Does Queue Correlation Matter in 5G Multi-Connectivity with Packet Duplication?

Cheng-Yeh Chen and Hung-Yun Hsieh

***Abstract*—Multi-connectivity (MC) with packet duplication allows users to connect to multiple base stations for ultra-reliability. While related work has investigated the impact of path correlation in terms of correlated channel fading or shadowing correlation, we argue in this letter that another source of correlation referred to as "queue correlation" needs to be considered to better profile the performance of MC. To proceed, we build a queue-theoretic framework to analyze the origin and impact of queue correlation. Evaluation results show that queue correlation could increase service outage *by an order of magnitude* compared to the baseline without considering such an effect under a standard 5G urban macro-cell environment.**

***Index Terms*—Multi-connectivity, packet duplication, queue correlation.**

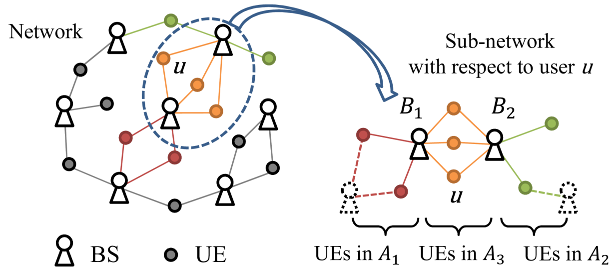

Fig. 1. Network scenario: For any given UE u enabled by MC, we extract the sub-network containing the two BSs that u connects to and all UEs connecting to the two BSs. All UEs included can be split into 3 sets as A_1, A_2, and A_3.

I. Introduction

5G new radio (NR) aims to facilitate a wide spectrum of challenging scenarios, including enhanced mobile broadband (eMBB) and ultra-reliable and low latency communications (URLLC). A promising technology to achieve those demanding requirements is multi-connectivity (MC) [1]. MC allows a single user equipment (UE) to establish multiple active links to multiple base stations (BSs) simultaneously for higher path diversity. Related work has shown that MC is suitable to reduce system cost by mobile data offloading [2], ensure ultra-high reliability in URLLC [3], and enhance resiliency and service continuity for edge computing [4].

A potential drawback of MC comes from the channel correlation between links. If the chosen links do not exhibit sufficient path diversity, the achievable diversity gain by MC would be reduced since an error in one channel could imply the same in another channel. Related work has hence investigated the impact of channel correlation on MC [5]–[7]. Shadowing correlation between device-to-device and cellular links is shown in [5] to cause degradation in available range of URLLC. Upper and lower bounds on reliability of correlated fading channels are analyzed in [6] with the impact of dependency and the gap between best- and worst-case revealed to be significant. [7] demonstrates that extra signal-to-noise ratio is required to offset the degradation caused by channel correlation over correlated fading channels.

While previous work has identified the impact of channel correlation on MC, the queueing effect at each BS has not been considered, where *individual queues are typically assumed to be independent over links*. However, when a large number of UEs are enabled by MC, the assumption of independent queues may not hold. For example, if packet duplication (PD), a common scheduling method in the 3GPP standards [1], is adopted for MC, duplicated packets would be transmitted simultaneously over multiple redundant links. Duplicated traffic could introduce correlation among the queues of individual BSs (referred to as "queue correlation" in this letter), thus reducing diversity gain of these redundant links. Clearly, correlated queues result in correlated packet drops (due to, say, queue overflow) and hence higher loss rate for MC-enabled UEs.

We argue in this letter that queue correlation has non-negligible impact on performance analysis of 5G multi-connectivity with packet duplication. Toward this objective, our contribution could be summarized as follows:

- To the best of our knowledge, *we are the first to identify and profile the impact of queue correlation on MC with PD* since a plethora of existing work [5]–[7] on path correlation focuses only on the dependency of channel fading such as shadowing and spatial correlation without regard to queue correlation.
- We propose to model MC with PD as a multi-queue system for analyzing the impact of queue correlation under arbitrary degrees of multi-connectivity. We apply level-dependent quasi-birth-death process (LDQBD) to numerically solve the joint system-size distribution of the proposed multi-queue system.
- We show through network simulation that *queue correlation does matter since it increases outage probability by a factor of 10 for more than 25% of UEs in 5G urban macro-cell environment*.

