Dimensionnement du fronthaul 5G : c'est simple comme un coût de file

Anne Bouillard, Fabien Mathieu, Philippe Sehier, Thomas Deiß

Dimensionnement du fronthaul 5G : c'est simple comme un coût de file

Anne Bouillard[1] †, Fabien Mathieu[1], Philippe Sehier[1] et Thomas Deiß[1] ‡

[1]*Nokia Bell Labs Fance, Villarceaux*

La technologie 5G a pour objectif le déploiement de nouveaux services, ayant des exigences différentes en termes de délai et de bande passante, dans des architectures flexibles. Ceci est rendu possible grâce à la virtualisation des réseaux. En particulier, en fonction des exigences des services, certaines fonctionnalités de la couche physique sont mutualisées et déplacées des antennes vers un contrôleur centralisé (C-RAN) à travers un *fronthaul*. L'agrégation du trafic des antennes permet de tirer partie du multiplexage statistique, mais requiert de fortes contraintes de délai. La décision de centraliser ou pas une fonctionnalité dépend donc des effets en termes de performances du gain de multiplexage et du coût de délocalisation.

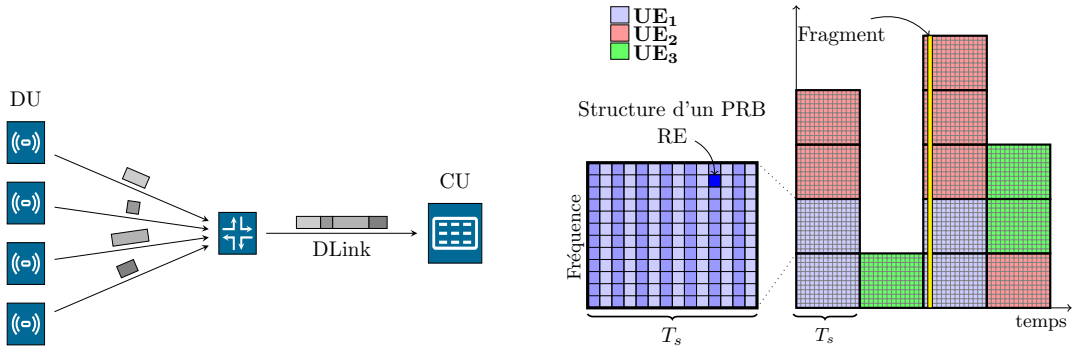
Dans cet article, nous proposons une modélisation du fronthaul par une file D/G/1. La forte corrélation entre les paquets, due au codage radio, est prise en compte. L'analyse de cette file donne des formules de dimensionnement du fronthaul, et permet de quantifier le gain apporté par le multiplexage. Ces résultats sont enfin évalués par simulation.

Une première version de cet article a été présentée à *BackNets 2017* [SBMD17].

Mots-clefs : File D/G/1, 5G, fronthaul, dimensionnement.

1 Modèle du fronthaul 5G

Le réseau fronthaul peut être composé de plusieurs routeurs, et a en général une topologie en arbre dirigée vers le C-RAN. Afin de nous concentrer sur l'impact du multiplexage, nous modélisons ce réseau selon la figure 1a : N_c antennes (*distributed units*, ou DU) sont multiplexées sur un seul lien (DLink). Le trafic ainsi agrégé est servi à taux constant μ par une unité centrale (CU).



(a) Modèle simplifié du fronthaul.

(b) Structure des paquets LTE.

FIGURE 1: Modélisation du fronthaul.

†A. Bouillard et F. Mathieu sont membres du LINCS `www.lincs.fr`

‡T. Deiß est partiellement financé par le projet H2020-ICT-2014-2 *5G-Crosshaul: The 5G Integrated fronthaul/backhaul* (671598).

Modélisation des paquets La structure des paquets LTE est représentée sur la figure 1b. Le bloc de transmission élémentaire est le PRB (*physical resource block*). Chaque PRB est composé de 168 éléments de ressource (RE) distribués sur 12 sous-porteuses et $m = 14$ symboles. Les PRBs sont émis tous les $T_s = 1ms$, et chaque cellule peut émettre au maximum N_p PRBs. Selon [DDM+13], la taille d'un RE est de 16 bits, donc la taille d'un PRB est $Q = 16 \times 168 = 2688$ bits. On appelle *case* (horaire) chaque intervalle de durée T_s. Une case est divisée en m symboles (de temps).

