On the design and performance of scheduling policies exploiting spatial diversity for URLLC

A. Chagdali, Salah Eddine Elayoubi, Antonia Maria Masucci, A. Simonian

On the design and performance of scheduling policies exploiting spatial diversity for URLLC

A. Chagdali[a,b], S. E. Elayoubi[a], A. M. Masucci[b,*], A. Simonian[b]

[a] *CentraleSupélec, 3 Rue Joliot Curie, 91190, Gif-sur-Yvette,*
[b] *Orange, 44 Avenue de la République, 92320, Chatillon,*

Abstract

In this paper, we study the performance of packet scheduling schemes for Ultra-Reliable Low-Latency Communications (URLLC) services. We exploit spatial diversity (i.e., redundant coverage of users) guaranteed in numerous 5G radio access network scenarios to examine the impact of multi-connectivity. Thus, we consider a set of URLLC users connected to two frequency layers or Radio Access Technologies (RATs) to ensure minimal queuing time. We review four packet scheduling and redundancy schemes, namely Join-the-Shortest-Queue (JSQ), the shortest expected delay (SED), systematic Redundancy (RED), and redundancy with Cancellation upon completion (CAN). We choose the outage probability as a metric, defined as the packet's probability of arriving after some given target delay. We show that RED performs well at low load, whereas JSQ and SED are better when the load rises. Besides, CAN outperforms all other schemes. We then discuss the trade-off between performance and implementation complexity.

Keywords: 5G, URLLC, Resource Allocation, Redundancy, Systems of Parallel Queues.

1. Introduction

The fifth generation of mobile networks (5G) marks a paradigm shift compared to previous generations. While the latter has focused on human-centric

*Corresponding author
Email address: antoniamaria.masucci@orange.com (A. M. Masucci)

and data-driven applications, 5G forecasts the deployment of novel mission-critical use cases tackling autonomous driving, remote surgery, and industrial automation, among others. These services belong to the Ultra-Reliable Low-Latency Communications (URLLC) family of services, expected to prompt many challenges on the current capacity-centered network given the stringent latency and reliability constraints.

Providing Ultra-Reliable Low-Latency Communications (URLLC) services, in particular, will instigate many challenges on the current capacity-centered network, because of the stringent latency and reliability constraints needed. Bearing in mind that a considerable part of the End-to-End (E2E) latency stems from the air interface, addressing radio resource allocation and packet scheduling is of utmost importance to respect the negotiated service level agreement (SLA) for URLLC. For instance, the reliability requirement for the transmission of a URLLC packet of size 32 bytes should be $1 - 10^{-5}$, with a data plane latency of 1 ms ([1, 2]). The reliability here refers to the ability to decode a large proportion of packets before the expiration of the latency budget, any packet received behind the target being considered as in outage or, equivalently, as lost.

5G New Radio (NR) is designed to provide the necessary network adaptability, offering flexible numerology, new frequency bands and sub-carrier spacing, along with mini-slots [3]. These technologies will reduce the air interface latency, especially for URLLC-driven configurations that allow short packets transmission, faster encoding and decoding times, flexible frame structure as well as instant and reservation-based scheduling [4].

The above-quoted 5G NR technologies ensure that a packet is quickly transmitted provided that resources are available and allocated to the device; otherwise, the scheduling and queuing delays may have a drastic impact on the E2E latency. Most of the literature on URLLC seems to disregard the queuing effect, either because it considers cyclic resource reservation for each user [5], or because it privileges a contention-based approach with no queuing [6]. The former approach is ineffective for highly random packet generation scenarios where a per-user cyclic reservation leaves the resources empty almost all the time. The latter contention-based approach is used for the uplink where waiting for a scheduling grant may be prohibitive, but it is not suitable for the downlink where the base station can provide orthogonal resources to the packets. When dealing with sporadic traffic in the downlink, fast and agile packet scheduling techniques are therefore essential for minimizing queuing delays and ensuring Quality of Service (QoS) targets, as will

be addressed in this paper.

When the amount of resources reserved for URLLC is limited, redundant scheduling over several available resources is a practical way for reducing queuing delays, since 5G is specifically designed to integrate multiple Radio Access Technologies (RAT), including 5G NR with several frequency bands, 4G and WiFi interfaces. In practice, a redundant coverage is ensured in almost all locations, especially in dense areas [7]; exploiting the presence of several base stations covering the same device is thus a way to lower the scheduling and queuing delays, by dynamically selecting the base station with the smallest instantaneous load or replicating the packet on several base stations.

In this paper, we exploit redundancy models addressed in the literature for cloud computing applications. These models duplicate arriving packets to a subset of available servers, chosen uniformly or following some load-based criterion. For instance, redundancy-d models uniformly chose $d \geqslant 1$ servers among the available ones. In contrast, Join-the-Shortest-Queue (JSQ) model corresponds to selecting the server that has the lowest number of packets waiting in his queue [8, 9]. [10] examines a queuing system that follows a routing policy where customers join the queue that ensures the Shortest Expected Delay (SED). For the redundancy models, three variants are possible. The basic variant serves all packet copies, and the "cancel-on-start" one consists in immediately deleting all remaining packet copies from other queues as soon as a copy is being served, whereas the "cancel-on-complete" one waits until the first copy has been completed [11].

While the literature on scheduling models with multiple servers is already rich, their focus is in general on the average service time of packets [11, 12]. This metric is not relevant for URLLC and, instead, the delay percentile has to be computed and compared for different schemes. In this paper, we thus consider the practical scenario of URLLC devices covered by two base stations belonging to different RATs and compare the different packet dispatching and redundancy policies, including JSQ, SED and Redundancy with and without cancellation upon completion. Our performance models are based on the aforementioned models which we complete by the evaluation of the reliability metrics.

This paper is an extension of [13], differing in the fact that it covered the homogeneous case (i.e., both base stations have the same service rates), whereas, in this paper, we study the heterogeneous case. In [13], we derived the decay rates for the tail of the sojourn time distribution to approximate

the outage probability, while here we base our analysis of JSQ, SED and Redundancy on equilibrium equations. The paper then makes the following contributions:

- we develop packet scheduling schemes for URLLC exploiting the integration of two RATs within the 5G RAN;
- we utilize theoretical and simulation results to evaluate the reliability metric for the different scheduling schemes.
- we discuss the trade-off between performance and implementation complexity of the analyzed schemes.

This paper is organized as follows. In Section 2, we introduce our system model and present four scheduling options for URLLC. Section 3 focuses on the performance evaluation of the allocation schemes using outage probability as a metric by the exploitation of the delay distribution found based on the resolution of the equilibrium equations or by simulation. Discussion about the adequate resource allocation policy to select is provided in Section 4, a resource dimensioning framework is provided in Section 5, and the paper is concluded in Section 6.

2. System Model

2.1. Resource allocation schemes

We consider the downlink of a wireless system with a set of URLLC users located within an area served by two RATs. The RATs may belong to different Infrastructure Providers (InP's) and can serve dynamically packets belonging to URLLC users. A possible architecture that allows this dynamic service of packets is the one depicted in Figure 1, where an entity connected to both base stations is accountable for the choice of the scheduling policy by either dispatching or duplicating packets. This dynamic packet scheduling is performed based on the instantaneous system state, following the policies outlined below.