II. System Models

A. Network Model

We focus on a scenario where UEs are capable of establishing dual connections with two BSs simultaneously, as specified in 3GPP [1]. All MC-enabled UEs adopt PD

The authors are with the Graduate Institute of Communication Engineering and Department of Electrical Engineering, National Taiwan University, Taipei 106, Taiwan (e-mail: r08942083@ntu.edu.tw; hungyun@ntu.edu.tw).
This work was supported in part by funds from the Ministry of Science and Technology and National Taiwan University under Grants MOST-110-2221-E-002-086 and MOST-111-2221-E-002-100.

to transmit duplicated data over all links, while other UEs establish a single connection to a BS as Fig. 1 shows. We consider one UE u that connects to two BSs named B_1 and B_2. We construct a sub-network containing B_1, B_2, and all UEs connecting to them. In the sub-network, all UEs can be split into three sets: A_1, A_2, and A_3. Set A_1 includes UEs connecting to B_1 but not to B_2. Set A_2 includes UEs connecting to B_2 but not to B_1. Finally, set A_3 includes UEs connecting to both B_1 and B_2. We discuss in Section III-D how we extend dual-connectivity to higher degrees of connectivity for queue-theoretic analysis.

B. Queueing Model

Each BS is considered as an $M/M/C/K$ queue [8] as follows: (1) The first M indicates that the packet arrival processes from all UEs are mutually independent Poisson processes. Let $\lambda_{i,j}$ denote the packet arrival rate of UE j in set A_i and $\lambda_i = \sum_{j=1}^{|A_i|} \lambda_{i,j}$ denote the compound arrival rate for A_i. As shown in Fig. 1, the arrival rates for B_1 and B_2 are $\lambda_1 + \lambda_3$ and $\lambda_2 + \lambda_3$, respectively. (2) The second M indicates the service process is Poisson process, which has also been adopted in recent work on URLLC applications [9]. We denote μ_i as the service rate for B_i determined by the amount of radio resource allocated for transmission. (3) C indicates the maximum number of concurrent transmissions supported by one BS, which is limited by the total available bandwidth. Let c_i denote the maximum number of concurrent transmissions for B_i. (4) K indicates the maximum system size introduced to capture the hard latency requirement in URLLC [8]. If the number of packets being queued and processed exceeds the maximum system size k_i for B_i, any new packet arrival is dropped.

C. Outage Model for Single Connection

We consider downlink transmission in this letter to keep the discussion focused, although the uplink scenario could be analyzed similarly. We characterize service outage by the following three types of errors: *queue blocking*, *delay violation*, and *channel decoding error*, and we denote $P_b^{(i)}$, $P_d^{(i)}$, and $P_e^{(i)}$ as the probabilities of queue blocking, delay violation, and channel decoding error for a connection in B_i.

1) Queue Blocking: Upon the arrival of a downlink packet in B_i, if the number of packets being processed or queued in B_i already reaches the maximum system size k_i, a blocking event occurs and the packet is dropped. Let N_i be the system size in B_i and $P_{N_i}(n)$ be the probability mass function of N_i. The blocking probability $P_b^{(i)}$ thus can be defined as

$$P_b^{(i)} = P_{N_i}(k_i). \tag{1}$$

2) Delay Violation: If the packet is not blocked, it is admitted to the queue and waits for resource allocation. If B_i fails to complete the transmission within the delay constraint, delay violation occurs, and the packet times out. Let T_i be the waiting time for a packet being queued and processed in B_i. Define $f_{T_i}(t)$ as the probability density function (PDF) of T_i and $F_{T_i}(t)$ as the corresponding cumulative density function (CDF). The delay violation probability with respect to a delay bound d can be written as

$$P_d^{(i)} = 1 - F_{T_i}(d). \tag{2}$$

3) Channel Decoding Error: If the packet is transmitted in time but the UE cannot decode the received data, a decoding error occurs. We apply the finite block-length model [10], a typical channel model for URLLC, to model the decoding error probability. In an additive white Gaussian noise channel, the decoding error probability to transmit L_i bits of information in r_i units of bandwidth with signal-to-noise ratio (SNR) γ_i is approximated by $P_e^{(i)} \approx Q\left(\left(r_i C(\gamma_i) - L_i\right)/\sqrt{r_i V(\gamma_i)}\right)$, where $C(\gamma) = \log_2(1+\gamma)$ represents the channel capacity under the infinite block length regime, $V(\gamma) = \log_2(e)^2 \left(1 - (1+\gamma)^{-2}\right)$ represents the channel dispersion, and Q is the complementary Gaussian CDF.

4) Overall Outage Probability: We define $P_{out}^{(i)}$ as the overall outage probability of the connection in B_i and it could be given as the summation of probabilities over disjoint failure events as follows:

$$P_{out}^{(i)} = P_b^{(i)} + \left(1 - P_b^{(i)}\right) P_d^{(i)} + \left(1 - P_b^{(i)}\right)\left(1 - P_d^{(i)}\right) P_e^{(i)}, \tag{3}$$

for $i = 1, 2$. Note that $\left(1 - P_b^{(i)}\right) P_d^{(i)}$ is the probability that B_i is not blocked but times out, while $\left(1 - P_b^{(i)}\right)\left(1 - P_d^{(i)}\right) P_e^{(i)}$ is the probability that B_i is not blocked and timed out, but the UE fails to decode the packet.