À chaque case, chaque équipement (UE) envoie des PRBs à une antenne (DU), et chaque antenne regroupe ces PRBs (au maximum N_p) en un paquet qui est transmis à la file d'attente tous les symboles de temps : tous les T_s/m un fragment de taille S/m est envoyé, où S est la taille du paquet.

Modèle du trafic On suppose que la distribution du nombre de PRBs envoyés à chaque case est la même pour chaque DU, et que ce nombre de PRB est i.i.d. pour chaque DU et chaque case. On note λ l'intensité du trafic (sous nos hypothèses, $\lambda \leq \Lambda := N_p N_c$). La distribution exacte du trafic est inconnue au delà de son intensité et de son maximum, et pour calculer des bornes de performances conservatives, on peut considérer la *pire* distribution : celle qui envoie N_p PRBs avec probabilité λ/Λ, et 0 PRB sinon. L'analyse qui suit est ainsi valable pour toute distribution des PRBs.

2 Analyse du modèle

Notre objectif est de déterminer la probabilité que le délai de transmission d'un PRB dépasse une certaine valeur critique, afin d'étudier la manière dont cette probabilité varie en fonction des paramètres de notre modèle. Pour faire cela, on calcule d'abord la distribution de la file d'attente.

Soit $A_n^{(i)}$ la taille du paquet émis par la cellule i au début d'une case n et $A_n = \sum_{i=1}^{N_c} A_n^{(i)}$ le nombre d'arrivée total pour cette case. Comme $(A_n^{(i)})$ est i.i.d. sur $\{0,1,\ldots,N_p\}$, (A_n) l'est sur $\{0,1,\ldots,\Lambda\}$. On a de plus $\lambda = \mathbf{E}[A_n]$.

À chaque symbole de temps de la case n, la quantité A_n/m est envoyée dans la file. Malgré cette corrélation, lorsqu'on regarde le système aux instants de fin de case (juste avant l'arrivée du premier fragment du paquet $n+1$), le système suit la dynamique d'une file D/G/1. Plus précisément, si X_n est l'état de la file à la fin de la case n, alors on a

$$\begin{cases} X_0 = 0 \\ X_{n+1} = \max\left(X_n + A_{n+1} - \mu, 0\right). \end{cases} \tag{1}$$

Si $\lambda < \mu$, cette équation définit une chaîne de Markov ergodique qui admet une unique distribution stationnaire, que l'on notera π dans la suite. L'analyse de cette chaîne en suivant le raisonnement de [Kin64] montre que π_k (la probabilité que la taille de la file soit k) décroît exponentiellement avec k.

Délai d'un paquet On s'intéresse maintenant à la probabilité qu'un paquet soit transmis avec un délai supérieur à δ. On suppose que $\delta > 1/m$: le délai cible est plus grand qu'un symbole de temps.

On dit qu'un fragment est *tardif* si son délai de transmission est supérieur à δ. Cela signifie que la taille de la file d'attente est supérieure à $\delta\mu$ après l'arrivée du fragment. Quand un fragment est tardif, on considère que le paquet correspondant est hors-délai (la reconstitution des PRBs est impossible). Comme la même quantité de données arrive à chaque symbole d'une case, si un fragment est tardif, alors au moins l'un des deux fragments parmi le premier et le dernier est également tardif. Cela permet de déterminer les cas où une case n est tardive. Il faut d'abord qu'il y ait eu une émission ($A_n > 0$), puis :

- si $X_n > \mu(\delta - \frac{1}{m})$, le dernier fragment de la case n, arrivé $1/m$ auparavant, voit une occupation supérieure à $\delta\mu$: la case n est tardive ;
- si $X_n \leq \mu(\delta - \frac{1}{m})$:
  - si $A_n > \mu$, la taille de la file augmente à chaque symbole de la case n. Au dernier symbole, la file a une longueur inférieure ou égale à $\delta\mu$, la case n n'est pas tardive ;
  - si $A_n \leq \mu$, alors la taille de la file décroît à chaque symbole, donc la case n est tardive si le premier symbole l'est, c'est-à-dire si $X_{n-1} + A_n/m > \delta\mu$.