Provided that network slicing, considered as a crucial enabler for URLLC use cases, is implemented, the RAN slice manager (defined in 3GPP as the Network Slice Subnet Management Function (NSSMF) [14]) is responsible for this dynamic management. It receives the packets from the vertical's application server and sends them to the base station's schedulers based on

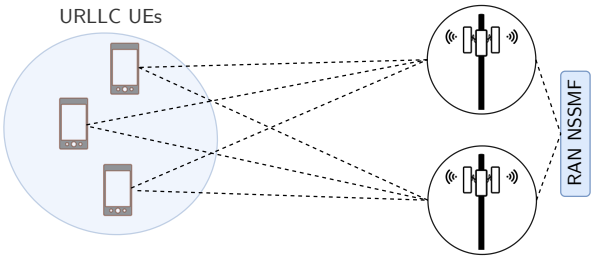

Figure 1: In the neighborhood of URLLC user equipment, two base stations connected to RAN NSSMF responsible for dispatching/duplicating packets.

the chosen policy. This decision depends on the periodical updates received from the base stations' schedulers regarding their load status.

As stated in [2], 3GPP decided that the average user plane latency for URLLC it takes to successfully deliver an application layer packet/message via the radio interface in both uplink and downlink should be 0.5 ms. When a packet belonging to a URLLC device arrives at the scheduler, different policies can be applied, for instance:

• **Join-the-Shortest Queue discipline:** the first scheme consists of sending the incoming URLLC packet to the queue with the least number of waiting packets. If both BS's are empty or have the same number of waiting packets, packets are equally likely to join either BS;

• **Shortest expected delay discipline:** This scheduling policy assigns an arriving customer to the queue that has the shortest expected delay, where delay refers to the sojourn time (waiting time plus the service time).

• **Redundancy discipline:** each incoming packet is independently duplicated in both queues. This scheme does not require any prior knowledge of the radio access channel; thus, extensive control plane information is not required;

• **Redundancy with Cancellation discipline:** as in the previous case, the scheduler sends the incoming packet to both BS's. This scheme necessitates eliminating the remaining copy, provided that one of the copies has been fully served.

Throughout this paper, we denote the aforementioned schemes by **JSQ**, **SED**, **RED** and **CAN**, respectively.

2.2. Queuing model

We model the network architecture by two parallel queues fed by a Poisson process of URLLC packets with mean arrival rate λ, the size of packets being denoted by W (bytes). Motivated by the flexibility of the 5G NR air interface, we consider a First Come First Serve (FCFS) discipline for each queue. This means that the base station adapts its mini-slot dynamically so that one URLLC packet is served by the base station during one mini-slot [1]. Service times of packets at either queue are assumed to be mutually independent as the two base stations are supposed to use different spectrum bands and to be located in different positions, making the channels independent. As of the distribution of service times, it depends on the MCS used by the UEs, determined based on the instantaneous channel (an MCS corresponds to a service time instance, knowing that URLLC packets are generally of a constant small size). In order to ease the analysis, we assume that the resulting distribution is approximated by an exponential with rate α and β, respectively such that $\alpha \leqslant \beta$ and we set $z = \beta/\alpha$. We will show in the numerical distribution how these rates are obtained using realistic assumptions.

Given these two $M/M/1$ queues coupled by either JSQ, SED, RED or CAN discipline, we denote by M (resp. N) the number of packets in the first (resp. the second) queue. The associated equilibrium distribution of the occupancy vector (M, N) is then defined by $p_{m,n} = \mathbb{P}(M = m,\, N = n)$, $(m, n) \in \mathbb{N}^2$. Following [8], [10], [15] and, [12] respectively, this stationary distribution is then shown to exist provided that

- for JSQ, $\alpha + \beta > \lambda$, that is, $(1 + z)\alpha > \lambda$;
- for SED, $\alpha + \beta > \lambda$, $(1 + z)\alpha > \lambda$;
- for RED, $\alpha > \lambda$;
- for CAN, $\alpha + \beta > \lambda$, that is, $(1 + z)\alpha > \lambda$;

[1] Note that, in cases where the amount of spectral resources is large, and the packet is small, several packets may be multiplexed in the frequency dimension in the mini-slot of smallest size (2 OFDMA symbols). Our assumption of a FCFS rule for each queue then gives an upper bound of the performance, assuming a maximal slot size flexibility.

3. Performance Evaluation

For each of the above allocation schemes, the performance indicator is the outage probability metric $\mathbb{P}(T > t_0)$, where T is the sojourn time of a packet in the system and t_0 is the delay budget. This can be completely characterized by the distribution of T, which is difficult to obtain explicitly for JSQ, SED, and RED. In fact, it is closely related to the occupancy distribution $(p_{m,n})$. To compare the four schemes' respective performance, we first compare the delay distribution obtained using the equilibrium equations to the simulations obtained using a discrete event simulator for JSQ and RED. Then, we compare the JSQ and SED schemes since they are very similar strategies. Throughout the remainder of the paper, we fix $1/\alpha = 0.064$ ms (corresponding to a B=2 MHz system bandwidth, spectral efficiency of e=2 bits/Hz/s, and packets of size W= 32 bytes, so it is straightforward that $\alpha = B \times e/W$). We vary z to account for the impact of heterogeneity on the performance of the scheduling schemes.

3.1. Delay distribution of JSQ scheduling scheme

We consider two M/M/1 ruled by the "Join the Shortest Queue" (JSQ) discipline. An arriving packet joins the shortest queue unless both queues have equal lengths, then he joins the first queue with probability $q^{'} = 1 - q$ and the second queue with probability q, where q is arbitrarily chosen in $[0, 1]$ [16].

The equilibrium equations for $p_{m,n}$ formulated below are found by equating for each state the rate into and the rate out of the same state, where $\kappa = \lambda + \alpha + \beta$.

$$
\begin{aligned}
\kappa p_{m,n} &= \lambda p_{m-1,n} + \alpha p_{m+1,n} + \beta p_{m,n+1} \quad m > 0,\, n > m+1 \\
\kappa p_{n-1,n} &= \lambda p_{n-2,n} + \alpha p_{n,n} + \beta p_{n-1,n+1} + q\lambda p_{n-1,n-1} \quad m > 0,\, n = m+1 \\
\kappa p_{m,n} &= \lambda p_{m,n-1} + \alpha p_{m+1,n} + \beta p_{m,n+1} \quad n > 0,\, m > n+1 \\
\kappa p_{m,m-1} &= \lambda p_{m,m-2} + \alpha p_{m+1,m-1} + \beta p_{m,m} + q^{'}\lambda p_{m-1,m-1} \quad n > 0,\, m = n+1 \\
\kappa p_{n,n} &= \lambda(p_{n-1,n} + p_{n,n-1}) + \alpha p_{n+1,n} + \beta p_{n,n+1} \quad n > 0 \\
(\lambda+\beta)p_{0,n} &= \alpha p_{1,n} + \beta p_{0,n+1} \quad n > 1 \\
(\lambda+\beta)p_{0,1} &= q\lambda p_{0,0} + \alpha p_{1,1} + \beta p_{0,2} \\
(\lambda+\alpha)p_{m,0} &= \alpha p_{m+1,0} + \beta p_{m,1} \qquad m > 1 \\
(\lambda+\alpha)p_{1,0} &= q^{'}\lambda p_{0,0} + \alpha p_{2,0} + \beta p_{1,1} \\
\lambda p_{0,0} &= \alpha p_{1,0} + \beta p_{0,1}
\end{aligned}
\tag{1}
$$

