de Ingeniería Telemática.
JITEL 2021.
Universidad de A Coruña.

Actas de las XV Jornadas de Ingeniería Telemática (JITEL 2021), A Coruña (España), 27-29 de octubre de 2021.

Retardo en redes fronthaul con split funcional flexible: un modelo basado en teoría de colas

Luis Diez, Ramón Agüero
Departamento de Ingeniería de Comunicaciones. Universidad de Cantabria.
{ldiez, ramon}@tlmat.unican.es

En este trabajo se estudia el retardo en topologías de red vRAN, considerando tanto las estaciones base, que se dividen entre un controlador y un cabezal de radio remoto, y la red de conmutación de paquetes fronthaul que los une. Se contempla el uso de functional split flexible, según el que las funciones que se ejecutan en cada una de las dos entidades se puede modificar dinámicamente. Se propone un modelo basado en teoría de colas, que es capaz de reflejar de manera precisa el comportamiento de estos nodos, que se valida tras una extensa campaña de medidas. Además, se utiliza la teoría de redes abiertas de Jackson para modelar el retardo extremo a extremo en la red fronthaul, lo que permite analizar el impacto de establecer diferentes políticas de red. Los resultados ponen de manifiesto que el modelo propuesto se puede emplear para establecer las configuraciones óptimas de red, pues los resultados que ofrece son prácticamente idénticos a los obtenidos mediante simulación.

***Index Terms*—vRAN, functional split, teoría de colas, redes de Jackson**

I. Introducción

Uno de los requisitos más exigentes en los sistemas 5G es el relativo al retardo, que resulta fundamental para soportar adecuadamente servicios de tipo *Ultra-Reliable Low Latency Communication* (URLLC), tales como los relativos a Internet táctil o a la conducción autónoma. Por otro lado, las arquitecturas de redes de acceso radio están sufriendo una continua evolución, incorporando, entre otros, elementos SDN y NFV, dando lugar a lo que ya se conoce como virtual RAN. Aunque la capacidad de virtualización de funciones de las estaciones base tiene grandes ventajas (como reducción de costes), también aparecen nuevos aspectos que deben ser analizados, tales como el retardo asociado a esta virtualización.

Inicialmente, las soluciones Cloud-RAN (C-RAN) proponían la completa virtualización de las unidades de banda base, con una conexión de gran capacidad con los antenas. En este artículo se considera una evolución de este solución, en la que existen diferentes grados de centralización (*functional-splits*) [1] y esta centralización puede ser modificada de manera dinámica, lo que da lugar a la funcionalidad conocida como *flexible functional-split*. Esta adaptación permite solventar algunas de las limitaciones de las soluciones C-RAN [2] iniciales, adaptando la red a las necesidades concretas y recursos que se tienen. En estas arquitecturas la estación base se divide en una *centralized unit* (CU) que contiene ciertas funciones y que se conecta, a través de una red de conmutación de paquetes *fronthaul*, a las *distributed units* (DU) en las que están el resto de funciones. A su vez las DUs poseen una conexión de gran capacidad con las antenas o *radio units*. A fin de soportar servicios que precisen URLLC, es necesario conocer el retardo asociado entre las CUs y DUs. En este artículo se extiende el modelo presentado inicialmente en [3], que permite analizar el retardo asociado a una CU o DU, para conocer la latencia extremo a extremo en la red *fronthaul* cuando se aplica una determinada política de selección de split. El modelo propuesto podría ayudar al dimensionado de este tipo de redes, y a establecer límites en el tráfico admisible ante ciertos requisitos de retardo.

El resto del documento se estructura de la siguiente manera. En la Sección II se presenta una revisión de la literatura relativa a *flexible functional-split*, resaltando el carácter innovador del modelo propuesto. A continuación, en la Sección III se presenta el modelo basado en cadenas de *Markov* y teoría de redes de *Jackson*, que se valida en la Sección IV sobre diferentes escenarios. en la Sección V se presentan las conclusiones más relevantes que se han obtenido, y se enumeran líneas de trabajo futuro.

II. Trabajos relacionados

El potencial de las arquitecturas *functional-split* flexibles se ha analizado ampliamente en la literatura [4], [3], y ya existen trabajos describiendo su implementación para posibilitar la selección dinámica del nivel de centralización, tales como [5], [6].

Más relacionados con este trabajo, han aparecido propuestas de políticas de selección de *split* centradas en diferentes aspectos. Por ejemplo, en [7], [8] Harutyunyan *et al.* modelan la selección de split como un problema de tipo *Virtual Network Embedding* (VNE) formulado como

un *Integer Linear Program* (ILP). De forma similar, los autores de[9] y [10] proponen algoritmos de selección de *split* que asegure el uso de técnicas de cooperación entre elementos de acceso, mientras se hace un uso eficiente de la red *fronthaul*, permitiendo el despliegue de servicios que requieran URLLC.

Otros trabajos prestan atención a métricas diversas en sus políticas de selección de *split*, tales como la tasa [11], o el retardo [12]. Entre los parámetros considerados, existen multitud de propuestas que se centran en la eficiencia energética, tales como [13], [14], [15]. Otro grupo de trabajos se centran en la interacción de la selección de *split* con la red *fronthaul* óptica. En este sentido se ha analizado tanto la reducción de latencia [16] como la limitación de capacidad de la red [17]. Asimismo, existen trabajos que presentan soluciones de *orquestación* [18], que permitan la reconfiguración global de la red de acceso ante cambios en el nivel de centralización.

Aunque la revisión de la literatura se podría extender, la mayoría de las investigaciones previas proponen soluciones para definir el nivel de centralización. Por el contrario, el modelo presentado en este trabajo tienen como objetivo modelar el comportamiento de la red de *fronthaul*, en función del retardo, cuando se aplica cualquier política.

III. Modelo de colas para el *fronthaul*

En esta sección se va a presentar el modelo, basado en teoría de colas, que considera dos tipos de nodos: (1) refleja el comportamiento del CU o DU; y (2) se usan para modelar *switches* y enlaces en la red. El segundo tipo se modelará mediante un nodo M/M/1, mientras que el primero precisa una aproximación más compleja. A continuación se describirá el modelo de los nodos que representan los CU y DU, para posteriormente establecer el retardo extremo a extremo esperado en la red de *fronthaul*. Para facilitar la lectura, la Tabla I enumera los símbolos que se utilizan en el resto de la sección.