En posant $B = \mu(\delta - \frac{1}{m})$ et en notant D_n le délai du dernier paquet introduit dans la file au temps $n-1$, on obtient

$$\{X_n \geq B\} \cap \{A_n > 0\} \subseteq \{D_n \geq \delta\} \subseteq \{X_n \geq B\} \cup \{X_{n-1} + \frac{A_n}{m} \geq \delta\mu \cap A_n \leq \mu\}$$
$$\subseteq \{X_n \geq B\} \cup \{X_{n-1} + \frac{\mu}{m} \geq \delta\mu\}, \text{ donc}$$
$$\mathbf{P}(X_n \geq B, A_n > 0) \leq \mathbf{P}(D_n \geq \delta) \leq \mathbf{P}(X_{n-1} \vee X_n \geq B).$$

On peut transformer le terme de gauche en utilisant l'équation (1) et l'indépendance de X_{n-1} et A_n :

$$\mathbf{P}(X_n \geq B) - \mathbf{P}(X_n \geq B, A_n = 0) = \mathbf{P}(X_n \geq B) - \mathbf{P}(X_{n-1} \geq B + \mu)\mathbf{P}(A_n = 0).$$

Le dernier terme peut être négligé car la distribution de la taille de la file décroît exponentiellement, ainsi que $\mathbf{P}(A_n = 0)$ quand le nombre de cellules devient grand : on peut alors simplifier le terme de gauche par $\mathbf{P}(X_n \geq B)$.

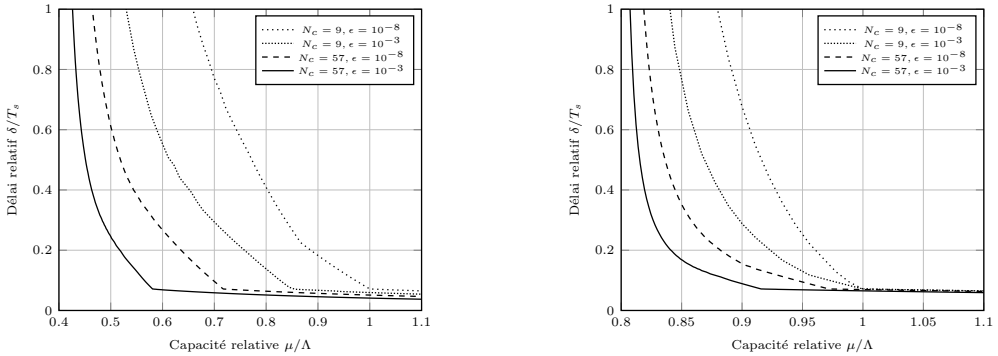
3 Évaluation numérique

Nous utilisons la méthodologie suivante : d'abord, la distribution A_n est calculée en convolant la distribution bimodale N_c fois. Puis, nous estimons la distribution stationnaire π de X_n en utilisant l'équation (1). Pour cela, on procède par itérations successives d'une distribution tronquée ($X_n < 1000$) jusqu'à ce que l'écart entre 2 itérations soit inférieur à 10^{-6} en norme 1. Enfin, en combinant A_n et π et en distinguant les cas $A_n > \mu$, $A_n < \mu$ et $A_n = 0$, on déduit la probabilité qu'une case soit tardive. Cette approche nous permet de calculer rapidement en fonction des paramètres le plus petit délai δ pour lequel le taux de paquets tardifs est inférieur à un taux de pertes autorisées ε (la transmission hors-délai est assimilée à une perte).

Impact du taux de service μ Les figures 2a et 2b montrent δ en fonction de la capacité relative μ/Λ pour $\lambda = 0.4\Lambda$ et $\lambda = 0.8\Lambda$. On observe les phénomènes suivants :

— De manière prévisible, le délai augmente de manière si l'on cherche un taux de pertes très faible ($\varepsilon = 10^{-8}$) ;