We denote by S and $S^{'}$ the duration of a test job service time at BS1 and BS2, respectively. Given the occupancy vector (M, N), the delay T of a given job is then given by

$$
T = \begin{cases}
S_1 + ... + S_M + S & if M < N \\
S'_1 + ... + S'_N + S' & if N < M \\
S_1 + ... + S_M + S & if M = N, \quad w.p. \quad 1-q \\
S'_1 + ... + S'_M + S' & if M = N, \quad w.p. \quad q
\end{cases}
\tag{2}
$$

All random variables S_1, ..., S_M, S are mutually independent and identically distributed, with exponential distribution with mean $1/\alpha$. in the same manner, all random variables S'_1, ..., S'_N, S' are mutually independent and identically distributed, with exponential distribution with mean $1/\beta$. For all

$t \geqslant 0$, the definition (2) of T thus entails

$$\begin{aligned}\mathbb{P}(T > t) &= \sum_{m,n\geqslant 0} p_{m,n} \cdot \mathbb{P}\left(T > t \mid M = m, N = n\right) \\ &= \sum_{n>m} p_{m,n} \cdot \mathbb{P}(S_1 + ... + S_m + S > t) + \\ &\quad + \sum_{m<n} p_{m,n} \cdot \mathbb{P}(S'_1 + ... + S'_m + S' > t) + \\ &\quad + (1-q) \sum_{m\geqslant 0} p_{m,m} \cdot \mathbb{P}(S_1 + ... + S_m + S > t) + \\ &\quad + q \sum_{n\geqslant 0} p_{n,n} \cdot \mathbb{P}(S'_1 + ... + S'_m + S' > t)\end{aligned}$$

Besides, given m and n, the identical exponential distribution of all variables S_i, $1 \leqslant i \leqslant m$, and S provides

$$\mathbb{P}(S_1 + ... + S_m + S > t) = \sum_{i=0}^{m} e^{-\alpha\, t} \frac{(\alpha\, t)^i}{i!} \tag{3}$$

and similarly for all variables and S'_j, $1 \leqslant j \leqslant n$ and S',

$$\mathbb{P}(S'_1 + ... + S'_n + S' > t) = \sum_{j=0}^{n} e^{-\beta\, t} \frac{(\beta\, t)^j}{j!} \tag{4}$$

So that the latter expression of $\mathbb{P}(T > t)$ further reads

$$\begin{aligned}\mathbb{P}(T > t) = \sum_{m=0}^{\infty} \sum_{n=m+1}^{\infty} p_{m,n} \cdot e^{-\alpha\, t} \sum_{i=0}^{m} \frac{(\alpha\, t)^i}{i!} + \sum_{n=0}^{\infty} \sum_{m=n+1}^{\infty} p_{m,n} \cdot e^{-\beta\, t} \sum_{j=0}^{n} \frac{(\beta\, t)^j}{j!} + \\ (1-q) \cdot \sum_{m\geqslant 0} p_{m,m} \cdot e^{-\alpha\, t} \sum_{i=0}^{m} \frac{(\alpha\, t)^i}{i!} + q \cdot \sum_{n\geqslant 0} p_{n,n} \cdot e^{-\beta\, t} \sum_{j=0}^{n} \frac{(\beta\, t)^j}{j!}, \qquad t \geqslant 0.\end{aligned} \tag{5}$$

Equation 5 requires the resolution of the equilibrium equations. We solve the system of equilibrium equations by adding blocking equations. We consider that all packets arriving while there are L waiting packets as lost. Thus,

we can write the blocking equations as follows:

$$\begin{aligned}
\kappa p_{m,L} &= \lambda p_{m-1,L} + \alpha p_{m+1,L} \qquad 0 < m < L-1 \\
\kappa p_{L-1,L} &= \lambda p_{L-2,L} + \alpha p_{L,L} + q\lambda p_{L-1,L-1} \\
\kappa p_{L,n} &= \lambda p_{L,n-1} + \beta p_{L,n+1} \qquad 0 < m < L-1 \\
\kappa p_{L,L-1} &= \lambda p_{L,L-2} + \beta p_{L,L} + (1-q)\lambda p_{L-1,L-1} \\
(\alpha+\beta) p_{L,L} &= \lambda (p_{L-1,L} + p_{L,L-1}) \\
(\lambda+\beta) p_{0,L} &= \alpha p_{1,L} \\
(\lambda+\alpha) p_{L,0} &= \beta p_{L,1}
\end{aligned} \tag{6}$$

Knowing that the sojourn time of a tagged job at BS1 or BS2 finding k jobs in the system follows an Erlang distribution with shape $k+1$ and rate α or β respectively and keeping in mind that the sojourn time should be less than 0.5 to avoid an outage event. Consequently, we have $(k+1)/\beta \leqslant (k+1)/\alpha < 0.5$ ms. Thus $k < 6.8125$ for BS1 and $k < 14.625$ for BS2. In the numerical application, we choose $L = 20$. Since our goal is to quantify the outage and compare it to a system without blocking, we pick $L > k$.

We solve the linear system described by equations 1 and 6 to find $\{p_{m,n}\}$, $(m,n) \in [0..L]^2$ and use equation 5 to plot an approximation of $\mathbb{P}(T > t_0)$. In Figure 2, we compare the outage probability obtained using equilibrium equations with its counterpart found using discrete-event simulations for different values of z. We focus on traffic regimes where the outage probability is lower than 10^{-5}, defined by 3GPP as a key performance indicator for a plethora of URLLC-centered services [1, 2].

3.2. Delay distribution of SED system

This section analyzes the performance of a system with two servers under the shortest expected delay (SED) scheduling scheme. This policy steers an arriving packet to the queue with the shortest expected delay, where delay refers to the waiting time plus the service time.

Let m and n be the number of customers in the first and second, respectively, including a possible customer in service. For an arriving packet, the expected delay in the first queue is $(m+1)/\alpha$ and in the second queue is $(n+1)/\beta$. The SED scheduling scheme assigns an arriving packet to queue 1 if $(m+1)/\alpha < (n+1)/\beta$ and to queue 2 if $(m+1)/\alpha > (n+1)/\beta$. When the expected delays in both queues are equal, i.e., $\beta(m+1) = \alpha(n+1)$, the arriving customer joins queue 1 with probability q and queue 2 with probability $1-q$.