Tabla I: Símbolos y variables

	Nodos CU/DU
s	Número de *splits*
λ	Tasa de llegada de tramas
μ_j	Tasa de servicios del *split* j^{th}
$\alpha_{j,k}$	Probabilidad de transitar del *split* j^{th} al k^{th} $\sum_{k=1}^{s}\alpha_{j,k}=1,\ \alpha_{j,j}=0$
γ_j	Tasa de cambio del *split* j^{th}
ξ_j	Inverso del tiempo de *stand-by* del *split* j^{th}
$\pi_i(t)$	Probabilidad del estado (i,t)
	Hay i tramas en el nodo:
	(1) t impar, usando el *split* $j:\ j=\frac{t+1}{2},\ (i,j)$
	(2) t par, *stand-by* tras *split* $j:\ j=\frac{t}{2},\ (i,\widetilde{j})$
π_i	Vector columna: $[\pi_i(1)\dots\pi_i(t)\dots\pi_i(2s)]$
$\mathcal{Q}$	matriz infinitesimal del proceso de QBD
F	matriz de re-envío
B	Matriz de transición hacia atrás
L, L_0	matrices de transición de estado con mismo número de tramas
	Red *fronthaul*
λ	Tasa de llegada el nodo (*switch*/enlace)
μ	Tasa de servicio del nodo (*switch*/enlace)
ρ	Ocupación del nodo (*switch*/enlace)
Λ	Vector de tasas de entrada en los nodos
Γ	Vector de tráfico externo
	$\gamma_i \neq 0$ solo para nodos CU
$\mathcal{R}$	Matriz de encaminamiento de la red *fronthaul*

A. Modelo de los nodos CU y DU

Como se ha mencionado anteriormente, el modelo de los nodos CU y DU es una extensión del presentado en [3]. Se considera una comunicación *downlink*, aunque se podría aplicar también al *uplink*. Se asume que las tramas llegan al CU siguiendo un proceso de *Poisson* de tasa $\lambda\ ms^{-1}$ y que se admiten s *splits*, cada uno de los cuales se caracteriza por un tiempo de servicio con distribución exponencial y valor medio $\mu_k^{-1}\ ms$ para cada *split* k^{th}. De acuerdo a la política adoptada se asume que el tiempo de permanencia en cada nivel de *split* k también está distribuido exponencialmente, con media $\gamma_k^{-1}\ ms$. Al cambiar de *split* se permanece en situación de *standby* durante el tiempo necesario para proceder a la reconfiguración, que también se asume exponencial, con media $\xi_k^{-1}\ ms$, para cada *split* k^{th}. Al abandonar un *split* k, y siempre de acuerdo con la política utilizada, el sistema usa el *split* l con probabilidad α_{kl}, y se define $\alpha_{kk}=0$, para asegurar que no se transita al mismo *split*.

Las principales mejoras con respecto al trabajo presentado en [3] son:

- Se consideran diferentes tiempos de permanencia para cada nivel de *split*.
- El tiempo en el estado *standby* es diferente para cada *split*.
- Se asegura que el *split* de destino es diferente al de origen en cada cambio.

Estas modificaciones permiten un modelado más realista y un mayor nivel de configuración, que da lugar a la cadena de *Markov* tri-dimensional que se muestra en la Figura 1.

Se definen dos tipos de estados con las duplas (i,j) y $(i,\widetilde{j})$ respectivamente. La primera dupla indica el estado de operación normal, donde i se corresponde con el número de tramas en el nodo, y j indica el índice asociado al *split* funcional. La segunda dupla representa el estado de *standby* al abandonar el *split* j. De este modo, la cadena de *Markov* resultante tiene s planos horizontales, cada uno representando un nivel de *split*. Si el nodo se encuentra activo en el *split* j, ante la llegada de una trama (tasa λ) o al terminar el procesado (tasa μ_j) se produce una transición a derecha o izquierda, respectivamente. Además, se puede transitar al estado de *stanby*, $(i,\widetilde{j})$, con tasa γ, de modo que las tramas pueden seguir llegando, pero no se procesan hasta abandonar ese estado, tal como se ve en la Figura 1.

Tras salir de *stanby* el nodo pasa a otro nivel de *split* con tasa ξ_j. En concreto el nuevo *split* k se selecciona con probabilidad α_{jk} con $k\in\{1,\dots,s\},\ k\neq j$, de modo que la tasa de transición desde $(i,\widetilde{j})$ a (i,k) es $\alpha_{jk}\cdot\xi_j$. Aunque durante la estancia en *standby* no se pueden procesar tramas, se asume que es posible almacenarlas hasta volver a estar en situación de atenderlas. El modelo presentado da lugar a un proceso de quasi nacimiento y muerte (quasi-birth-death, QBD), en el que cada nivel se corresponde con todos los estados con el mismo número de tramas: (i,j) y