III. Queue Correlation for Multi-connectivity

Based on the outage model for a single connection described in Section II, we formulate the outage probability of MC with PD to capture the impact of queue correlation.

A. Outage Model for MC with PD

Since PD provides redundant links for transmissions, an error occurs only when all of the links fail. Let $P_{N_1,N_2}(n_1, n_2)$ and $f_{T_1,T_2}(t_1, t_2)$ denote the joint system-size and waiting-time distribution respectively for B_1 and B_2. To distinguish notation for marginal and joint error probability, we use $P_x^{(i)}$ to represent the marginal error probability for BS B_i and P_x to represent the joint error probability, where x stands for different error types including b, d, e, and out.

1) Queue Blocking: The marginal blocking probability $P_b^{(i)}$ defined in (1) has an analytical solution for an $M/M/C/K$ queue [11]. Take $P_b^{(1)}$ for instance. $P_b^{(1)}$ can be written as the summation of the joint system-size distribution at the maximum value $n_1 = k_1$ for B_1:

$$P_b^{(1)} = \sum_{j=0}^{k_2} P_{N_1,N_2}(k_1, j) = \frac{(\lambda_1 + \lambda_3)^{k_1}}{c_1^{k_1 - c_1} c_1! \mu_1^{k_1}} P_0^{(1)}, \tag{4}$$

where $P_0^{(1)}$ is the normalization term to make the total probability equal to 1:

$$P_0^{(1)} = \left(\sum_{n=0}^{c_1 - 1} \frac{(\lambda_1 + \lambda_3)^n}{n! \mu_1^n} + \sum_{n=c_1}^{k_1} \frac{(\lambda_1 + \lambda_3)^n}{c_1^{n - c_1} c_1! \mu_1^n} \right)^{-1}. \tag{5}$$

However, the joint blocking probability P_b defined as

$$P_b = P_{N_1,N_2}(k_1, k_2) \tag{6}$$

has no analytical solution. To proceed, we convert the two-queue system into a special kind of birth-death process and propose an algorithm to solve it numerically in Section III-B.

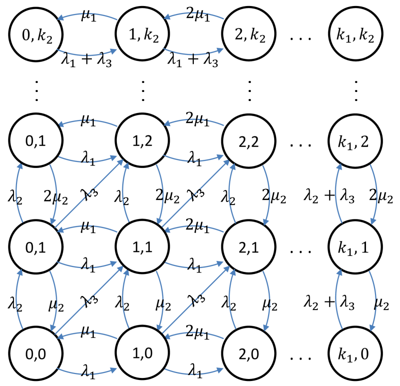

Fig. 2. Equivalent state transition diagram of the queueing system with two $M/M/C/K$ queues and three independent arrival processes.

2) Delay Violation: Similarly, the marginal delay violation probability $P_d^{(i)}$ defined in (2) is analytical for an $M/M/C/K$ queue but the joint delay violation probability:

$$P_d = \Pr[T_1 > d, T_2 > d] = \int_d^\infty \int_d^\infty f_{T_1,T_2}(t_1,t_2)dt_1dt_2, \tag{7}$$

has no analytical solution. We will solve it in terms of the joint system-size distribution in Section III-C.

3) Channel Decoding Error: While analytical expressions derived in previous work for decoding errors under channel correlation [5]–[7] could be applied in our framework to model the joint decoding error probability, in this letter we consider independent decoding errors to proceed with the analysis since our primary focus is on queue correlation. Hence, the joint decoding error probability is calculated as the product of the two marginal error probabilities $P_e = P_e^{(1)}P_e^{(2)}$, where $P_e^{(i)}$ is given in Section II-C3.

4) Overall Outage Probability: If B_1 and B_2 are independent, $P_{out} = P_{out}^{(1)}P_{out}^{(2)}$, where $P_{out}^{(i)}$ is given in (3). However, if correlation exists in queue, calculation of P_{out} is more involved. We derive P_{out} from the marginal and joint error probabilities by considering queue blocking, delay violation, and channel decoding error as follows:

$$\begin{aligned} P_{out} = P_b &+ \left(P_b^{(1)} - P_b\right)\left(P_d^{(2)} + P_e^{(2)} - P_d^{(2)}P_e^{(2)}\right) \\ &+ \left(P_b^{(2)} - P_b\right)\left(P_d^{(1)} + P_e^{(1)} - P_d^{(1)}P_e^{(1)}\right) \\ &+ \left(1 - P_b^{(1)} - P_b^{(2)} + P_b\right)\Big[P_d + \left(P_d^{(1)} - P_d\right)P_e^{(2)} \\ &+ \left(P_d^{(2)} - P_d\right)P_e^{(1)} + \left(1 - P_d^{(1)} - P_d^{(2)} + P_d\right)P_e\Big]. \end{aligned} \tag{8}$$

The derivation of P_{out} is obtained by the inclusion-exclusion principle, whose details are omitted due to lack of space.