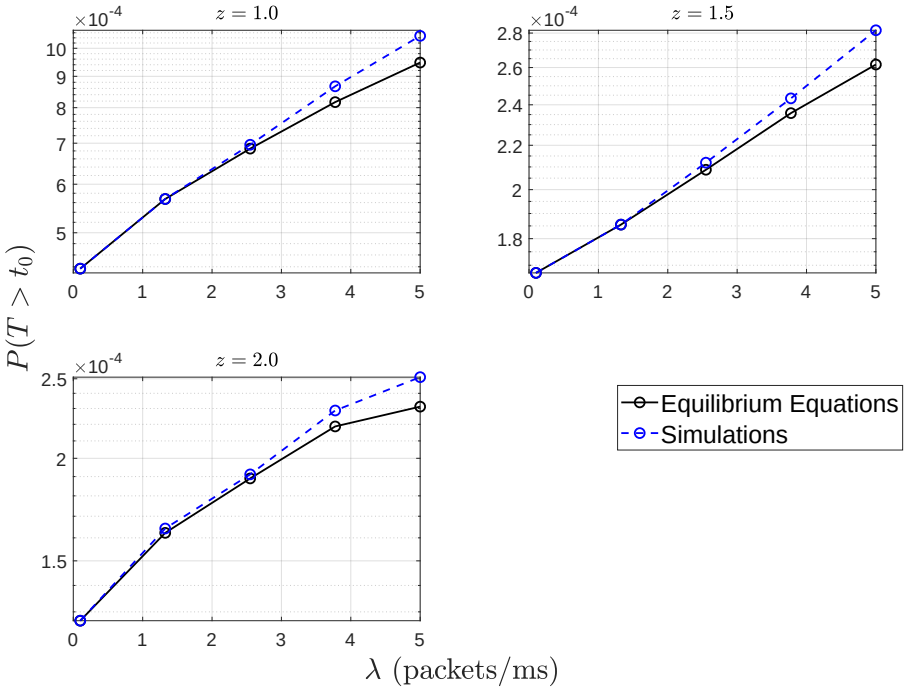

Figure 2: Outage probability for the JSQ scheme using equilibrium equations and simulations with $t_0 = 0.5$ ms, increasing arrival rate lambda for different values of z.

The equilibrium equations for $p_{m,n}$ formulated below are found by equating for each state the rate into and the rate out of the same state, where $\kappa = \lambda + \alpha + \beta$. We write here the equations for the case $z \geq 1$ (similarly, it is possible to write the equations also for the case $z < 1$).

$$
\begin{aligned}
\lambda p_{0,0} &= \alpha p_{1,0} + \beta p_{0,1} \quad m = 0, \quad n = 0 \\
(\lambda+\beta) p_{0,z} &= q\lambda p_{0,z-1} + \alpha p_{1,z} + \beta p_{0,z+1} \quad m = 0, \quad n = z \\
(\lambda+\beta) p_{0,n} &= \lambda p_{0,n-1} + \alpha p_{1,n} + \beta p_{0,n+1} \quad m = 0, \quad n \leq z-1 \\
(\lambda+\beta) p_{0,n} &= \lambda p_{0,n-1} + \alpha p_{1,n} + \beta p_{0,n+1} \quad m = 0, \quad z-1 < n < z \\
(\lambda+\beta) p_{0,n} &= \alpha p_{1,n} + \beta p_{0,n+1} \quad m = 0, \quad n > z \\
(\lambda+\alpha) p_{m,0} &= q^{'}\lambda p_{m-1,0} + \alpha p_{m+1,0} + \beta p_{m,1} \quad n = 0, \quad m = \frac{1}{z} \\
(\lambda+\alpha) p_{m,0} &= \alpha p_{m+1,0} + \beta p_{m,1} \qquad n = 0 \qquad m \neq \frac{1}{z}
\end{aligned}
$$

$$
\begin{aligned}
\kappa p_{m,n} &= \lambda p_{m-1,n} + \alpha p_{m+1,n} + \beta p_{m,n+1} + q\lambda p_{m,n-1} \qquad 0 < m < L, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad n = (m+1)z \\
\kappa p_{m,n} &= \lambda p_{m-1,n} + \lambda p_{m,n-1} + \alpha p_{m+1,n} + \beta p_{m,n+1} \qquad 0 < m < L, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (m+1)z - 1 < n < (m+1)z \\
\kappa p_{m,n} &= \lambda p_{m-1,n} + \alpha p_{m+1,n} + \beta p_{m,n+1} \qquad 0 < m < L, \quad n > (m+1)z \\
\kappa p_{((n+1)\frac{1}{z}-1,n} &= \lambda \left[p_{(n+1)\frac{1}{z}-2,n} + p_{(n+1)\frac{1}{z}-1,n-1} \right] + \alpha p_{((n+1)\frac{1}{z},n} + \\
&\quad + \beta p_{((n+1)\frac{1}{z}-1,n+1} \qquad 0 < n < L, \qquad m = (n+1)\frac{1}{z} - 1 \\
\kappa p_{m,n} &= \lambda p_{m,n-1} + \alpha p_{m+1,n} + \beta p_{m,n+1} + q^{'}\lambda p_{m-1,n} \qquad 0 < n < L, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad m = (n+1)\frac{1}{z} \\
\kappa p_{m,n} &= \lambda p_{m,n-1} + \lambda p_{m-1,n} + \alpha p_{m+1,n} + \beta p_{m,n+1} \qquad 0 < n < L, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (n+1)\frac{1}{z} - 1 < m < (n+1)\frac{1}{z} \\
\kappa p_{m,n} &= \lambda p_{m,n-1} + \alpha p_{m+1,n} + \beta p_{m,n+1} \qquad 0 < n < L, \quad m > (n+1)\frac{1}{z}
\end{aligned}
\tag{7}
$$

Given the occupancy vector (M, N), the delay T of a given job is then given by

$$T = \begin{cases} S_1 + ... + S_M + S & if M < (N+1)\frac{1}{z} - 1 \\ S'_1 + ... + S'_N + S' & if N < (M+1)z - 1 \\ S_1 + ... + S_M + S & if M = (N+1)\frac{1}{z} - 1, \quad w.p. \quad 1-q \\ S'_1 + ... + S'_N + S' & if N = (M+1)z - 1, \quad w.p. \quad q \end{cases} \tag{8}$$

Applying similar reasoning as JSQ, $\mathbb{P}(T > t)$ further reads

$$\begin{aligned} \mathbb{P}(T > t) &= \sum_{m,n \geqslant 0} p_{m,n} \cdot \mathbb{P}\left(T > t \mid M = m, N = n\right) \\ &= \sum_{m=0}^{\infty} \sum_{n=(m+1)z-1}^{\infty} p_{m,n} \cdot e^{-\alpha t} \sum_{i=0}^{m} \frac{(\alpha t)^i}{i!} + \\ &+ \sum_{n=0}^{\infty} \sum_{m=(n+1)\frac{1}{z}-1}^{\infty} p_{m,n} \cdot e^{-\beta t} \sum_{j=0}^{n} \frac{(\beta t)^j}{j!} + \\ &+ (1-q) \cdot \sum_{m \geqslant 0} p_{m,((m+1)z-1)} \cdot e^{-\alpha t} \sum_{i=0}^{m} \frac{(\alpha t)^i}{i!} + \\ &+ q \cdot \sum_{n \geqslant 0} p_{(n+1)\frac{1}{z}-1,n} \cdot e^{-\beta t} \sum_{j=0}^{n} \frac{(\beta t)^j}{j!}, \qquad t \geqslant 0. \end{aligned} \tag{9}$$

