IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 5, 2018, accepted April 12, 2018, date of publication May 1, 2018, date of current version July 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2832130

Packet-Level Modeling of Cooperative Diversity: A Queueing Network Approach

NAVID TADAYON[1], (Member, IEEE), AND GEORGES KADDOUM[2], (Member, IEEE)
[1]Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 2E4, Canada
[2]Département de génie électrique, École de Technologie Supérieure, Montreal, QC H3C 1K3, Canada

Corresponding author: Navid Tadayon (navid.tadayon@utoronto.ca)

This work was supported in part by the NSERC Discovery under Grant 435243-2013, in part by the NSERC PDF Fellowship, and in part by the Fund de Recherche du Quebec-Nature et Technologies.

ABSTRACT Cooperative communications are expected to be a centerpiece for 5G cellular networks. Using cooperation, a wider and more uniform network coverage is attained, network capacity is enhanced, and power consumption is drastically reduced. While these latter heavily rely on collaboration among network elements, there is barely a packet-level perspective on what tangible gains cooperation is able to achieve. Motivated by this fact, different from studies on the capacity analysis of cooperative protocols, this paper tackles cooperation as a packet-level problem. The latter perspective empowers us to broaden our understanding of cooperation through characterizing high-level quantities, such as delay, throughput, fairness, and buffer length, as more tangible measures of the instant service quality that users/devices experience. In the pursuit of achieving this goal, the theory of BCMP queueing networks is leveraged. The proposed modeling approach can be used to analyze networks with an arbitrary number of relays, traffic classes, generic service time distributions, and several serving disciplines. To showcase the generality of this framework, we establish the queueing models for some of the most well-known cooperative protocols, such as amplify and forward amplify-forward, decode-forward, selection-relaying, incremental-relaying, and opportunistic-relaying, and evaluate their performances through the above-mentioned metrics. Moreover, a distributed cooperative protocol based on space–time block codes is proposed, its corresponding BCMP model is derived, and its performance is compared with other cooperative protocols.

INDEX TERMS Cooperative communications, amplify-forward, decode-forward, selection-relaying, incremental-relaying, space-time coded cooperation, opportunistic-relaying, BCMP theory, delay, throughput, fairness.

I. INTRODUCTION

Cooperation is a communication paradigm whereby a transmitting terminal is not necessarily the source of traffic. Instead, wireless terminals, hereinafter denoted by user equipment-UE, may relay each others' packets. This is an important, often very effective, concept in wireless communications, where channel impairments with unpredictable nature, such as frequency-selective fading and shadowing, prevent two UEs to successfully communicate with each other. The promise is that if the source UE (hereafter, dubbed S-UE) and relaying UE (hereafter, dubbed R-UE) are spatially well separated, it is very unlikely that the path between them and their common destination is simultaneously in a deep fading state.

The benefits of cooperation goes beyond increasing communications range. Cooperation can be employed (a) to deal with spectrum crunch problem in conjunction with cognitive radio (CR) (b), and to enhance capacity due to the spatial diversity. But, above all, is its application to 5G: For the first time, 5G is expected to connect *things* to network. These devices are battery-operated, thus, should run for long without needing to be recharged/maintained. Transmitting information to a distant base station (BS) can quickly drain a device battery. By forming a cooperating cluster with close-by relay terminals, battery endurance can be substantially enhanced [1]. Cooperation is even a live topic in cloud radio access networks (C-RANs) where the cost of baseband processing and power consumption can

be lowered by cooperative processing of signals in several BSs [2].

Notwithstanding its significance, it is unknown how this enhanced diversity gain translates into user-level experience. As a matter of fact, in almost all cases, cooperation was evaluated in the physical layer in the form of capacity and outage analysis [3]. To the best of our knowledge, this study is the first one to model cooperative communications from an upper layer perspective where performance is characterized by user-level metrics such as delay, throughput, jitter, etc. To that end, the BCMP[1] theory, introduced in [4] and recapitulated in this paper, is adopted.

A. MOTIVATIONS AND OBJECTIVES

Modeling and analysis of cooperative networks is an imperative undertaking. Over the years, this problem has been dealt with on different fronts. From an *Information Theory* perspective, cooperative communication has been extensively investigated through capacity and outage analysis [3], [5]–[8]. Later on, many studies used *Stochastic Geometry* to gain insight about connectivity and outage analysis of large collaborative networks [9]. While these models bring a new perspective into the problem, their accuracy heavily rely on assumptions made w.r.t. the topology and scale of the modeled network. Modeling cooperation using *Game Theory* is a completely different methodology [10], [11]. However, game theory is unable to determine how that deigned protocol performs in terms of packet-level metrics such as delay and throughput.