B. Joint System-Size Distribution

To the best of our knowledge, there is no existing solution for $P_{N_1,N_2}(n_1,n_2)$, where the two queues share a portion of the common duplicated arrival process. To solve $P_{N_1,N_2}(n_1,n_2)$, we formulate the two-queue system into a finite two-dimensional Markov chain having various transition structures for each state. This corresponds to a class of vector state process called level-dependent quasi-birth-death process (LDQBD). We derive the generator matrix and apply an efficient and stable numerical algorithm proposed in [12] for finite LDQBD to solve $P_{N_1,N_2}(n_1,n_2)$.

We start from the equivalent state transition diagram of the two-queue system shown in Fig. 2. Each state represents a two-dimension state vector of system sizes in B_1 and B_2. Note that although the two-queue system is constructed for a given UE u, the system sizes in both BSs reflect the influence of all UEs involved in A_1, A_2, and A_3. To solve the two-dimensional Markov chain, we need to derive the block-partitioned tri-diagonal form of generator matrix $\boldsymbol{Q}$ and all of its submatrices. $\boldsymbol{Q}$ can be written in terms of $(k_1+1)\times(k_1+1)$ submatrices as follows:

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{0,0} & \boldsymbol{Q}_{0,1} & & & \\ \boldsymbol{Q}_{1,0} & \boldsymbol{Q}_{1,1} & \boldsymbol{Q}_{1,2} & & \\ & \boldsymbol{Q}_{2,1} & \boldsymbol{Q}_{2,2} & \cdots & \\ & & \boldsymbol{Q}_{3,2} & \cdots & \boldsymbol{Q}_{k_1-1,k_1} \\ & & & \cdots & \boldsymbol{Q}_{k_1,k_1} \end{pmatrix}, \tag{9}$$

where each submatrix $\boldsymbol{Q}_{i,j}$ is a $(k_2+1)\times(k_2+1)$ matrix. The (i',j')-th element in $\boldsymbol{Q}_{i,j}$ represents the transition rate of birth-death process from system size $(n_1 = i, n_2 = i')$ to $(n_1 = j, n_2 = j')$ for an LDQBD.

More precisely, we define the following $(k_2+1)\times(k_2+1)$ matrices for $j = 1,\cdots,k_1-1$:

$$\begin{aligned} \boldsymbol{Q}_{0,0} &= \lambda_2(\boldsymbol{N}+\boldsymbol{C}) - \Sigma_{i=1}^3\lambda_i\boldsymbol{I} + \boldsymbol{B}(\boldsymbol{N}^T - \boldsymbol{I}), && (10)\\ \boldsymbol{Q}_{j-1,j} &= \lambda_3(\boldsymbol{N}+\boldsymbol{C}) + \lambda_1\boldsymbol{I}, && (11)\\ \boldsymbol{Q}_{j,j-1} &= \min\{j,c_1\}\,\mu_1\boldsymbol{I}, && (12)\\ \boldsymbol{Q}_{j,j} &= \boldsymbol{Q}_{0,0} - \boldsymbol{Q}_{j,j-1}, && (13)\\ \boldsymbol{Q}_{k_1,k_1} &= (\lambda_2+\lambda_3)(\boldsymbol{N}+\boldsymbol{C}) - (\lambda_2+\lambda_3+c_1\mu_1)\boldsymbol{I} \\ &\quad + \boldsymbol{B}(\boldsymbol{N}^T - \boldsymbol{I}), && (14) \end{aligned}$$

where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{N}$ is the nilpotent matrix, $\boldsymbol{B}$ is a diagonal matrix with $\boldsymbol{B}_{i,i} = \min\{i,c_2\}\,\mu_2$, and $\boldsymbol{C}$ is a matrix with $\boldsymbol{C}_{k_2+1,k_2+1} = 1$ and all other elements being 0.