We can write the blocking equations as follows:

$$\begin{aligned} \kappa p_{L,n} &= \lambda p_{L,n-1} + \beta p_{L,n+1} + q^{'} \lambda p_{L-1,n} \quad m = L, \quad n = Lz - 1 \\ \kappa p_{L,n} &= \lambda p_{L,n-1} + \beta p_{L,n+1} \quad m = L, \quad n \neq Lz - 1 \\ (q^{'}\lambda + \alpha + \beta) p_{m,L} &= \lambda p_{m-1,L} + \lambda p_{m,L-1} + \alpha p_{m+1,L} \quad n = L, \ m = (L+1)\frac{1}{z} - 1 \\ \kappa p_{m,L} &= \lambda p_{m-1,L} + q\lambda p_{m,L-1} + \alpha p_{m+1,L} \quad n = L, \quad m = \frac{L}{z} - 1 \\ \kappa p_{m,L} &= \lambda p_{m-1,L} + \lambda p_{m,L-1} + \alpha p_{m+1,L} \quad n = L, \\ & \qquad \frac{L}{z} - 1 < m < (L+1)\frac{1}{z} - 1 \end{aligned}$$

$$
\begin{aligned}
\kappa p_{m,L} &= \lambda p_{m-1,L} + \alpha p_{m+1,L} \quad n = L, \quad m < \frac{L}{z} - 1 \\
(\alpha+\beta) p_{m,L} &= q^{'} \lambda p_{m-1,L} + \lambda p_{m,L-1} + \alpha p_{m+1,L} \quad n = L, \quad m = (L+1)\frac{1}{z} \\
(\alpha+\beta) p_{m,L} &= \lambda p_{m-1,L} + \lambda p_{m,L-1} + \alpha p_{m+1,L} \quad n = L, \\
&\qquad\qquad (L+1)\frac{1}{z} - 1 < m < (L+1)\frac{1}{z} \\
(\alpha+\beta) p_{m,L} &= \lambda p_{m,L-1} + \alpha p_{m+1,L} \quad n = L, \quad m > (L+1)\frac{1}{z} \\
(\lambda+\beta) p_{m,L} &= q\lambda p_{m,L-1} + \alpha p_{m+1,L} \quad m = 0, \quad n = L = z \\
(\lambda+\beta) p_{m,L} &= \lambda p_{m,L-1} + \alpha p_{m+1,L} \quad m = 0, \quad z - 1 < n = L < z \\
(\lambda+\beta) p_{m,L} &= \alpha p_{m+1,L} \quad m = 0, \quad n = L > z \\
(\lambda+\alpha) p_{L,0} &= \beta p_{L,1} \quad m = L, \quad n = 0 \\
(\alpha+\beta) p_{L,L} &= \lambda p_{L,L-1} + \lambda p_{L-1,L} \quad m = L, \quad n = L, \quad z = 1 \\
(\alpha+\beta) p_{L,L} &= \lambda p_{L,L-1} + q^{'} \lambda p_{L-1,L} \quad m = L, \quad n = L, \quad z = (L+1)\frac{1}{L} \\
(\alpha+\beta) p_{L,L} &= \lambda p_{L,L-1} + \lambda p_{L-1,L} \quad m = L, \quad n = L, \quad 1 < z < (L+1)\frac{1}{L} \\
(\alpha+\beta) p_{L,L} &= \lambda p_{L,L-1} \quad m = L, \quad n = L, \quad z > (L+1)\frac{1}{L}
\end{aligned}
\tag{10}
$$

We solve the linear system described by the above equilibrium equations to find $\{p_{m,n}\}$, $(m,n) \in [0..L]^2$ and use equation (9) to plot an approximation of $\mathbb{P}(T > t_0)$.

In Figure 3, we plot the outage probability using equilibrium equations and simulations. Similar to the JSQ case, we notice that the outage probability decreases with increasing values of z and increases with rising arrival rates.

Figure 4 compares the outage probability obtained using JSQ and SED to validate the results achieved for SED. First, we notice that, for $z = 1$, the outage probabilities coincide. This behavior is foreseen given that for the homogeneous case, SED is identical to JSQ (i.e., comparing the number of waiting packets is equivalent to compare the expected delay in each queue). On the other hand, for $z = 1.5$ and $z = 2$, SED performs better than JSQ which can be explained by the fact that the JSQ scheme does not exploit the

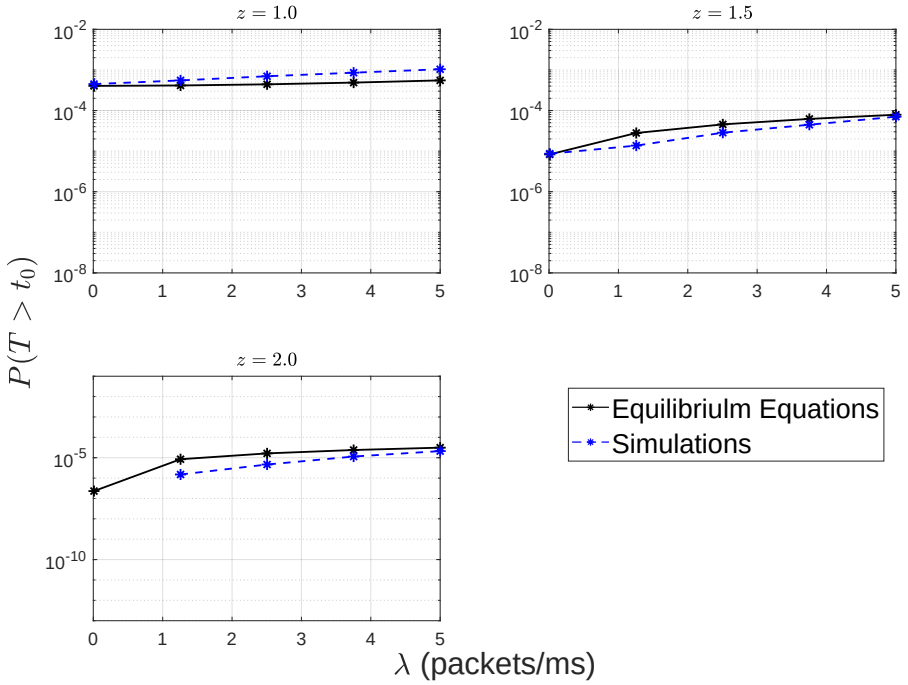

Figure 3: Outage probability for the SED scheme using equilibrium equations and simulations with $t_0 = 0.5$ ms , increasing arrival rate λ and for different values of z.

higher second base station' service rate.

3.3. Delay distribution of the RED system

The systems mentioned above of two coupled queues can be compared to that of two parallel FCFS queues created by arrivals with two demands, as analyzed in [15] and [17]. The incoming packet is duplicated and sent to both queues, where each copy is served independently.

In a similar manner to the JSQ scheme, we formulate the equilibrium equations for $p_{m,n}$ by equating for each state the rate into and the rate out of the same state, where $\kappa = \lambda + \alpha + \beta$.