Using *Queuing Theory* to analyze wireless networks has a long history. Yet, there has been a recent surge of interest to apply queueing theory in order to comprehend the behavior of CR networks (CRNs). In particular, the fusion of cooperative communication and CR led to the emergence of a new trend of research works. Most available studies in this domain aim at a basic setting where there is only one primary and one secondary terminal in contention to gain access to a single channel. For instance, [12]–[14] use First-Come First-Serve (FCFS) queuing discipline whereas [15] and [16] employ priority queuing approach. The reason for the simplifying assumption in these works is obvious: Solving a model of interconnected queues is an NP-hard problem in most cases as it involved finding the steady-state solution (SSS) of the underlying Markovian process whose states expand exponentially with the number of queues and number of queue servers. While these constraints have been partially tackled in [17]–[20], they either rely on approximations such as fluid analysis [18] and large deviation [19] to find the first few moments of the queue statistics, or make explicit assumption on the exponentiality of queue service times to leverage the product-form solution [17], [20].

Over the years, open/closed BCMP theory has been employed to address important research questions from finding optimal network routing, CPU scheduling, and load balancing to modeling supply chains and manufacturing processes. Despite its generality and widespread applicability in other fields, BCMP theory has barely been leveraged to further our understanding of emerging wireless networks and protocols.

With this introduction, this paper's goal is to develop a packet-level understanding on cooperative communications. In this vein, the theory of BCMP queueing networks is employed [4]. The advantage of BCMP theory is in that it eliminates some of the aforementioned restrictions on other queueing models such as Jackson networks. In particular, BCMP theory enables us to model networks with (i) arbitrary UEs' connectivity (ii) arbitrary number of UEs (iii) state-dependent traffic arrival to UEs (iv) FCFS/non-FCFS queueing disciplines (v) generic/state-dependent UEs' service-time distributions (vi) closed/open structure. This study is the first to use BCMP theory to model cooperation in wireless networks. Some of the well-known cooperative protocols are adopted and their performances are evaluated: This includes amplify-forward (AF), decode-forward (DF), selection-relaying (SR), incremental-relaying (IR) [5], space-time coded cooperation (introduced in this paper), and multi-relay diversity (a.k.a. opportunistic-relaying [21]). To summarize, this paper's **contributions** are as follows:

- To recapitulate the classic BCMP theory in understandable terms.
- To introduce a streamlined methodology for mapping protocol features to queueing network parameters.
- To introduce a new space-time coded cooperation protocol that uses space-time block codes (STBC) to collaboratively forward source information.
- To build queueing models for three main categories of cooperative protocols: Repetition-based, space-time coded, and opportunistic.
- To formulate a comprehensive set of metrics: throughput, fairness, total latency and queue occupancy.
- To conduct extensive numerical simulations and discussions.

Section II provides the required background to understand the BCMP queuing model. Section III restates the product-form solution of a BCMP network. In Section IV, the BCMP theory is applied to model some of the most popular cooperative protocols. Section V deals with finding closed-form expressions for throughput, fairness metric, sojourn delay, and joint queue length PMF in the network. Finally, case studies are conducted and numerical results are presented in Section VI.

Readers' Guide: Readers exposed to queuing theory can skip Section II (except Subsection II-A). Readers advanced in queuing theory can skip forward to Corollary II, in Section III.

Notation: Unless otherwise stated, vectors (matrices) are shown in uppercase and random (deterministic) quantities in boldface (regular) font.

[1]The abbreviation BCMP is an initialism for authors' names in [4].

TABLE 1. Notation system in this paper.