We then define the system-size distribution matrix $\boldsymbol{\pi}$ as

$$\boldsymbol{\pi} = [\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_{k_1}], \tag{15}$$

where $\boldsymbol{\pi}_j = [P_{N_1,N_2}(j,0), P_{N_1,N_2}(j,1), \cdots, P_{N_1,N_2}(j,k_2)]$. To find the steady state, one has to solve $\boldsymbol{\pi}\boldsymbol{Q} = \boldsymbol{0}$ subject to $\boldsymbol{\pi}\boldsymbol{e} = 1$, where $\boldsymbol{e}$ is a column vector of ones. Note that $\boldsymbol{\pi}\boldsymbol{Q} = \boldsymbol{0}$ is equivalent to the following system of equations,

$$\boldsymbol{\pi}_0\boldsymbol{Q}_{0,0} + \boldsymbol{\pi}_1\boldsymbol{Q}_{1,0} = \boldsymbol{0}, \tag{16}$$

$$\boldsymbol{\pi}_{j-1}\boldsymbol{Q}_{j-1,i} + \boldsymbol{\pi}_j\boldsymbol{Q}_{j,j} + \boldsymbol{\pi}_{j+1}\boldsymbol{Q}_{j+1,j} = \boldsymbol{0}, \tag{17}$$

$$\boldsymbol{\pi}_{k_1-1}\boldsymbol{Q}_{k_1-1,k_1} + \boldsymbol{\pi}_{k_1}\boldsymbol{Q}_{k_1,k_1} = \boldsymbol{0}, \tag{18}$$

for $j = 1,\cdots,k_1-1$. An efficient and stable way to compute the numerical solution for finite LDQBD was proposed in [12] by introducing the rate matrix $\boldsymbol{R}_j$ as the solution of

$$\boldsymbol{\pi}_{j+1} = \boldsymbol{\pi}_j\boldsymbol{R}_j, \tag{19}$$

for $j = 0,\cdots,k_1-1$. Using the notation of the rate matrix, (16), (17), and (18) can be transformed into

$$\boldsymbol{\pi}_0\left(\boldsymbol{Q}_{0,0} + \boldsymbol{R}_0\boldsymbol{Q}_{1,0}\right) = \boldsymbol{0}, \tag{20}$$

$$\boldsymbol{\pi}_{j-1}\left(\boldsymbol{Q}_{j-1,j} + \boldsymbol{R}_{j-1}\boldsymbol{Q}_{j,j} + \boldsymbol{R}_{j-1}\boldsymbol{R}_j\boldsymbol{Q}_{j+1,j}\right) = \boldsymbol{0}, \tag{21}$$

$$\boldsymbol{\pi}_{k_1-1}\left(\boldsymbol{Q}_{k_1-1,k_1} + \boldsymbol{R}_{k_1-1}\boldsymbol{Q}_{k_1,k_1}\right) = \boldsymbol{0}, \tag{22}$$

for $j = 1, \cdots, k_1 - 1$. With these equations, one can first compute $\boldsymbol{R}_{k_1-1}$ in (22) as

$$\boldsymbol{R}_{k_1-1} = -\boldsymbol{Q}_{k_1-1,k_1}\boldsymbol{Q}_{k_1,k_1}^{-1}, \tag{23}$$

and then find all the other rate matrices recursively using (21) for $j = k_1 - 1, \cdots, 1$, as follows:

$$\boldsymbol{R}_{j-1} = -\boldsymbol{Q}_{j-1,j}\left(\boldsymbol{Q}_{j,j} + \boldsymbol{R}_j\boldsymbol{Q}_{j+1,j}\right)^{-1}. \tag{24}$$

Starting from a non-zero matrix $\boldsymbol{\pi}_0$ in (20), we can construct the whole steady-state matrix using the relation $\boldsymbol{\pi}_{j+1} = \boldsymbol{\pi}_j\boldsymbol{R}_j$ for $j = 0, \cdots, k_1 - 1$. Finally, we can normalize the summation of $\boldsymbol{\pi}$ to 1 using $\boldsymbol{\pi}\boldsymbol{e} = \boldsymbol{1}$. Algorithm 1, as adapted from [12], summarizes the whole procedure for obtaining $\boldsymbol{\pi}$, or equivalently $P_{N_1,N_2}(n_1, n_2)$, needed to calculate P_b in (6).

C. Joint Waiting-Time Distribution

To find the joint waiting-time distribution for the two BSs, we introduce two random variables M_1 and M_2 to represent the system sizes observed by a packet not dropped by B_1 and B_2. The joint probability mass function $P_{M_1,M_2}(m_1, m_2)$ can be obtained from the conditional probability of the joint system-size distribution as follows,

$$P_{M_1,M_2}(m_1, m_2) = \frac{P_{N_1,N_2}(m_1, m_2)}{\sum_{n_1=0}^{k_1-1}\sum_{n_2=0}^{k_2-1} P_{N_1,N_2}(n_1, n_2)}, \tag{25}$$