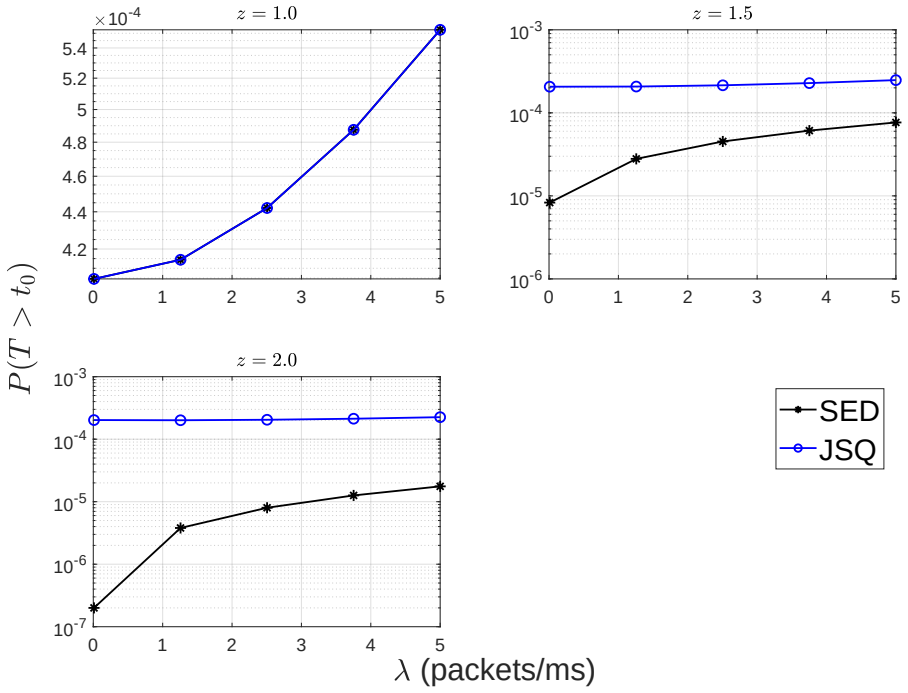

Figure 4: Outage probability comparison of the JSQ and SED schemes using equilibrium equations with $t_0 = 0.5$ ms, increasing arrival rate λ and for different values of z.

$$\begin{aligned} \kappa p_{m,n} &= \lambda p_{m-1,n-1} + \alpha p_{m+1,n} + \beta p_{m,n+1} \quad m > 0, n > 0 \\ (\lambda + \beta) p_{0,n} &= \alpha p_{1,n} + \beta p_{0,n+1} \quad n > 0 \\ (\lambda + \alpha) p_{m,0} &= \alpha p_{m+1,0} + \beta p_{m,1} \quad m > 0 \\ \lambda p_{0,0} &= \alpha p_{1,0} + \beta p_{0,1} \end{aligned} \tag{11}$$

Given the occupancy vector (M, N), the delay T of a given job is given by the minimum

$$T = \min\left(S_1 + ... + S_M + S,\ S'_1 + ... + S'_N + S'\right) \tag{12}$$

where all random variables S_1, ..., S_M, S and S'_1, ..., S'_N, S' are mutually independent and identically distributed, with exponential distribution with mean $1/\alpha$ and $1/\beta$, respectively. Given $t \geqslant 0$, the definition (12) of T entails

$$\begin{aligned}\mathbb{P}(T > t) &= \mathbb{P}\left(S_1 + ... + S_M + S > t,\ S'_1 + ... + S'_N + S' > t\right) \\ &= \sum_{m,n \geqslant 0} p_{m,n}\, \mathbb{P}\left(S_1 + ... + S_m + S > t,\ S'_1 + ... + S'_n + S' > t\right) \\ &= \sum_{m,n \geqslant 0} p_{m,n}\, \mathbb{P}(S_1 + ... + S_m + S > t) \times \mathbb{P}(S'_1 + ... + S'_n + S' > t)\end{aligned}$$

by the independence assumption. Besides, using equations 3 and 4, the latter expression of $\mathbb{P}(T > t)$ further reads

$$\mathbb{P}(T > t) = \sum_{m,n \geqslant 0} p_{m,n} \cdot e^{-(\alpha+\beta)\, t} \sum_{i=0}^{m} \frac{(\alpha\, t)^i}{i!} \sum_{j=0}^{n} \frac{(\beta\, t)^j}{j!}, \qquad t \geqslant 0. \tag{13}$$

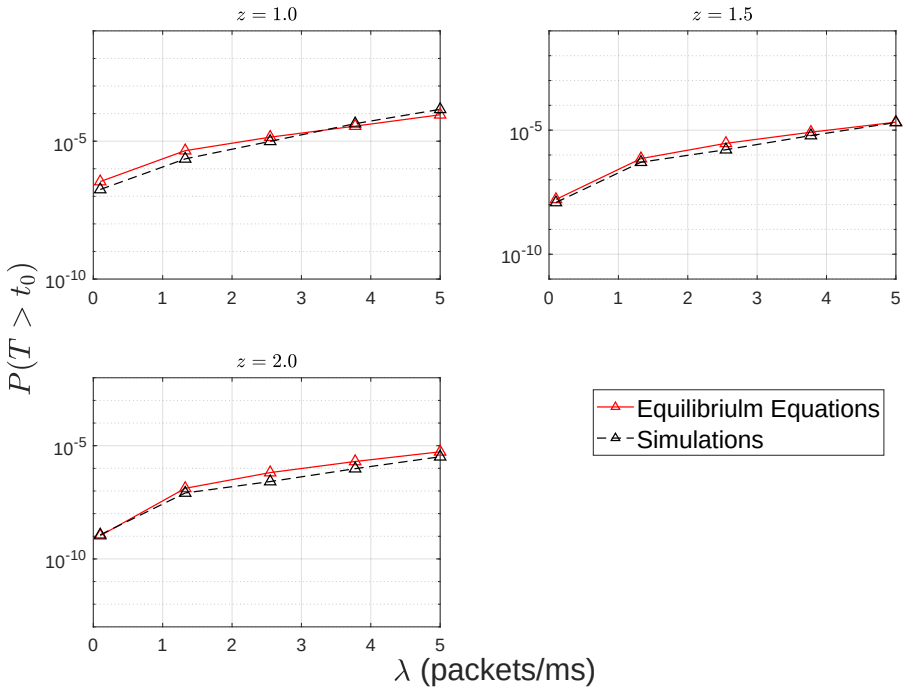

Figure 5: Outage probability for the RED scheme using equilibrium equations and simulations with $t_0 = 0.5$ ms, increasing arrival rate λ and for different values of z.

Applying the same reasoning as the JSQ case, we write the blocking equations as follows:

$$
\begin{aligned}
\kappa p_{L,n} &= \lambda p_{L-1,n-1} + \beta p_{L,n+1} \qquad 0 < n < L \\
\kappa p_{m,L} &= \lambda p_{m-1,L-1} + \alpha p_{m+1,L} \qquad 0 < m < L \\
(\lambda + \beta) p_{0,L} &= \alpha p_{1,L} \\
(\lambda + \alpha) p_{L,0} &= \beta p_{L,1} \\
(\alpha + \beta) p_{L,L} &= p_{L-1,L-1}
\end{aligned} \tag{14}
$$

We solve the linear system described by equations 11 and 14 to find $\{p_{m,n}\}$, $(m,n) \in [0..L]^2$ with $L = 15$. We use equation 13 to plot an approximation of $\mathbb{P}(T > t_0)$. Figure 5 compares the outage probability obtained using equilibrium equations and using discrete-event simulations for different values of z. The outage probability found using the equilibrium equations for a system with blocking represents a lower bound to the outage probability for our system. For the next section, we keep the outage probability found using the equilibrium equations due to the fast computational time compared to simulations.