Symb.	Definition	Symb.	Definition
N	Num. of queues (name $\mathcal{Q}_i$)	R	Num. of traffic classes (name $\mathcal{C}_i$)
$\mathbf{P}^{\mathrm{TPM}}$	Transition prob. matrix	$\mathbf{P}^{\mathrm{RPM}}$	Routing prob. matrix
$\mathbf{S}$ (S)	State of the net. (its realization)	$\mathbf{S}_i$ (S_i)	State of $\mathcal{Q}_i$ (its realization)
$T_s \backslash T_{i,j}$	Length of a time slot	T	Period of state of a CTMC
$n_i \backslash n_{i,r}$	Num. of pks. in $\mathcal{Q}_i$ (and from $\mathcal{C}_r$)	$m_{l,r}$	Num. of pks. from $\mathcal{C}_r$ in l^{th} stage of service
$b_{i,r}^k$	Coeff. in *Coxian* dist.	$a_{i,r}^{(j)}$	Complementary coeff. in *Coxian* dist.
$p_i^j (p_{i,r}^{j,s})$	Prob. that a pk. in $\mathcal{Q}_i$ (and from $\mathcal{C}_r$) transits to $\mathcal{Q}_j$ (and changes to $\mathcal{C}_s$)	$p_0^{i,r}$	Prob. that a pk. of $\mathcal{C}_r$ joins $\mathcal{Q}_i$
θ_k	k^{th} Eigenvalue of $\mathbf{P}^{\mathrm{TPM}}$	$P_i(n_i)$	Marginal length dist. of $\mathcal{Q}_i$
$P_{i,j}^{\mathrm{BER}}$	Bit error rate btw. UE_i and UE_j	$P_{i,j}^{\mathrm{Out}}$	Outage prob. btw. UE_i and UE_j
$\mathcal{M}(\mathsf{S}\|\mathcal{E}_j)$	Num. of pks. in sub-chain E_j	$\mathcal{M}(\mathsf{S})$	Total num. of pks. in the net.
$\lambda_i (\lambda_{i,r})$	Exogenous arrival rate to $\mathcal{Q}_i$ (from $\mathcal{C}_r$)	$\mathtt{T}_{i,r}$	Service time of a pk. of $\mathcal{C}_r$ in $\mathcal{Q}_i$
$\mu_i (\mu_{i,r})$	Service-rate of a pk. in $\mathcal{Q}_i$ (from $\mathcal{C}_r$)	$\Lambda_{i,r}$ $(\mathtt{e}_{i,r})$	Gross (Normalized) arrival rate of $\mathcal{C}_r$ to $\mathcal{Q}_i$
$\beta_{i,j}$	Channel gain btw. UE_i and UE_j	$\gamma_{i,j}$	SNR btw. UE_i and UE_j
ρ_i	Utilization factor of $\mathcal{Q}_i$	$\mathfrak{R}_p$	Transmission rate
$\mathtt{k}$	Const. in product-form	$D(\cdot)$	Multiplicative factor in product-form
R_i, F_i, S_i	Aux. queues for $\mathcal{Q}_i$	g_i, h_i, f_i	Multiplicative terms in product-form solution
$\mathcal{B}_{i,r}^{(l)}$	Factor in PS queues service time dist.	$\mathcal{A}_i^{\langle \mathrm{name} \rangle}$	Norm. factor of $\mathcal{Q}_i$ in protocol $\langle \mathrm{name} \rangle$

Abbrev.	Expansion	Abbrev.	Expansion	Abbrev.	Expansion
AF	Amplify-Forward	BC	Broadcast	BER	Bit Error Rate
CoMP	Coordinated Multipoint	CTMC	Continuous-time Markov Chain	DF	Decode-Forward
DTMC	Discrete-time Markov Chain	D2D	Device-to-Device	SD	Standard deviation
FCFS	Fist-Come First-Serve	GBE	Global Balance Equation	IR	Incremental-Relaying
LCFS	Last-Come First-Serve	LT	Laplace Transform	OR	Opportunistic-Relaying
PS	Processor Sharing	PDF	Prob. Density Function	PMF	Prob. Mass Function
QoS	Quality-of-Service	RPM	Routing Prob. Matrix	SR	Selection-Relating
R-UE	Relaying UE	RV	Random Variable	SSS	Steady State Solution
STBC	Space-Time Block Coding	S-UE	Source UE	STT	Space-Time Transmission
STCC	Space-Time Coded Cooperation	TBE	Traffic Balance Equation	TPM	Transition Prob. Matrix

II. BCMP NETWORKS: PRELIMINARIES

In this section, different BCMP queuing elements are introduced to prepare readers for the statement of the BCMP theory in the ensuing section. Our discussion includes BCMP queue model and its corresponding discrete-time Markov chain (DTMC), queue arrival processes, queue disciplines, queue departure process, and continuous-time Markov chain (CTMC) representation.

In a cooperating cluster, the number of queues and traffic classes are denoted by N and R, respectively. The probability that a packet from class r (call $\mathcal{C}_r$) changes type to class s (call $\mathcal{C}_s$) by transiting from i^{th} queue (call $\mathcal{Q}_i$) to j^{th} queue (call $\mathcal{Q}_j$) is represented by $p_{i,r}^{j,s}$. This is illustrated in the network of queues ($N = 2$ and $R = 2$) in Fig. 1a where only four out of 16 transitions are annotated with probabilities.

These transitions can be compactly represented by the routing probability matrix (RPM) $[p_{i,r}^{j,s}] = \mathbf{P}^{\mathrm{RPM}} \in \mathbb{R}^{NR \times NR}$. From a macroscopic perspective, the network in Fig. 1a is modeled as a discrete-time Markov chain (DTMC) comprising of states (i, r) and transitions $(i, r) \rightarrow (j, s)$, whose associated probabilities are $p_{i,r}^{j,s}$.[2] Fig. 1b depicts the corresponding DTMC model of Fig. 1a (16 transitions).

[2]Later on, we will see that there is a microscopic perspective into this network as a continuous-time Markov chain (CTMC), which characterize the interaction between queues in a more elaborate manner.