for $m_i = 0, \cdots, k_i - 1$. The joint CDF $F_{T_1,T_2}(t_1, t_2)$ can be written as the sum of the waiting-time distribution observed by each arrival conditioned on $P_{M_1,M_2}(m_1, m_2)$ as follows:

$$\begin{aligned} &F_{T_1,T_2}(t_1, t_2) \\ &= \sum_{m_1=0}^{k_1-1}\sum_{m_2=0}^{k_2-1} F_{T_1,T_2|M_1,M_2}(t_1, t_2|m_1, m_2)P_{M_1,M_2}(m_1, m_2) \\ &= \sum_{m_1=0}^{k_1-1}\sum_{m_2=0}^{k_2-1} F_{T_1|M_1}(t_1|m_1)F_{T_2|M_2}(t_2|m_2)P_{M_1,M_2}(m_1, m_2). \end{aligned} \tag{26}$$

Notice that the service processes of the two BSs are independent, so the conditional waiting times $T_1|M_1$ and $T_2|M_2$ are independent if the system sizes in both BSs are given. Such relation leads to the final result of (26). Furthermore, $F_{T_1|M_1}(t_1|m_1)$ and $F_{T_2|M_2}(t_2|m_2)$ can be written as the CDF of the sum of independent exponential (service time) and Erlang (waiting time) random variables as follows:

$$F_{T_i|M_i}(t_i|m_i) = \begin{cases} F_{S_i}(t_i), & m_i = 0, \cdots, c_i - 1, \\ F_{S_i+W_{i,m_i}}(t_i), & m_i = c_i, \cdots, k_i - 1, \end{cases} \tag{27}$$

where $S_i \sim \text{Exponential}(\mu_i)$ and $W_{i,m_i} \sim \text{Erlang}(m_i - c_i + 1, c_i\mu_i)$ for $i = 1, 2$. After obtaining $F_{T_1,T_2}(t_1, t_2)$ by (26), we can find $f_{T_1,T_2}(t_1, t_2)$ via the derivative in t_1 and t_2 needed for calculating P_d in (7) and (8).

D. Extension from Dual-Connectivity to Multi-Connectivity

We have focused on dual-connectivity so far for sake of expression clarity in this letter. However, the proposed analytical framework can be extended to higher degrees of connectivity. If a given UE u establishes g-connectivity ($g \in \mathcal{N}$) with BSs labeled as B_1 to B_g, our main goal becomes solving $P_{N_1,\cdots,N_g}(n_1, \cdots, n_g)$, whose numerical solution could be obtained by Algorithm 1 since the applied LDQBD is not restricted by certain dimensionality [12]. Note that UEs other than u could have any degree of connectivity to individual BSs regardless of g. To proceed with the analysis, one could first construct a sub-network similar to Fig. 1 including all the UEs connected by B_1 to B_g, and partition these UEs into $2^g - 1$ disjoint UE sets: A_1 to A_{2^g-1}. UEs connected by the same subset of BSs would be classified into the same UE set. The compound arrival rates (λ_1 to λ_{2^g-1}) could be obtained by summing all the packet arrival rates within each UE set. Then, one could define the g-dimensional Markov chain similar to the one in Fig. 2 and derive the corresponding block-partitioned tri-diagonal generator matrix $\boldsymbol{Q}$, whose dimension is still 2 but matrix size is $\prod_{i=1}^{g}(k_i+1)$ by $\prod_{i=1}^{g}(k_i+1)$. The system-size distribution $\boldsymbol{\pi}$ is solved by the system of equations indicated by $\boldsymbol{\pi}\boldsymbol{Q} = \boldsymbol{0}$ following the same procedure in Algorithm 1.

Algorithm 1 Algorithm for calculating $P_{N_1,N_2}(n_1, n_2)$

- **Input:** $\boldsymbol{Q}$
- Compute $\boldsymbol{R}_{k_1-1} = -\boldsymbol{Q}_{k_1-1,k_1}\boldsymbol{Q}_{k_1,k_1}^{-1}$.
- **for** $j = k_1 - 1$ to 1 **do**
  - Compute $\boldsymbol{R}_{j-1} = -\boldsymbol{Q}_{j-1,j}\left(\boldsymbol{Q}_{j,j} + \boldsymbol{R}_j\boldsymbol{Q}_{j+1,j}\right)^{-1}$.
- **end for**
- Solve $\boldsymbol{\pi}_0\left(\boldsymbol{Q}_{0,0} + \boldsymbol{R}_0\boldsymbol{Q}_{1,0}\right) = \boldsymbol{0}$ for $\boldsymbol{\pi}_0 \neq 0$.
- **for** $j = 0$ to $k_1 - 1$ **do**
  - Compute $\boldsymbol{\pi}_{j+1} = \boldsymbol{\pi}_j\boldsymbol{R}_j$.
- **end for**
- Normalize $\boldsymbol{\pi}$ to satisfy $\boldsymbol{\pi}\boldsymbol{e} = \boldsymbol{1}$.
- **return** $\boldsymbol{\pi}$