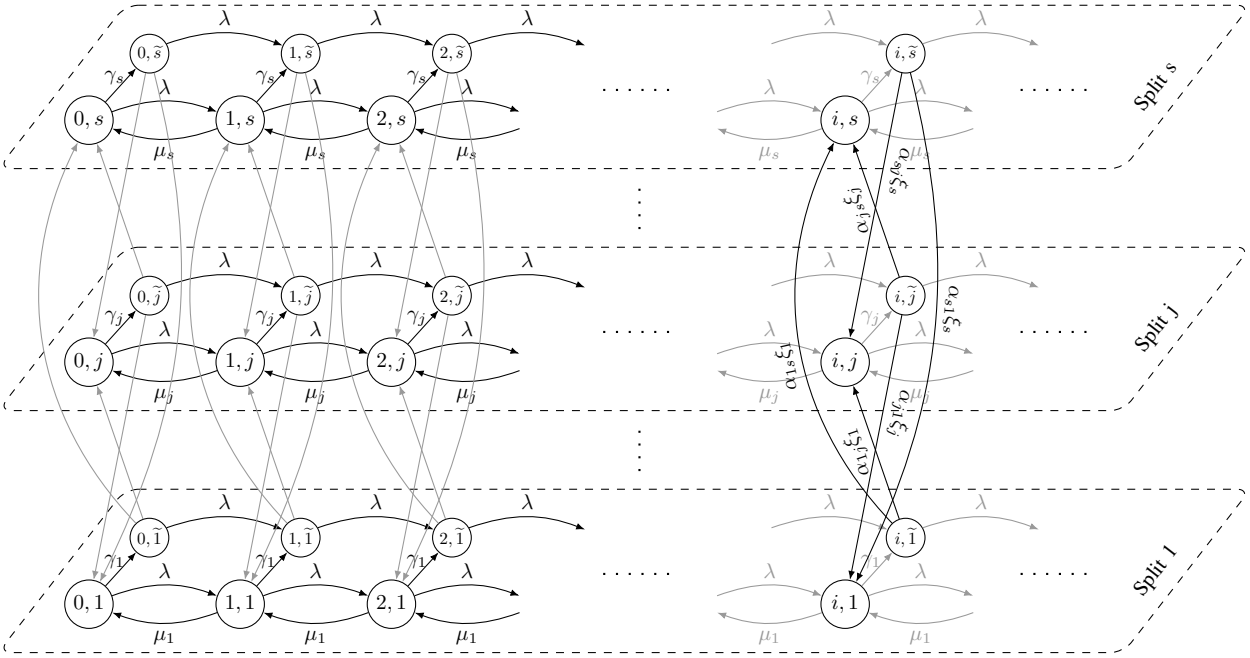


Fig. 1: Cadena de *Markov* para los nodos CU y DU

$(i, \widetilde{j})$, para $j, \widetilde{j} \in \{1, \ldots, s\}$. Se propone aplicar el método Matrix Geometric para definir el retardo de procesado de cada trama, tal como se muestran los trabajos de Neuts y Hajek [19], [20]. La matriz infinitesimal que caracteriza el QBD se define como:

$$\mathcal{Q} = \begin{bmatrix} L_0 & F & 0 & 0 & \cdots \\ B & L & F & 0 & \cdots \\ 0 & B & L & F & \cdots \\ \vdots & & \ddots & & \ddots \end{bmatrix} \tag{1}$$

donde $L_0, B, L, F \in \mathbb{R}^{2s \times 2s}$. Las matrices B, F se definen en la ecuación (2), L en la ecuación (3) y $L_0 = L + B$.

$$F = \begin{bmatrix} \lambda & 0 & \cdots & 0 \\ 0 & \lambda & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda \end{bmatrix},$$

$$B = \begin{bmatrix} \mu_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \mu_2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \mu_s & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \tag{2}$$

Se define la distribución estacionaria del proceso QBD como $\Pi = [\pi_0, \pi_1, \pi_2, \ldots]$, donde π_i es un vector columna de longitud $2s$, y $\pi_i(t),\ t \in \{1, \ldots, 2s\}$ es la probabilidad de tener i tramas en el nodo cuando: (1) t es impar, el nodo se encuentra en el *split* j, y $j = \frac{t+1}{2}$, (2) t es par, el nodo está en *standby* tras pasar por el *split* j, $j = \frac{t}{2}$. Si el nodo está trabajando en régimen de estabilidad, la distribución estacionario existe, y hay una matriz constante R que cumple la siguiente relación [19, Theorem 3.1.1]

$$R^2 \cdot B + R \cdot L + F = 0, \tag{4}$$

donde $R \in \mathbb{R}^{2s \times 2s}$. Aunque no hay una solución cerrada para la ecuación cuadrática (4), se puede utilizar un método iterativo para encontrar R. Además, se sabe que existe una única solución positiva, con la que se puede obtener π_0:

$$\begin{aligned} \pi_0^\intercal \left(L_0 + R \cdot B\right) &= \mathbf{0}^\intercal, \\ \pi_0^\intercal \left(I - R\right)^{-1} \mathbf{1} &= 1, \end{aligned} \tag{5}$$

donde $\mathbf{0}, \mathbf{1}$ son vectores de ceros y unos respectivamente, de longitud $2s$. Así, la distribución estacionaria $\mathbf{\Pi} = [\pi_0, \pi_1, \ldots]$ se obtiene como:

$$\pi_i^\intercal = \pi_0^\intercal \cdot R^i. \tag{6}$$

A partir de la distribución estacionaria, se pude obtener fácilmente el número medio de tramas en el nodo $\overline{N_{cu/du}}$ como:

$$\overline{N_{cu/du}} = \left\| \frac{\pi_1}{(I-R)^2} \right\|_1 = \left\| \frac{\pi_0^\intercal \cdot R}{(I-R)^2} \right\|_1 \tag{7}$$

donde $\| \cdot \|_1$ es la norma-1. Finalmente, usando la ley de *Little* se puede encontrar el retardo medio por trama $\tau_{\text{cu/du}}$, que tiene en cuenta tanto el tiempo de espera como de procesado:

$$\tau_{\text{cu/du}} = \frac{\overline{N_{\text{cu/du}}}}{\lambda} \tag{8}$$

Como se ha mencionado, la distribución estacionaria existe si la tasa de servicio media en el nodo es superior a la tasa de entrada. Por lo tanto, se puede establecer la

$$L = \begin{bmatrix} -(\lambda+\mu_1+\gamma_1) & \gamma_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -(\lambda+\xi_1) & \alpha_{12}\cdot\xi_1 & 0 & \alpha_{13}\cdot\xi_1 & \cdots & \alpha_{1s}\cdot\xi_1 & 0 \\ 0 & 0 & -(\lambda+\mu_2+\gamma_2) & \gamma_2 & 0 & \cdots & 0 & 0 \\ \alpha_{21}\cdot\xi_2 & 0 & 0 & -(\lambda+\xi_2) & \alpha_{23}\cdot\xi_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & -(\lambda+\mu_s+\gamma_s) & \gamma_s \\ \alpha_{s1}\cdot\xi_s & 0 & \alpha_{s2}\cdot\xi_s & 0 & \alpha_{s3}\cdot\xi_s & \cdots & 0 & -(\lambda+\xi_s) \end{bmatrix} \tag{3}$$

tasa máxima $\lambda_{\max}$ que garantice la estabilidad del sistema como:

$$\lambda_{\max} = \sum_{i=1}^{s} \theta_i \cdot \mu_i \tag{9}$$

donde θ_i es la probabilidad de que el nodo se encuentre en un nivel de *split*, la cual se puede calcular resolviendo el siguiente sistema:

$$\Theta^{\intercal} \cdot A = \mathbf{0}^{\intercal} \quad ; \quad \Theta^{\intercal} \cdot \mathbf{1} = 1 \tag{10}$$

donde Θ es un vector columna de longitud $2s$ con las probabilidades de los *splits* y estado de *standby*, $A = L + B + F$, y $\mathbf{0}$ y $\mathbf{1}$ son vectores columna de ceros y unos respectivamente de longitud $2s$.