The sparsity of the RPM may result in a DTMC that is reducible to multiple disjoint sub-chains $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_m$, as shown in Fig. 1c. Representing the number of packets of class $\mathcal{C}_r$ in $\mathcal{Q}_i$ by $n_{i,r}$, the total number of packets in that j^{th} sub-chain is given by $\mathcal{M}\left(S|\mathcal{E}_j\right) = \sum_{(i,r)\in\mathcal{E}_j} n_{i,r}$. Consequently, the total number of packets in the network (called system state) is expressed by the sum of the number of packets in m disjoint sub-chains $\mathcal{E}_1, \ldots, \mathcal{E}_m$ (see Fig. 1c and 1d) as $\mathcal{M}(S) = \sum_{j=1}^{m} \mathcal{M}\left(S\middle|\mathcal{E}_j\right) = \sum_{j=1}^{m} \sum_{(i,r)\in\mathcal{E}_j} n_{i,r}$. For a soon to be clarified reason, $\mathcal{M}(S)$ is called the system state.

A. QUEUE ARRIVAL PROCESS

In classic queueing theory modeling, the arriving flow of traffic into the network is described with stochastic arrival processes. The theory of BCMP networks goes one step further by allowing arrival processes to be *state-dependent*. The assumption is that the arrival process of $\mathcal{C}_s$ traffic type into $\mathcal{Q}_k$ is *Poissonian*, hence, the exogenous arrival rate $\lambda_{k,s}$ sufficiently describes the whole process. For the sake of simplicity, it is assumed that there is a mainstream source of traffic; its intensity can depend on the total system state, i.e. $\lambda(\mathcal{M}(S))$, which splits the traffic between queues or sub-chains based on and external routing probabilities $p_0^{i,r}$, where $\sum_{(i,r)} p_0^{i,r} = 1$ (as shown in Fig. 1d).

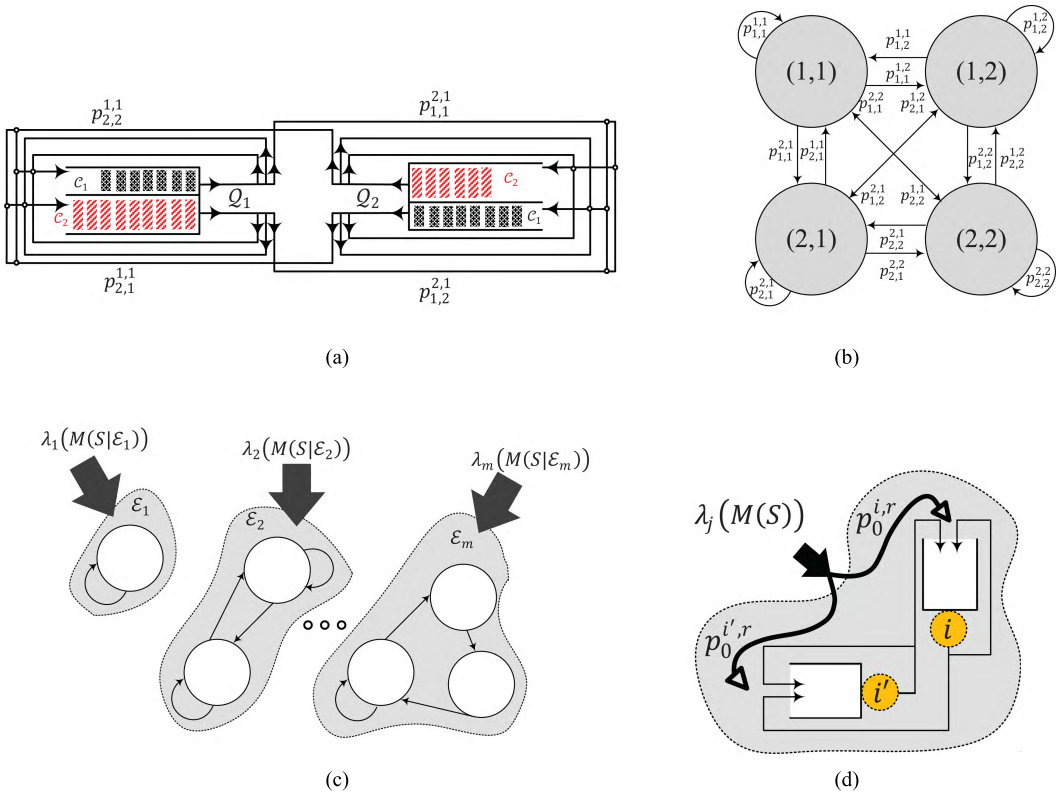

FIGURE 1. Two exemplary queue models and their corresponding DTMCs. (a) An exemplary network of $N = 2$ queues with $R = 2$ classes and 16 possible transitions. (b) DTMC associated with the network on the left. Each state is represented by (i, r), $1 \le i \le N$ and $1 \le r \le R$. (c) Disjoint sub-chains in a DTMC of a larger network. (d) Traffic routing into an open sub-chain with $N = 2$ queues.