IV. Performance Evaluation

In this section, we show the distribution of queue correlation in a standard 5G urban macro-cell environment [13] and its impact on the performance of MC with PD. Simulation parameters are specified in line with common assumptions for URLLC [8]. The bandwidth is 5 MHz, the sub-carrier spacing is 60 kHz, and the transmission time interval (TTI) is 0.25 ms. The packet arrival process is Poisson with a rate of 1000 packets per second and the packet size is set as 32 bytes. The delay bound used in (2) for each packet is 1 ms.

We measure queue correlation based on the Pearson correlation coefficient of P_b, P_d and P_{out} between B_1 and B_2. Take ρ_b as an example. Since P_b is the joint probability and $P_b^{(1)}$ and $P_b^{(2)}$ are the marginal probabilities, by the derivation in [14], we have

$$\rho_b = \left(P_b - P_b^{(1)}P_b^{(2)}\right)/\sqrt{P_b^{(1)}\left(1 - P_b^{(1)}\right)P_b^{(2)}\left(1 - P_b^{(2)}\right)}. \tag{28}$$

The same equation holds for ρ_d and ρ_{out}. A key factor in our evaluation is the control of the duplication ratio α defined as

$$\alpha = \frac{\lambda_3}{\min\{\lambda_1 + \lambda_2\} + \lambda_3}, \tag{29}$$

which is used to quantify the level of duplication between any pair of BSs between 0 and 1. $\alpha = 0$ indicates no duplicated traffic shared between the pair of BSs while $\alpha = 1$ indicates at least one BS shares all of its traffic with the other BS due to traffic duplication. We conduct our evaluation under various traffic loads $\beta_i = \lambda_i/\mu_i$ for BS B_i.

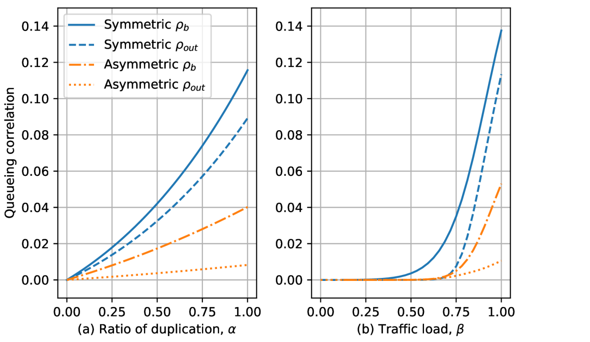

Fig. 3. Blocking correlation versus (a) duplication ratio at traffic load $\beta = 0.95$ and (b) traffic load at duplication ratio $\alpha = 1$ for both the symmetric and asymmetric pair of BSs.

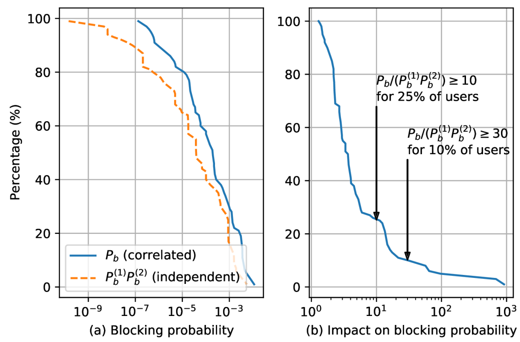

Fig. 4. Comparison of blocking probability in a standard 5G urban macro-cell environment. (a) CCDF for blocking probability (P_b for correlated case and $P_b^{(1)}P_b^{(2)}$ for independent case). (b) CCDF for impact on blocking probability ($P_b/P_b^{(1)}P_b^{(2)}$).

A. *Impact of Queue Correlation*

Both α and β can affect the level of queue correlation. Fig. 3 shows how these two parameters affect ρ_b and ρ_{out}. (The value of ρ_d is much smaller, but the trend is similar.) The symmetric or asymmetric pairing of BSs in Fig. 3 denotes the condition that the total traffic for the two BSs is *balanced or unbalanced.* From Fig. 3(a), one could observe that ρ_b increases almost linearly as duplication ratio α increases. In contrast, ρ_b is negligible for small β but shoots high for large β in Fig. 3(b). These observations provide an insight that *queue correlation is evident when the traffic load is high* ($\beta \geq 0.75$), where P_b is high compared to P_e and therefore dominates P_{out}. We observe that, for example, P_b could be 59 times higher than $P_b^{(1)}P_b^{(2)}$, the values under independent assumption, when $\beta = 0.75$. As Fig. 3 shows, it is beneficial to *connect to an asymmetric pair of BSs (e.g. micro and macro BSs)* since a much smaller queue correlation is found compared to the symmetric case.