B. Retardo extremo a extremo en el fronthaul

Como se ha mencionado, se asumen que los nodos CU y DU están conectados por una red de conmutación de paquetes formada por *switches* y enlaces que podrían utilizar tecnologías diferentes. Estos elementos (enlaces y *switches*) se modelan como sistemas M/M/1, lo que permite aplicar teoría de redes abiertas de *Jackson*. Se modela la topología de red como un grafo dirigido $\mathcal{G} = (\mathbb{V}, \mathbb{E})$, donde $\mathbb{V}$ y $\mathbb{E}$ son el conjunto de nodos y enlaces, respectivamente. Si se asume que existen c CUs, d DUs, n *switches* y l enlaces, se puede definir $V \triangleq |\mathbb{V}| = c + d + n + l$. Se define la matriz de encaminamiento $\mathcal{R}$, de tamaño $V \times V$, que indica cómo las tramas recorren la red entre los CUs y sus DUs correspondientes. En la Figura 2 se muestra, a modo de ejemplo, una conexión entre la CU_x y DU_x a través de in *switch* S_x y los enlaces correspondientes. Como se puede ver, el modelo basado en el proceso QBD se utiliza en los nodos CU y DU, mientras que el *switch* y los enlaces se modelan como sistemas M/M/1.

Si se asume que se respetan las condiciones de los teoremas de Burke y Jackson [21], [22], se puede establecer el retardo extremo a extremo como la suma de los retardos asociados a cada nodo en la ruta. Estas condiciones implican que el proceso de tráfico a la salida de cada nodo sea estadísticamente idéntico al de entrada. En el caso de los nodos M/M/1 el retardo se puede calcular como $\tau_{\text{mm1}} = \frac{1}{\mu-\lambda}$, donde μ y λ son las tasas de servicio y de tráfico de entrada al nodo. En este caso, se garantiza la estabilidad si $\mu > \lambda$. Se asume que únicamente los CUs reciben tráfico, y que la matriz de encaminamiento $\mathcal{R}$ indica la ruta hasta el DU correspondiente. Además, los *switches* y enlaces pueden ser compartidos por varios flujos de tráfico. Con ello, se define Λ como el vector de tasas de entrada λ_v de cada nodo $v \in \mathbb{V}$, que se puede calcular como [23], [22]:

$$\Lambda = \Phi \cdot (\mathcal{I} - \mathcal{R})^{-1} \tag{11}$$

donde Φ es otro vector fila que contiene el tráfico externo en la red, de modo que $\phi_v = 0$ para los *switches*, enlaces y DUs, y $\phi_v \neq 0$ para los CUs. Por lo tanto, usando la matriz de encaminamiento $\mathcal{R}$ y las tasas de entrada en los CUs se puede calcular la la tasa de entrada y la ocupación en cada nodo y, a partir de ello, el retardo correspondiente. Finalmente, el retardo extremo a extremo para cada flujo $f \in \mathbb{F}$, siendo $\mathbb{F}$ el conjunto de flujos de entrada, se obtiene como:

$$\overline{\tau_f} = \sum_{v \in \mathcal{P}(f)} \tau_v \quad ; \quad \mathcal{P}(f) : \mathbb{F} \longrightarrow \mathbb{V} \tag{12}$$

donde $\mathcal{P}(f)$ es una función que devuelve los nodos que atraviesa el flujo f.

Además, es posible establecer el retardo promedio en la red (sin necesidad de calcular los retardos individuales por flujo), aplicando la ley de *Little*:

$$\overline{\tau} = \frac{\sum_{v \in \mathbb{V}} n_v}{\lambda_0} \tag{13}$$

donde λ_0 es la tasa total de tráfico externo en la red: $\lambda_0 = \sum_{v \in \mathbb{V}} \phi_v$. Por otro lado n_v es el número medio de tramas en el nodo v, definido (para *switches* y enlaces) por:

$$n_v = \frac{\rho}{1-\rho} \tag{14}$$

En la ecuación (14) ρ representa la ocupación del nodo, calculada como $\rho = \frac{\lambda}{\mu}$. Como se discutirá a continuación, el proceso de salida de las CUs y DUs no es estrictamente de *Poisson* y esto puede dificultar el uso de la teoría de redes abiertas de *Jackson*. Como se verá, bajo situaciones razonables (tiempos de *standby* pequeños), los resultados siguen siendo válidos y cercanos al rendimiento real.

IV. Validación del modelo

En esta sección se validará el modelo descrito anteriormente, comparando los resultados teóricos con los obtenidos mediante simulación. Para ello se hará uso de un simulador por eventos implementado en `C++`, que ha sido implementado *ad-hoc*. La razón de utilizar una nueva herramienta en lugar de soluciones existentes (p.e. `ns-3`) es que el objetivo es la validación del modelo, por lo que se ha buscado tener un mayor control sobre el