Since there is circulating traffic within the network, the queue's received net arrival rate is larger than the exogenous arrival rate into it. In order to find the normalized gross arrival rate for $\mathcal{C}_s$ traffic in $\mathcal{Q}_k$, denoted by $\mathsf{e}_{k,s}$, a traffic balance equation (TBE) is to be written. With N queues in the network, N independent traffic balance equations are obtained to find N normalized gross arrival rates $\mathsf{e}_{k,s}$ and gross arrival rates $\Lambda_{k,s}$ as follows:

$$\begin{cases} \mathsf{e}_{k,s} = \mathbb{1}(\lambda_{k,s}) + \sum\limits_{(i,r)\in\mathcal{E}_j} \mathsf{e}_{i,r} p_{i,r}^{k,s}, \\ \Lambda_{k,s} = \mathsf{e}_{k,s}\lambda_{k,s}, \quad 1 \le k \le N,\ 1 \le s \le R,\ 1 \le j \le m, \end{cases} \tag{1}$$

where $\mathbb{1}(\lambda_{k,s})$ is the indicator function and is defined as

$$\mathbb{1}(\lambda_{k,s}) = \begin{cases} 1, & \lambda_{k,s} > 0 \\ 0, & \lambda_{k,s} = 0, \end{cases} \tag{2}$$

and m is the number of isolated sub-chains in the corresponding DTMC of the network (Fig. 1c) or, from another perspective, the number of block sub-matrices within the block diagonal form of $\mathbf{P}^{\mathrm{RPM}}$. Note that it was sufficient to restrict the summation in (1) to transitions within the j^{th} sub-chain since $p_{i,r}^{k,s} = 0$ if $i \in \mathcal{E}_a,\ k \in \mathcal{E}_{a'},\ a \ne a'$. In matrix form, (1) can be written as

$$\begin{aligned} \left(\mathbf{I} - \mathbf{P}^{\mathrm{RPM}}\right)^{\mathrm{T}} \times \mathsf{E} &= \mathbb{1}(\boldsymbol{\lambda}), \\ \Lambda &= \mathsf{E} \circ \boldsymbol{\lambda}, \end{aligned} \tag{3}$$

where $[\mathbf{I}]_{NR\times NR}$ is the identity matrix, $\Lambda = [\Lambda_1, \ldots, \Lambda_{NR}]^{\mathrm{T}}$, $\mathsf{E} = [\mathsf{e}_1, \cdots, \mathsf{e}_{NR}]^{\mathrm{T}}$, and $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_{NR}]^{\mathrm{T}}$ are the gross arrival rate vector, normalized gross arrival rate vector, and exogenous arrival rate vector, respectively. Also vector $\mathbb{1}(\boldsymbol{\lambda}) = [\mathbb{1}(\lambda_1), \cdots, \mathbb{1}(\lambda_{NR})]^{\mathrm{T}}$ is the binarized version of vector $\boldsymbol{\lambda}$. Finally, $^{\mathrm{T}}$ and $\circ$ are the matrix transpose and *Hadamard* vector product (element-wise multiplication), respectively. Since it is counterintuitive to have a transition from the traffic source (queue) back to its destination (queue), such transition should be marked with zeroed probability in $\mathbf{P}^{\mathrm{RPM}}$.

B. QUEUE DISCIPLINE

The theory of BCMP networks permits analyzing networks consisting of queues from four different queuing disciplines, namely, FCFS, processor sharing (PS), Last-Come First-Serve (LCFS), and queues with infinite number of servers. Hereinafter, we only discuss FCFS and PS, as these are relevant in the context of wireless networks. The main distinction between PS and FCFS disciplines is in how the packets of different lengths are treated.

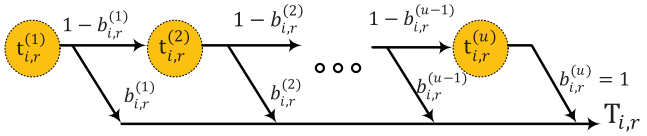

FIGURE 2. Structural representation of a RV with rational Laplace transform.

1) FCFS

As its name implies, the FCFS discipline gives strict serving priority to packets that arrive earlier. There is a number of restrictions imposed by the BCMP theory on queues with FCFS discipline: **(i)** even if it is possible to have multiple traffic classes in the network ($R > 1$), packets from different classes in $\mathcal{Q}_i$ are treated similarly, that is, served with service times $\mathtt{T}_{i,r}$ drawn from identical service time probability distribution functions (PDFs), $f_{\mathtt{T}_{i,r}}(t) = f_{\mathtt{T}_i}(t)$, **(ii)** $f_{\mathtt{T}_i}(t)$ should be exponentially distributed[3] with mean service rate $\mu_i^{-1} = \mathbb{E}\langle \mathtt{T}_i \rangle$ (not class-dependent), where **(iii)** μ_i can be load-dependent, i.e., $\mu_i\left(n_{i,r}\right)$.