3.4. Delay distribution of the CAN system

We use [12, Theorem 5], to derive the sojourn time distribution for the CAN case. In [12], the authors study a two-server system, with two classes of jobs (A and B) that do not use redundancy (class A has an arrival rate of λ_A and is allocated to queue 1 of service rate α and class B with arrival rate λ_B is allocated to queue 2 whose service rate is β). If a third class of jobs (R) whose jobs arrive at rate λ is allocated to both servers, with cancel upon completion of the redundant copy, [12, Theorem 5] states that the response time of class R jobs is distributed $Exp(\alpha + \beta - \lambda_A - \lambda_B - \lambda)$. When $\lambda_A = \lambda_B = 0$, the system corresponds to our CAN scheme. Therefore, the distribution of the sojourn time T of a packet is exponentially distributed with rate ξ^* equal to the sum of the service rates at each queue minus the input rate of the class of redundant jobs, that is, $\xi^* = \alpha + \beta - \lambda$ or, equivalently, $\xi^* = (1+z)\alpha - \lambda$ for $\lambda < \alpha + \beta$. Thus, the outage probability $\mathbb{P}(T > t_0)$ corresponding to the delay threshold t_0 can be simply expressed as

$$
\mathbb{P}(T > t_0) = \exp(-\xi^* t_0) \tag{15}
$$

This result underlines the fact that the redundant class experiences a response time distribution identical to that in an M/M/1, even though the system is not an M/M/1 queue.

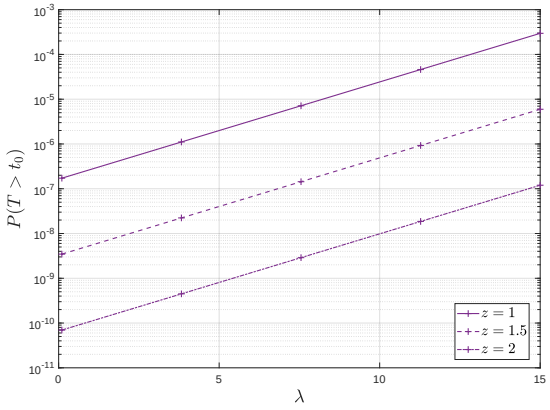

Figure 6: Outage probability $P(T > t_0)$ with $t_0 = 0.5$ ms for CAN increasing arrival rate λ and for different values of z.

Figure 6 plots the outage probability as a function of λ and z using equation 15. Likewise, we see that the outage probability decreases with increasing values of z and increases with growing arrival rates.

4. Numerical experiments

In this section, we assess the performance of the schemes mentioned above by comparing the outage probability for different values of λ and z. As mentioned before, we fix $1/\alpha = 0.064$ ms which corresponds to a 2 MHz system bandwidth, spectral efficiency of 2 bits/Hz/s and packet size of 32 bytes and vary z. Figure 7 plots the outage probability using the equilibrium equations for RED, JSQ and SED, and equation 15 for CAN for a target latency of $t_0 = 0.5$ ms.

As expected, CAN outperforms JSQ, SED, and RED for all values of λ and z. While the outage probability obtained using JSQ coincides with the one achieved SED for $z = 1$, the latter outperforms JSQ for $z = 1.5$ and $z = 2$ for all values of λ. Additionally, RED outperform SED for low arrival rates (for λ less or equal than $\lambda^* = 7$, $\lambda^* = 9$ and $\lambda^* = 11$ packets/ms for $z = 1$, $z = 1.5$ and $z = 2$, respectively). λ^* represents the intersection point between SED and RED. Note that for a target outage 10^{-5}, RED and CAN are more suitable.

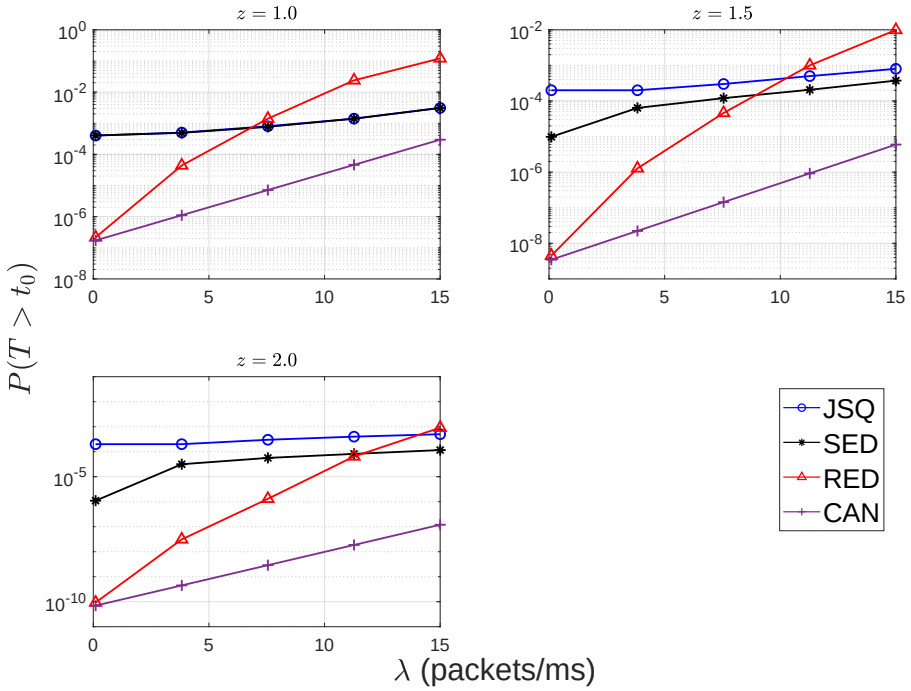

Figure 7: Outage probability $P(T > t_0)$ for JSQ, SED, RED and CAN with $t_0 = 0.5$ ms, increasing arrival rate λ and for different values of z.

4.1. Qualitative Analysis

The decision to deploy one or another scheduling scheme depends on the achieved performance and the feasibility of the different solutions. We omit JSQ from our analysis given that it is equivalent to SED for the homogenous case and substandard for the heterogeneous case. Although the CAN allocation scheme displays good results compared to the RED and SED policies, it remains intricate to implement. First, the base station should notify the RAN NSSMF upon each packet completion. Then, the RAN NSSMF should forward this information to the other BS to remove the remaining copy from the queue. This paradigm requires seamless knowledge about the system and can be achieved if the base stations are co-located. If not, the delay prompted by the communication links may destroy the advantage of CAN and degrade it to a RED scheme.

When base stations are connected through a limited backhaul, the NSSMF selection reduces to RED and SED schemes. Note that RED does not neces-

sitate that base stations share any coordination or control data. Conversely, SED needs to estimate both base stations' instantaneous load upon packet arrival to steer it appropriately, introducing a limited communication overhead on the backhaul. Therefore, an efficient policy consists in dynamically altering the allocation scheme depending on the arrival rate.

As we can see in Figure 7, RED is favorable for low arrival rates up to λ^*. If the instantaneous system load exceeds λ^*, a shift in resource allocation schemes is thus needed to respect the URLLC reliability requirements.