B. *Distribution of Queue Correlation*

To evaluate how queue correlation is distributed among all users, we conduct network simulation over a network layout compliant to standard 5G urban macro-cell test environment [13] with 19 BSs with 100 UEs distributed uniformly inside the service coverage. Each UE connects to two BSs with the strongest SNRs. One could observe in Fig. 4(a) that the curve of P_b (with queue correlation) deviates from that of $P_b^{(1)}P_b^{(2)}$ (independent queues), and hence queue correlation does alter the distribution of the blocking probability among UEs. Fig. 4(b) further reveals that over 25% of UEs experiencing 10 times higher blocking probability and 10% of UEs experiencing 30 times higher blocking probability compared to the value under the independent assumption. Thus, *queue correlation could substantially decrease the availability of URLLC service and should not be overlooked.*

V. Conclusion and Future Work

Our evaluation has shown that queue correlation plays a non-negligible role in MC performance, especially when the traffic load is high. We believe that the proposed framework for analyzing MC with PD has potential to be used as the basis for further performance optimization. For example, when performing admission control or handover in an MC-enabled network, the number of UEs connecting to the same pair or group of BSs should be reduced to avoid queue correlation. The user association problem in MC with PD should consider the increased possibility of collision caused by duplicated traffic. With the aid of the proposed analytical framework, one could formulate an optimization problem aiming to jointly mitigate queue correlation and channel correlation to further guarantee the performance for URLLC.

References

[1] 3GPP, "Multi-connectivity; Overall description," 3rd Generation Partnership Project (3GPP), Technical Specification 37.340, v16.8.0, 2021.

[2] Y. Wu, Y. He, L. P. Qian, J. Huang, and X. Shen, "Optimal Resource Allocations for Mobile Data Offloading via Dual-Connectivity," *IEEE Trans. on Mobile Computing*, vol. 17, no. 10, pp. 2349–2365, 2018.

[3] P. Popovski, . Stefanovi, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A. Bana, "Wireless Access in Ultra-Reliable Low-Latency Communication (URLLC)," *IEEE Trans. on Communications*, vol. 67, no. 8, pp. 5783–5801, 2019.

[4] M. F. zko, A. Koutsaftis, R. Kumar, P. Liu, and S. S. Panwar, "The Impact of Multi-Connectivity and Handover Constraints on Millimeter Wave and Terahertz Cellular Networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 6, pp. 1833–1853, 2021.

[5] C. She, Z. Chen, C. Yang, T. Q. S. Quek, Y. Li, and B. Vucetic, "Improving Network Availability of Ultra-Reliable and Low-Latency Communications With Multi-Connectivity," *IEEE Trans. on Communications*, vol. 66, no. 11, pp. 5482–5496, 2018.

[6] K.-L. Besser and E. A. Jorswieck, "Reliability Bounds for Dependent Fading Wireless Channels," *IEEE Trans. on Wireless Communications*, vol. 19, no. 9, pp. 5833–5845, 2020.

[7] Y. Chen, A. Wolf, M. Drpinghaus, J. C. S. S. Filho, and G. P. Fettweis, "Impact of Correlated Fading on Multi-Connectivity," *IEEE Trans. on Wireless Communications*, vol. 20, no. 2, pp. 1011–1022, 2021.

[8] Chih-Ping Li, Jing Jiang, W. Chen, Tingfang Ji, and J. Smee, "5G Ultra-Reliable and Low-Latency Systems Design," in *European Conference on Networks and Communications (EuCNC)*, 2017, pp. 1–5.

[9] S. Bakri, P. A. Frangoudis, and A. Ksentini, "Dynamic Slicing of RAN Resources for Heterogeneous Coexisting 5G Services," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[10] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel Coding Rate in the Finite Blocklength Regime," *IEEE Trans. on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[11] D. Gross, *Fundamentals of queueing theory*. John Wiley & Sons, 2008.

[12] H. Baumann and W. Sandmann, "Numerical solution of level dependent quasi-birth-and-death processes," *Procedia Computer Science*, vol. 1, no. 1, pp. 1561–1569, 2010.

[13] M. Series, "Guidelines for evaluation of radio interface technologies for IMT-2020," 2017.

[14] M. Ganjalizadeh, P. Di Marco, J. Kronander, J. Sachs, and M. Petrova, "Impact of Correlated Failures in 5G Dual Connectivity Architectures for URLLC Applications," in *2019 IEEE Globecom Workshops (GC Wkshps)*, 2019, pp. 1–6.