2) PS

In this queuing discipline, there is no such thing as queuing delay since all packets in a queue are served immediately and simultaneously. More specifically, if the service rate is μ and there are k unit in the queue, each one is served with μ/k tasks per unit time. Subsequently, any addition (elimination) of a packet or a task reduces (increases) others' allocated capacities to $\mu/(k+1)$ (, $\mu/(k-1)$). In a similar fashion to FCFS, **(i)** there can be multiple traffic classes flowing through the network ($R > 1$). However, despite FCFS, **(ii)** each PS queue only has a single server, **(iii)** packets from different classes in a queue can be treated distinctly by that server (i.e. with service times drawn from non-identical PDFs $f_{\mathtt{T}_{i,r}}(t)$), **(iv)** and service times $\mathtt{T}_{i,r}$ can have generic distribution, albeit, with rational Laplace transform (LT). In contrary to FCFS discipline, the mean service rate $\mu_{i,r}^{-1} = \mathbb{E}\langle \mathtt{T}_{i,r} \rangle$ cannot be state-dependent in PS queues unless all classes in a queue are treated similarly. This will be clearer later on when we discuss the product-form solution of the network.

C. QUEUE DEPARTURE PROCESS

As aforementioned, one of the mild constraints of the BCMP theory with respect to PS queues is having service time distributions with *rational* LT. A class of probabilistic PDFs, known as *Coxian*, has the rational LT property. More concretely, the service time $\mathtt{T}_{i,r}$ for class $\mathcal{C}_r$ in $\mathcal{Q}_i$ is a random variable (RV) with Coxian distribution if it can be expressed as the parallel/serial combination of u exponential RVs $\mathtt{t}_{i,r}^{(1)}, \ldots, \mathtt{t}_{i,r}^{(u)}$ as shown in Fig. 2.

[3]In many circumstances, the exponential assumption leads to an upper bound on the performance. This is particularly true in the prevalent class of contention-free networks, such as TDMA, FDMA, and OFDM, where the service that a user receives in a given interval has little variability. It is crucial for readers to note that it is the variability in packet inter-departure and inter-arrival times that causes the queue build-up.

D. CTMC REPRESENTATION

For N UEs and R traffic classes, the network of queues can be modeled by using a continuous-time Markov chain (CTMC). Note that this CTMC is a packet-level model whereby each state represents the number of packets of different classes in a given queue. A state of this CTMC is denoted by $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_N)$, where $\mathbf{S}_i$ is the state of $\mathcal{Q}_i$. For FCFS queues, $\mathbf{S}_i^{\text{FCFS}} = \left(\zeta_{i,1}, \ldots, \zeta_{i,n_i}\right)$, where n_i is the number of packets at $\mathcal{Q}_i$ and $1 \leq \zeta_{i,j} \leq R$ indicates which traffic class j^{th} packet in $\mathcal{Q}_i$ belongs to. According to this definition, the state of the exemplary FCFS queue in Fig. 3a is $\mathbf{S}_i^{\text{FCFS}} = (1, 2, 2, 1, 1, 2, 2, 1, 2)$.

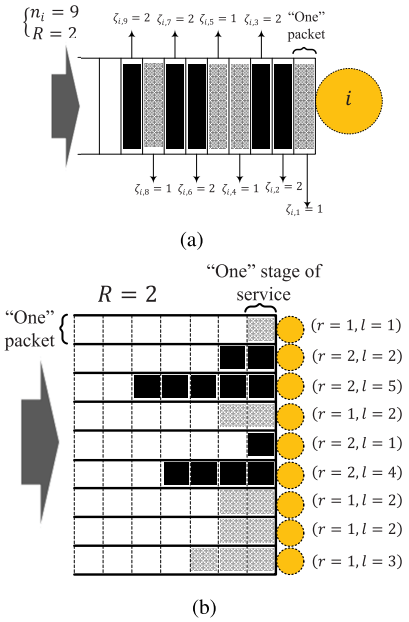

FIGURE 3. BCMP state definition for FCFS and PS queues. (a) FCFS queue. (b) PS queue.