We now examine the feasibility of the different schemes in the uplink. Recall that we argue that the NSSMF has access to information about the instantaneous load at each BS. It can decide the strategy to steer the downlink packets via the backhaul/fronthaul to the suitable BS. As for the uplink, the process differs significantly. The end-users generate packets that send scheduling requests to the different BSs. The latter replies with a grant indicating the time/frequency resource to be used for the packet. While RED is directly applicable in this case, SED and CAN need some additional signaling that can be specified as follows:

- For SED, when the base station receives the scheduling request, it forwards it to the NSSMF that indicates whether it has to issue a scheduling grant for the user;
- For CAN, both BSs issue a scheduling grant, as if a RED scheme were applied. However, when a base station finishes serving a packet, it signals it to the other base station. The latter may then delete the scheduling grant and reschedule another packet on the liberated resource, provided it has the necessary time and flexibility. Such a fast rescheduling is possible in 5G NR due to the dynamic in-resource scheduling feature, where an uplink scheduling grant may accompany the data intended for a user in the downlink [18].

5. Resource Dimensioning

In this section, we address a resource dimensioning problem. We are interested in finding necessary resources to achieve a given performance. In particular, we want to find

$$\begin{aligned} \min \quad & B_1 + B_2 \\ s.t. \quad & P(T > t_0) = \varepsilon \end{aligned}$$

where $P(T > t_0)$ can be analytically expressed in function of α and β for JSQ, SED, RED and CAN. Moreover, $\alpha = \frac{B_1 e_1}{W}$, $\beta = \frac{B_2 e_2}{W}$, and ε is the outage probability target. In Figure 8, we plot the minimum total bandwidth needed

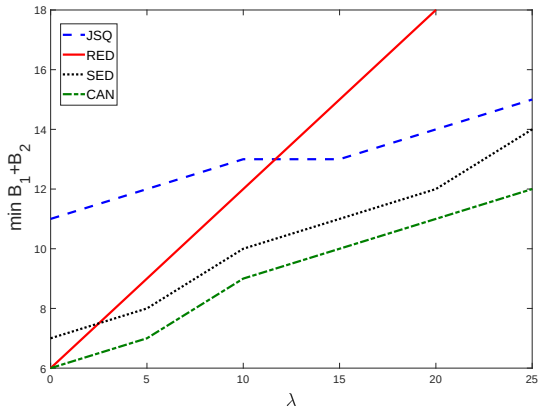

Figure 8: Minimum total bandwidth increasing arrival rate λ for $t_0 = 0.5$ ms and outage probability target $\varepsilon = 10^{-5}$.

to achieve a given outage probability target $\varepsilon = 10^{-5}$ for JSQ, SED, RED and CAN. We observe that the discipline CAN requires less minimum total bandwith with respect to RED, SED and JSQ. RED outperforms SED until a certain arrival rate and behind that value of λ SED requires less bandwidth to achieve the performance.

6. Conclusion

In this paper, we have evaluated the performance of scheduling schemes for URLLC traffic in the context of 5G networks. We exploit the redundant coverage of two frequency layers or RATs to reduce the packets' sojourn time. The most straightforward scheme, called RED, continuously duplicates the packets on both base stations. In contrast, the other schemes exploit the instantaneous state of the base stations' queues and make decisions per packet.

Particularly, JSQ allocates the arriving packet to the queue with the smallest number of waiting packets. Similarly, SED sends the packet to the base station with the shortest expected delay, whereas the CAN scheme always duplicates the packet but cancels the remaining copy upon service of

the other one. We derived explicit expressions for the performance of the different schemes and show that CAN outperforms the other policies for all arrival rates. However, CAN needs strict coordination between the two base stations. In the absence of such coordination, RED is preferred at low arrival rates, while SED is better otherwise.

Finally, as part of future works, it would be interesting to consider the cancel-on-start discipline, that is expected to behave between RED and CAN schemes. A comparison of the theoretical results based on the exponential service rate with practical traces with general service distributions is also an important extension. Finally, we would like to study adaptive multi-connectivity scheduling policies using deep Q-learning, thus allowing users to learn the optimal approach to achieve the required reliability.

Acknowledgements

This work is supported by MAESTRO-5G project funded by the French Agence Nationale de la Recherche (ref. ANR-18-CE25-0012).

References

[1] 3GPP, Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC), 3GPP TR 38.824 V16.0.0 (2019).

[2] 3GPP, Study on scenarios and requirements for next generation access technologies, 3GPP TR 38.913 v15.0.0, Tech. Rep. (2018).

[3] 3GPP, Study on new radio (NR) access technology, 3GPP TR 38.912 V15.0.0 (2018).

[4] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, B. Shim, Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects, IEEE Wireless Communications 25 (2018) 124–130.

[5] Y. Han, S. E. Elayoubi, A. Galindo-Serrano, V. S. Varma, M. Messai, Periodic radio resource allocation to meet latency and reliability requirements in 5G networks, in: 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), IEEE, 2018, pp. 1–6.

[6] R. Combes, S. E. Elayoubi, T. Varela, Optimal retransmission policies for ultra-reliable low latency communications with delayed feedback, in: 2019 IEEE Global Communications Conference (GLOBECOM), IEEE, 2019, pp. 1–6.

[7] P. Marsch, I. Da Silva, O. Bulakci, M. Tesanovic, S. E. El Ayoubi, T. Rosowski, A. Kaloxylos, M. Boldi, 5G radio access network architecture: Design guidelines and key considerations, IEEE Communications Magazine 54 (2016) 24–32.

[8] L. Flatto, H. McKean, Two queues in parallel, Communications on Pure and Applied Mathematics XXX (1977) 255–263.

[9] G. Brightwell, M. Luczak, The supermarket model with arrival rate tending to one, arXiv preprint arXiv:1201.5523 (2012).

[10] J. Selen, I. Adan, S. Kapodistria, J. Leeuwaarden, Steady-state analysis of shortest expected delay routing, Queueing Syst. Theory Appl. 84 (2016) 309–354. URL: https://doi.org/10.1007/s11134-016-9497-7. doi:10.1007/s11134-016-9497-7.

[11] U. Ayesta, T. Bodas, I. M. Verloop, On a unifying product form framework for redundancy models, Performance Evaluation 127 (2018) 93–119.

[12] K. Gardner, S. Zbarsky, S. Doroudi, M. Harchol-Balter, E. Hyytia, Reducing latency via redundant requests: Exact analysis, ACM SIGMETRICS Performance Evaluation Review 43 (2015) 347–360.

[13] A. Chagdali, S. E. Elayoubi, A. M. Masucci, A. Simonian, Performance of urllc traffic scheduling policies with redundancy, in: 2020 32nd International Teletraffic Congress (ITC 32), 2020, pp. 55–63. doi:10.1109/ITC3249928.2020.00015.

[14] 3GPP, Study on management and orchestration of network slicing for next generation network, 3GPP TR 28.801 V15.1.0, Tech. Rep. (2018).

[15] L. Flatto, S. Hahn, Two parallel queues created by arrivals with two demands I, SIAM Journal on Applied Mathematics 44 (1984) 1041–1053.

[16] I. J. B. F. Adan, J. Wessels, W. Zijm, Analysis of the asymmetric shortest queue problem, Queueing Systems 8 (1991) 1–58.

[17] L. Flatto, Two parallel queues created by arrivals with two demands II, SIAM Journal on Applied Mathematics 45 (1985) 861–878. URL: http://www.jstor.org/stable/2101634.

[18] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, A. Szufarska, A flexible 5G frame structure design for frequency-division duplex cases, IEEE Communications Magazine 54 (2016) 53–59.