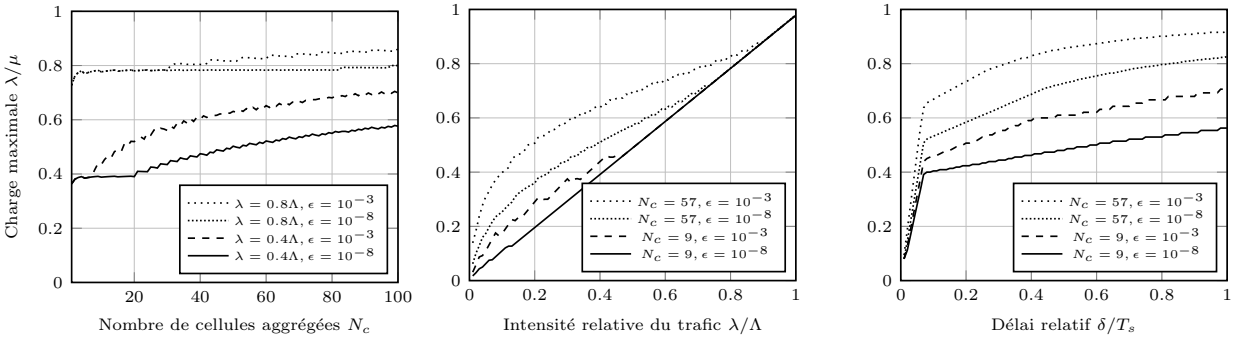
— le multiplexage statistique a un impact significatif qui permet de viser un taux de pertes plus faible. Par exemple, le délai avec $N_c = 57$ et $\varepsilon = 10^{-8}$ est plus faible qu'avec $N_c = 9$ et $\varepsilon = 10^{-3}$.



(a) $\lambda = 0.4\Lambda$ (b) $\lambda = 0.8\Lambda$

FIGURE 2: Délai atteint par une fraction $1-\varepsilon$ de cases non vides avec N_c cellules agrégées.

Atteindre un délai δ Fixons maintenant δ et ε et intéressons-nous à la charge maximale possible pour que le délai δ ne soit dépassé que pour une proportion au plus ε des cases. La figure 3a montre l'impact du nombre de cellules : il y a un seuil au-delà duquel le gain est visible et cet impact est plus important pour une charge faible des cellules ($\lambda = 0.4\Lambda$). La figure 3b montre l'impact de l'intensité du trafic. L'impact est d'autant plus fort que le nombre de cellules est élevé. Enfin, la figure 3c montre l'impact du choix de δ. Il apparaît clairement que cibler une latence inférieure à $1/m$ (un symbole de temps) dégrade énormément les performances.



(a) Impact du multiplexage (b) Impact de l'intensité du trafic (c) Impact du délai cible

FIGURE 3: Charge maximale pouvant atteindre un délai cible en fonction des paramètres. Valeurs par défaut : $\delta = 0,07T_s, \lambda = 0,4\Lambda$.

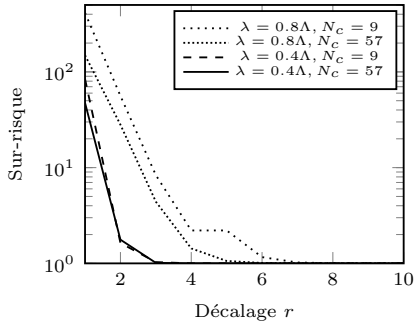


FIGURE 4: Sur-risque d'un paquet tardif après émission d'un paquet tardif ($\delta = 0.14T_s, \varepsilon = 10^{-3}$).

Rafales de paquets tardifs La perte de plusieurs cases consécutives peut être dommageable. Pour conclure, nous étudions donc à quel point une case tardive, par son occupation de la file, augmente les chances que les cases suivantes soient tardives. La figure 4 montre le coefficient multiplicateur de sur-risque qu'une case $n+r$ soit tardive sachant que la case n est tardive (voir détails dans [SBMD17]). On observe que le sur-risque est important pour la case qui suit immédiatement une case tardive, mais qu'il décroît rapidement et devient négligeable au bout de quelques cases de décalage. Le paramètre le plus impactant semble être l'intensité de trafic relative au niveau des cellules, λ/Λ : plus elle est proche de 1, plus il y a un sur-risque de larges rafales.

Références

[DDM+13] U. Doetsch, M. Doll, H.-P. Mayer, F. Schaich, J. Segel, and P. Sehier. Quantitative analysis of split base station processing and determination of advantageous architectures for LTE. *Bell Labs Tech. J.*, May 2013.

[Kin64] J. F. C. Kingman. A martingale inequality in the theory of queues. *Proceedings of the Cambridge Philosophical Society*, 60 :359, 1964.

[SBMD17] P. Sehier, A. Bouillard, F. Mathieu, and T. Deiß. Transport Network Design for FrontHaul. In *3rd IEEE Workshop on Next Generation Backhaul/Fronthaul Networks* , 2017.