The BCMP theory adopts a different state definition for PS network of queues. Once again, the overall system state is defined by $\mathbf{S}^{\text{PS}} = (\mathbf{S}_1, \ldots, \mathbf{S}_N)$. Each queue's state is an R-tuple represented by $\mathbf{S}_i^{\text{PS}} = (\mathbf{N}_{i,1}, \ldots, \mathbf{N}_{i,r}, \ldots, \mathbf{N}_{i,R})$. One layer inside, r^{th} element of $\mathbf{S}_i^{\text{PS}}$ is, by itself, a vector $\mathbf{N}_{i,r} = (\mathbf{m}_{1,r}, \ldots, \mathbf{m}_{l,r}, \ldots, \mathbf{m}_{u_{i,r},r})$ whose l^{th} element, $\mathbf{m}_{l,r}$, denotes the number of packets from $\mathcal{C}_r$ in $\mathcal{Q}_i$ that are in l^{th} stage of their Coxian service time. The length of $\mathbf{N}_{i,r}$ vector, i.e. $u_{i,r}$, is the total number of stages that the Coxian service time of $\mathcal{C}_r$ packets in $\mathcal{Q}_i$ is broken down into when it is expressed in the form shown in Fig. 2. Remember that the BCMP theory allows generic service time distributions with rational LT for PS queues explaining why the above state definition is more involved than FCFS. Fig. 3b illustrates this state definition for a PS queue, where nine packets from $R = 2$ classes (with $u_{i,1} = u_{i,2} = 7$) are simultaneously served. According to the above state definition for PS, one can see that $\mathbf{N}_{i,1} = (1, 3, 1, 0, 0, 0, 0)$ and $\mathbf{N}_{i,2} = (1, 1, 0, 1, 1, 0, 0)$.

To solve this CTMC, sufficient number of independent global balance equations (GBE) should be written. For the

above extremely elaborate state definitions of PS and FCFS queues, the number of states grows **exponentially** with N and R. Therefore, even for a moderate size network of queues, it is computationally unmanageable to evaluate the performance by directly solving the CTMC through solving GBEs. The BCMP theroy [4], proves that a unique product-form solution exists for the joint PDF of $\mathbf{S}$ in the network of queues with characteristics delineated in subsection II-B.

III. BCMP NETWORKS: PRODUCT-FORM

In a network of queues of FCFS and PS types, the equilibrium steady state PDF of the network state $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_N)$ has the following product-form:

$$f_{\mathbf{S}_1,\ldots,\mathbf{S}_N}\left(\mathsf{S}_1, \ldots, \mathsf{S}_N\right) = \mathtt{k}_1\, D\left(\mathcal{M}\left(S\right)\right) f_1\left(\mathsf{S}_1\right) \cdots f_N\left(\mathsf{S}_N\right), \tag{4}$$

where $\mathtt{k}_1$ is a constant, $D\left(\mathcal{M}\left(S\right)\right)$ is a multiplicative factor that is a function of total number of packets in the network, i.e. $\mathcal{M}\left(S\right) = \sum_{i=1}^{N} n_i$, and terms $f_i\left(\cdot\right)$ only depend on the state of $\mathcal{Q}_i$ and its serving discipline. Based on the state definition in subsection II-D, for FCFS queuing network,

$$f_i^{\text{FCFS}}\left(\mathsf{S}_i\right) = \left(\frac{1}{\mu_i}\right)^{n_i} \prod_{j=1}^{n_i} \mathsf{e}_{i,\zeta_{i,j}}, \tag{5}$$

where $n_i = \sum_{r=1}^{R} n_{i,r}$, and $\mathsf{e}_{i,\zeta_{i,j}}$ is the net arrival rate obtainable from (1) and (3). For the PS queuing network,

$$f_i^{\text{PS}}\left(\mathsf{S}_i\right) = n_i! \prod_{r=1}^{R} \prod_{l=1}^{u_{i,r}} \left(\left(\frac{\mathsf{e}_{i,r}\mathcal{B}_{i,r}^{(l)}}{\mu_{i,r}^{(l)}}\right)^{m_{l,r}} \cdot \frac{1}{m_{l,r}!}\right). \tag{6}$$

where $n_i = \sum_{r=1}^{R} \sum_{l=1}^{u_{i,r}} m_{l,r}$. Factors $\mathcal{B}_{i,r}^{(l)} = \prod_{j=1}^{l}\left(1 - b_{i,r}^{(j)}\right)$ reflect the fact that the PS queues' service time distribution is expressed in the Coxian form (see Fig. 2). Similarly, normalized gross arrival rates $\mathsf{e}_{i,r}$ are obtained from solving TBE in (3). Using the second axiom of the probability, it can be shown that, for an open network, $D(\mathcal{M}(S))$ in (4) is obtained as

$$D\left(\mathcal{M}\left(S\right)\right) = \prod_{j=1}^{m} \prod_{i=0}^{\mathcal{M}(S|\mathcal{E}_j)-1} \lambda_j\left(i\right), \tag{7}$$

where, according to the arrival model introduced in subsection II-A, $\lambda_j\left(i\right)$ is the state-dependent aggregate packet arrival rate to the j^{th} sub-chain.