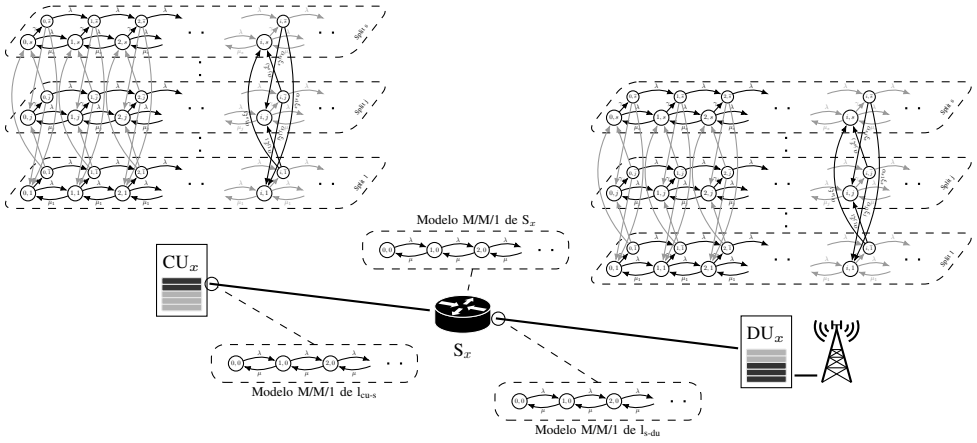


Fig. 2: Modelo extremo a extremo basado en cadenas de *Markov*

comportamiento simulado y evitar la complejidad añadida por soluciones más completas, tales como lógica de protocolos, que tendrían algún impacto en los resultados. A modo de resumen, el simulador implementa los dos tipos de nodo utilizados (M/M/1 y QBD) y cuatro tipos de eventos. Todos los nodos gestionan dos clases de eventos: (1) llegada de una trama y (2) finalización de procesado de trama. Además, en los nodos QBD hay otros dos tipos de eventos: (3) cambio de *split* y (4) finalización de *stanby*. Se pueden configurar varios flujos de entrada, y la matriz de encaminamiento indica las rutas que las tramas pertenecientes a estos flujos siguen. En la Tabla II se muestran los parámetros de configuración utilizados en todos los escenarios. Se han considerado 4 niveles de *split* ($s = 4$), con tasas de servicio $\mu_{1,2,3,4} = \{1, 1.5, 2, 4\}$ ms^{-1}. Estos valores se han seleccionado para ilustrar el potencial del modelo, y reflejan diferentes capacidades de procesado de tramas de los niveles de *split*.

Por otro lado, el tiempo medio de permanencia en cada *split* es $\gamma_{1,2,3,4} = \{\frac{1}{100}, \frac{2}{100}, \frac{3}{100}, \frac{4}{100}\}$ ms^{-1}. Como se puede ver, las tasas para las DUs son "complementarias", ya que el procesado total ha de repartirse entre la CU y la DU. Además, se asume que $\xi_j = \xi \ \forall j$ de modo que el tiempo de *standby* es el mismo para todos los *splits*. La matriz A establece las probabilidades de selección del siguiente nivel de *split*, siendo la probabilidad de transitar al mismo estado $\alpha_{i,i} = 0$, y asegurando que una vez iniciado el cambio de *split* este finalice, $\sum_{j=1^s} \alpha_{i,j} = 1$. Como se puede observar en la Tabla II la matriz A en los DUs también es la "complementaria" de la correspondiente a los CUs, para reflejar los cambios de *split* correspondientes.

En cuanto a los nodos M/M/1, utilizados para modelar los *switches* y enlaces, se han definido varias tasas de servicio, para reflejar diferentes situaciones y tecnologías. Inicialmente la tasa de servicio de los *switches* será $\mu_n = 5\ ms^{-1}$ y se reducirá a 3 ms^{-1} en el último escenario. Asimismo, las tasas de los enlaces representan dos tecnologías: fibra óptica con una tasa de $\mu_{\text{of}} = 8\ ms^{-1}$ y ondas milimétricas, cuya tasa μ_{mmw} se variará (1, 2, 4, 6 ms^{-1}) para analizar su impacto.

Tabla II: Configuración del escenario

Nodos CU y DU	
Tasas de servicio	$\mu = \{1, 1.5, 2, 4\}$ (ms^{-1})
Tasas de cambio de *split*	$\gamma_{\text{cu}} = \{\frac{1}{100}, \frac{2}{100}, \frac{3}{100}, \frac{4}{100}\}$ (ms^{-1})
	$\gamma_{\text{du}} = \{\frac{4}{100}, \frac{3}{100}, \frac{2}{100}, \frac{1}{100}\}$ (ms^{-1})
Duración de *standby*	$\xi^{-1} = 1, 5, 10, 20, 50$ (ms)
Probabilidades de transición entre *splits*	$A_{\text{cu}} = \begin{pmatrix} 0 & 0.6 & 0.2 & 0.2 \\ 0.1 & 0 & 0.3 & 0.6 \\ 0.3 & 0.3 & 0 & 0.4 \\ 0.2 & 0.3 & 0.5 & 0 \end{pmatrix}$
	$A_{\text{du}} = \begin{pmatrix} 0 & 0.5 & 0.3 & 0.2 \\ 0.4 & 0 & 0.3 & 0.3 \\ 0.6 & 0.3 & 0 & 0.1 \\ 0.2 & 0.2 & 0.6 & 0 \end{pmatrix}$
Red *fronthaul*	
Tasas de servicio de los *switches*	$\mu_n = 5, 3$ (ms^{-1})
Tasa de servicio de enlaces de fibra óptica	$\mu_{\text{of}} = 8$ (ms^{-1})
Tasa de servicio de enlaces mmWave	$\mu_{\text{mmw}} = 1, 2, 4$ (ms^{-1})

A. *Nodos CU/DU*

En el primer escenario a evaluar se validará el comportamiento de los nodos CU y DU. Se usará la configuración indicada en la Tabla II y se estudiará el tiempo de permanencia en el nodo al incrementar la tasa de entrada para los diferentes valores de tiempo de *standby*. En la Figura 3 se muestran los resultados teóricos con línea continua, y los obtenidos con el simulador con marcadores. Los valores de simulación se han obtenido a partir de 100 simulaciones independientes, en cada una de las cuales se han generado 10^6 tramas, para asegurar resultados estadísticamente fiables. En primer lugar, se puede observar que con el modelo teórico se obtienen valores casi idénticos a los simulados, lo que permite validar el modelo de los nodos CU/DU, así como la correcta implementación del simulador. Por otro lado, los resultados también indican que el tiempo de *standby* tiene un gran impacto, ya que el tiempo medio de permanencia crece de forma acusada al aumentar el valor de ξ^{-1}. Merece la pena indicar que en sistemas reales, es esperable que el tiempo de *stanby* necesario para la reconfiguración de las estaciones base sea varios órdenes de magnitud menor que el de permanencia en cada uno de los *split*,

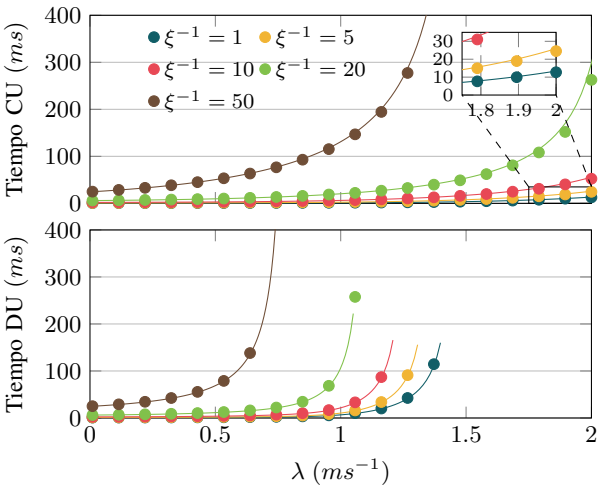


Fig. 3: Tiempo de permanencia en los nodos CU/DU incrementando la tasa de entrada λ y con diferentes tiempos de *standby*

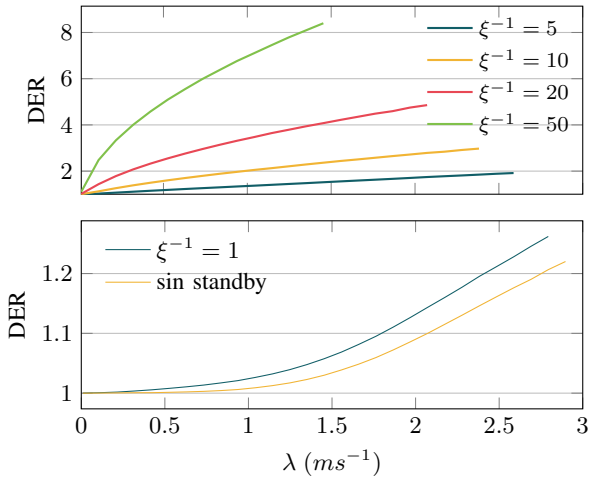


Fig. 4: Desviación estándar relativa (DER) del tiempo entre salidas en el CU para diferentes valores de la tasa de entrada y tiempos de *standby*

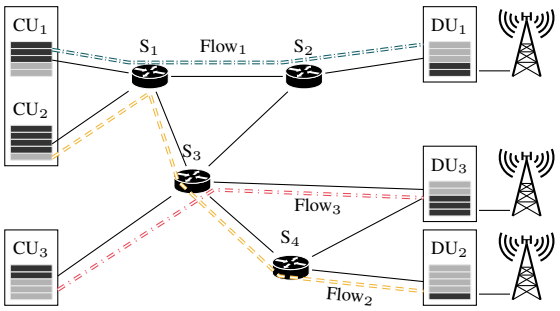


Fig. 5: Red *fronthaul* para validar el modelo extremo a extremo

como se ha visto en [5].

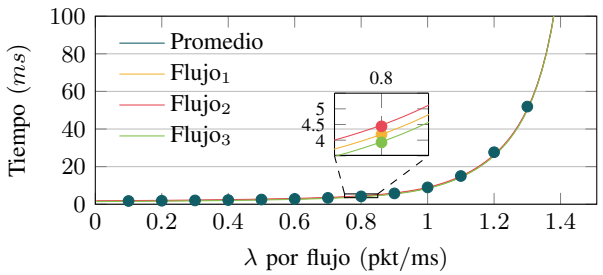
Como se ha mencionado previamente, para utilizar la teoría de redes abiertas de *Jackson* en la caracterización del retardo extremo a extremo se requiere que se cumpla el teorema de Burke, de modo que el proceso de tráfico a la salida sea estadísticamente idéntico al de la entrada [21], [23]. Por ello, se necesita asegurar que el tráfico de salida en los CU sea un proceso de *Poisson*, o de otro modo, que el tiempo entre salidas consecutivas sigue una distribución exponencial. Incluso si el tráfico de entrada sea tal que se asegure la estabilidad del sistema, definida por la ecuación (9), podría haber circunstancias en las que el teorema de Burke no se cumpliera. Por ello, se debe asegurar que: (i) el tráfico de entrada sea menor que la tasa de servicio del *split* más lento y (ii) que el tiempo de *standby* pueda considerarse despreciable en comparación con los tiempos de permanencia en los *split*.

A fin de analizar si estas dos condiciones han de respetarse de manera estricta, se ha usado el simulador para estudiar el tiempo entre salidas en el CU. En la Figura 4 se muestra la desviación estándar relativa (DER) de estos tiempos, que se define como la relación entre su desviación estándar y su media. Si la salida del nodo CU fuera un proceso de *Poisson*, la DER debe tomar valor 1. Se puede observar que la DER es notablemente mayor que 1 cuando el tiempo de *standby* es alto, por lo que en esas circunstancias el proceso de salida del tráfico no podría ser considerado de *Poisson*, incluso cuando la tasa de entrada está por debajo de su posible valor máximo, aquel que asegura estabilidad. Por otro lado, cuando el valor del tiempo de *standby* es menor, la DER está muy próxima a la unidad. Dado que esta situación es la más verosímil, se puede considerar que en condiciones realistas el tráfico a la salida del CU se corresponderá con un proceso de *Poisson*, y que por lo tanto el modelo presentado será válido.

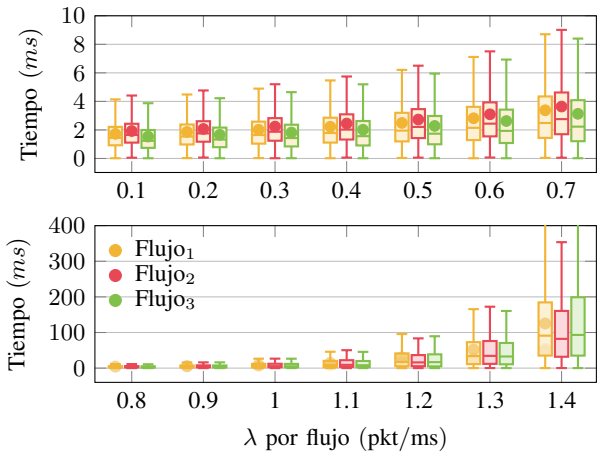
B. Retardo extremo a extremo

A continuación se analizará el retardo extremo a extremo esperado sobre el escenario representado en la Figura 5 que incluye tres pares CU/DU y cuatro *switches*. Sobre este escenario se establece un flujo entre cada par CU/DU, el cual sigue la ruta mostrada en la figura: (i) $CU_1 \rightarrow S_1 \rightarrow S_2 \rightarrow DU_1$; (ii) $CU_2 \rightarrow S_1 \rightarrow S_3 \rightarrow S_4 \rightarrow DU_2$; (iii) $CU_3 \rightarrow S_3 \rightarrow DU_3$.

En primer lugar se asume que todos los enlaces tienen alta capacidad (fibra óptica), por lo que únicamente se incluye en la evaluación del retardo el efecto de los nodos CU/DU y de los *switches*. En la Figura 6a se muestran los retardos medios de cada flujo y el promedio global, al incrementar el valor de la tasa de entrada. Los resultados teóricos se han obtenido con las ecuaciones (13) y (14), y se han comparado con los obtenidos mediante simulación. Nuevamente, para cada configuración se han realizado 100 simulaciones independientes generando en cada una de ellas 10^6 tramas. Como se puede observar, nuevamente los resultados teóricos son casi idénticos a los simulados, aumentando el retardo en ambos casos al incrementar la tasa de entrada. También se puede observar que el modelo teórico proporciona resultado prácticamente idénticos, en su valor medio, a los simulados, incluso cuando el tráfico de entrada es superior a la tasa de servicio del *split* más lento (1 pkt/ms), lo que implica que no se



(a) Retardo medio



(b) Distribución del retardo

Fig. 6: Retardo extremo a extremo al incrementar λ por flujo. La figura superior muestra el retardo medio y la inferior representa la variabilidad de los resultados obtenidos.

cumplen estrictamente los requisitos para aplicar la teoría de *Jackson*.

Usando el simulador se puede extender el análisis para conocer, no solo los valores medios, sino la distribución del retardo, que puede tener un impacto notable en el rendimiento de los servicios. En la Figura 6b se usan diagramas de caja (*boxplots*) para representar la variabilidad del retardo por flujo para varios valores de tasa de entrada (λ). Cada diagrama indica la mediana (percentil del 50%) con una línea horizontal, así como los percentiles del 25 y 75%, que corresponden a los límites de la cada. Por otro lado, las líneas superior e inferior indican los percentiles del 5 y 95%. También se indica en cada caja el valor medio mediante un marcador. Como se puede observar el retardo crece al aumentar la tasa de entrada, tal como se vio anteriormente. Estos resultados muestran que para tasas de entrada bajas el retardo máximo se encuentra por debajo de 10ms, el cual aumenta bruscamente cuando la tasa de entrada supera la tasa de servicio del *split* más lento (1 pkt/ms).

En el siguiente escenario se fija la tasa de los flujos 1 y 3 (f_1 y f_3) a 0.8 pkt/ms, y se va incrementando la correspondiente al flujo 2 (f_2). Como se puede apreciar en la Figura 5, f_2 atraviesa los *switches* S_1, que también es usado por f_1, y S_3, que se comparte con f_3. En la Figura 7 se muestra el retardo medio extremo a extremo. Al igual

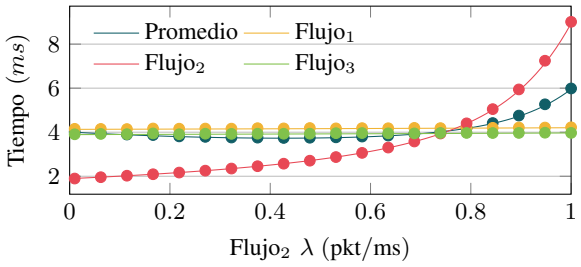


Fig. 7: Retardo extremo a extremo incrementando λ_{f_2}

que en los resultados anteriores los valores teóricos se indican con línea continua, mientras que las resultados obtenidos en la simulación se representan con marcadores, que indican el valor medio tras 100 experimentos independientes. En este caso, dado que las tasas de los flujos son diferentes, cada simulación genera tramas hasta asegurar que el flujo con menor tasa envía 10^6 tramas, y que el resto no ha dejado de generar tráfico, para que las condiciones de la red no cambien a lo largo de cada experimento. Además de comprobarse nuevamente que los resultados de la simulación y teóricos son casi idénticos, se puede observar que el incremento de la tasa λ_{f_2} prácticamente no influye en los otros flujos con los que comparte *switch*, ya que los *switches* en este escenario tienen poca ocupación en relación a su capacidad. Por otro lado, los resultados también muestran el incremento del retardo de f_2 aumentar su tasa, por lo que se deduce que, con esta configuración, el retardo es debido principalmente al procesado en los nodos CU/DU.

C. Impacto de estrategia de encaminamiento y enlaces heterogéneos

Se analiza a continuación el impacto sobre el retardo al modificar la tecnología de los enlaces que conforman la red que se está analizado (Figura 5), así como al adaptar la configuración de encaminamiento. Se asume que todos los enlaces tienen capacidad alta ($\mu_{\text{fo}} = 8\ ms^{-1}$), excepto el que conecta los *switches* S_1 y S_2, que emula un enlace mmWave. Bajo estas condiciones se ha variado la política de encaminamiento de S_1, de modo que con probabilidad φ se usa el camino corto (atravesando el enlace entre S_1 y S_2) y con probabilidad $1-\varphi$ el tráfico se reenvía por la siguiente ruta: $\text{CU}_1 \rightarrow S_1 \rightarrow S_3 \rightarrow S_2 \rightarrow \text{DU}_1$. La Figura 8 muestra que el retardo medio global (considerando todos lo flujos) varía a medida que se modifica el valor de φ. Las tasas para todos los flujos son 0.8 pkt/ms, y los resultados se representan como en las figuras anteriores. Los valores que se obtienen a través del simulador también se han obtenido de 100 simulaciones independientes, en las que se han generado 10^6 tramas por flujo. Los resultados muestran que la estrategia de encaminamiento, como era de esperar, tiene un impacto evidente en el rendimiento. En este sentido, se puede ver que hay un punto de operación óptimo (respecto a φ) donde el retardo presente el valor más bajo. En concreto, con la configuración descrita, cuando la tasa del enlace entre los *switches* S_1 y S_2 es 1 pkt/ms, el valor que optimiza el rendimiento es $\varphi \approx 0.6$.

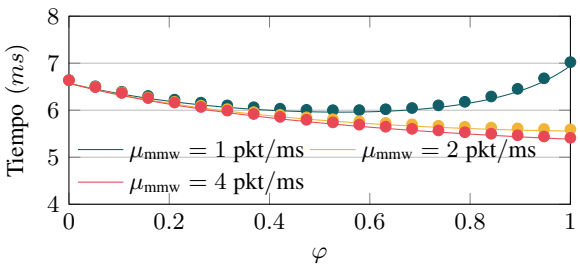


Fig. 8: Retardo extremo a extremo con enlaces heterogéneos y diferentes políticas de encaminamiento

V. Conclusiones

En este artículo se ha presentado un modelo basado en teoría de colas para analizar el rendimiento de la red *fronthaul* en redes de acceso con selección dinámica de *functional split*. El modelo considera diferentes tasas de servicio para cada uno de los niveles de centralización, así como diferentes tiempos de *standby*, que se contempla para las tareas de reconfiguración de los nodos CU y DU al modificar su split. Se plantea el uso de un proceso QBD, cuyo comportamiento se ha obtenido usando el método de la matriz geométrica. También se ha analizado bajo qué circunstancias el modelo de los nodos CU/DU se puede utilizar junto con teoría de redes de *Jackson* para evaluar el retardo extremo a extremo en la red *fronthaul*. Se ha visto que, para regímenes de operación realistas, los resultados obtenidos por el modelo teórico son casi idénticos a los proporcionados mediante simulación.

Posteriormente se ha estudiado el retardo extremo a extremo de la red *fronthaul*, observando nuevamente que los resultados proporcionados por el modelo y los obtenidos mediante simulación son prácticamente idénticos. Finalmente, se ha modificado la configuración del escenario de evaluación para mostrar el potencial del modelo ante diferentes circunstancias. En concreto, se ha analizado el rendimiento de la red al aplicar diferentes políticas de encaminamiento sobre enlaces heterogéneos, poniendo de manifiesto que el modelo puede ser utilizado para obtener puntos óptimos de operación.

Se han identificado dos líneas de trabajo que se abordarán en el futuro. Por un lado, utilizando el simulador se va a analizar el impacto que tiene limitar el tamaño de los *buffer* en los diferentes nodos, así como el efecto de cambiar los patrones de tráfico. Por otro lado, se pretende utilizar el modelo para evaluar diferentes políticas de *split*, y esquemas de gestión de los *buffer*.

Agradecimientos

Los autores agradecen la financiación de Gobierno de España (Ministerio de Economía y Competitividad, Fondo Europeo de Desarrollo Regional, MINECO-FEDER) por medio del proyecto *FIERCE: Future Internet Enabled Resilient smart CitiEs* (RTI2018-093475-AI00).

References

[1] 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.

[2] 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.

[3] L. Diez, C. Hervella, and R. Agüero, "Understanding the performance of flexible functional split in 5g vran controllers: A markov chain-based model," *IEEE Transactions on Network and Service Management*, vol. 18, no. 1, pp. 456–468, 2021.

[4] 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, 2019.

[5] A. M. Alba, J. H. G. Velásquez, and W. Kellerer, "An adaptive functional split in 5g networks," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2019, pp. 410–416.

[6] C.-Y. Chang, N. Nikaein, R. Knopp, T. Spyropoulos, and S. S. Kumar, "Flexcran: A flexible functional split framework over ethernet fronthaul in cloud-ran," in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–7.

[7] 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, 2018.

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

[9] V. Q. Rodriguez, F. Guillemin, A. Ferrieux, and L. Thomas, "Cloud-ran functional split for an efficient fronthaul network," in *2020 International Wireless Communications and Mobile Computing (IWCMC)*, 2020, pp. 245–250.

[10] Y.-T. Huang, C.-H. Fang, L.-H. Shen, and K.-T. Feng, "Optimal functional split for processing sharing based comp for mixed embb and urllc traffic," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.

[11] A. Martinez Alba and W. Kellerer, "A dynamic functional split in 5g radio access networks," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[12] A. Alabbasi, M. Berg, and C. Cavdar, "Delay constrained hybrid cran: A functional split optimization framework," in *2018 IEEE Globecom Workshops (GC Wkshps)*, 2018, pp. 1–7.

[13] T. Ismail and H. H. M. Mahmoud, "Optimum functional splits for optimizing energy consumption in v-ran," *IEEE Access*, vol. 8, pp. 194 333–194 341, 2020.

[14] L. Wang and S. Zhou, "Flexible functional split and power control for energy harvesting cloud radio access networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 1535–1548, 2020.

[15] H. Gupta, M. Sharma, A. Franklin A., and B. R. Tamma, "Apt-ran: A flexible split-based 5g ran to minimize energy consumption and handovers," *IEEE Transactions on Network and Service Management*, vol. 17, no. 1, pp. 473–487, 2020.

[16] S. Zhou, X. Liu, F. Effenberger, and J. Chao, "Mobile-pon: A high-efficiency low-latency mobile fronthaul based on functional split and tdm-pon with a unified scheduler," in *2017 Optical Fiber Communications Conference and Exhibition (OFC)*, 2017, pp. 1–3.

[17] 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)*, 2018, pp. 1–6.

[18] M. P. Amaral, J. Gomes, H. R. O. Rocha, J. A. L. Silva, and M. E. V. Segatto, "Processing resource allocation in 5g fronthaul," in *2019 SBMO/IEEE MTT-S International Microwave and Optoelectronics Conference (IMOC)*, 2019, pp. 1–3.

[19] M. Neuts, "Markov Chains with Applications in Queueing Theory, Which Have a Matrix-Geometric Invariant Probability Vector," *Advances in Applied Probability*, vol. 10, no. 1, pp. 185–212, 1978.

[20] B. Hajek, "Birth-and-death processes on the integers with phases and general boundaries," *Journal of Applied Probability*, vol. 19, no. 3, p. 488–499, 1982.

[21] P. J. Burke, "The Output of a Queuing System," *Operations Research*, vol. 4, no. 6, 1956.

[22] J. R. Jackson, "Jobshop-like Queueing Systems," *Management Science*, vol. 10, no. 1, 1963.

[23] L. Kleinrock, *Queueing Systems. Volume 1: Theory*. Wiley-Interscience, 1975.