A. COROLLARY I

At the cost of losing some information, aggregation of the form $n_{i,r} = \sum_{j=1}^{n_i} \mathbb{1}_{(\zeta_{i,j}=r)}\left(j\right)$ (for FCFS model) and $n_{i,r} = \sum_{l=1}^{u_{i,r}} m_{l,r}$ (for PS model) leads to the simpler state definition $\mathbf{S} = (\mathbf{S}_1, \ldots, \mathbf{S}_N)$ where $\mathbf{S}_i = \left(n_{i,1}, \ldots, n_{i,R}\right)$. This choice transforms the product-form solution to [4]:

$$f_{\mathbf{S}_1,\ldots,\mathbf{S}_N}\left(\mathsf{S}_1, \ldots, \mathsf{S}_N\right) = \mathtt{k}_2 D\left(\mathcal{M}\left(S\right)\right) g_1\left(\mathsf{S}_1\right) \cdots g_N\left(\mathsf{S}_N\right), \tag{8}$$

wherein, for FCFS queues with state-dependent service-rates,

$$g_i^{\text{FCFS}}\left(\mathsf{S}_i\right) = \left(n_i! \prod_{r=1}^{R} \frac{\mathsf{e}_{i,r}{}^{n_{i,r}}}{n_{i,r}!}\right) \prod_{j=1}^{n_i} \mu_i(j), \tag{9}$$

and, for the PS queues,

$$g_i^{\text{PS}}\left(\mathsf{S}_i\right) = n_i! \prod_{r=1}^{R} \frac{1}{n_{i,r}!} \left(\frac{\mathsf{e}_{i,r}}{\mu_{i,r}}\right)^{n_{i,r}}, \tag{10}$$

where $n_i = \sum_{r=1}^{R} n_{i,r}$ is the total number of packets in $\mathcal{Q}_i$ and $D\left(\mathcal{M}\left(S\right)\right)$ is given by (7).

An intriguing observation in (10) is that the join state PDF of the PS queuing network only depends on the mean service rates $\mu_{i,r}$ and not any higher moment, a rare attribute a.k.a. insensitivity [22]. As will be shown in Subsection V-B, when the queue input processes are state-independent and $N = R$, the closed-form in (8) for PS queues can be simplified to a very intuitive relation.

B. COROLLARY II

The state space of queues can even be further shrunk (aggregated). In fact, by agreeing to loose class information, network state becomes $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_N)$, where $\mathbf{s}_i = n_i = \sum_{r=1}^{R} n_{i,r}$, and the joint PDF holds the following product-form:

$$f_{\mathbf{s}_1,\ldots,\mathbf{s}_N}\left(\mathsf{s}_1, \ldots, \mathsf{s}_N\right) = \mathtt{k}_3 D\left(M\left(\mathsf{S}\right)\right) h_1\left(n_1\right) \cdots h_N\left(n_N\right), \tag{11}$$

where, for the FCFS queues,

$$h_i^{\text{FCFS}}\left(\mathsf{s}_i = n_i\right) = \left(\sum_{r \in R} \mathsf{e}_{i,r}\right)^{n_i} \frac{1}{\mu_i^{n_i}}, \tag{12}$$

and, for the PS queues,

$$h_i^{\text{PS}}\left(\mathsf{s}_i = n_i\right) = \sum_{r \in R} \left(\frac{\mathsf{e}_{i,r}}{\mu_{i,r}}\right)^{n_i}. \tag{13}$$

Once again, $D\left(\mathcal{M}\left(S\right)\right)$ is obtained using (7). For an open network with state-independent arrival rate, factor $\mathtt{k}_3$ is calculable using the second probability axiom, resulting in the following simple product-form solution:

$$f_{\mathbf{s}_1,\ldots,\mathbf{s}_N}\left(\mathsf{s}_1, \ldots, \mathsf{s}_N\right) = \prod_{i=1}^{N} P_i\left(n_i\right), \tag{14}$$

where $P_i\left(n_i\right) = \left(1 - \rho_i\right)\rho_i^{n_i}$ and ρ_i is called the utilization factor of $\mathcal{Q}_i$. Note that $\rho_i^{\text{FCFS}} = \left(1/\mu_i\right)\sum_{r \in R} \Lambda_{i,r}$ and $\rho_i^{\text{PS}} = \sum_{r \in R} \Lambda_{i,r}/\mu_{i,r}$. Gross arrival rates $\Lambda_{i,r}$ are obtained from (3). Needless to mention that, for the network of queues to be stable and have SSS in (14) (ergodic CTMC), $0 < \rho_i < 1, \forall i \in \{1, \ldots, N\}$.

IV. MODELING COOPERATION USING BCMP THEORY

In this section, the BCMP theory is applied to model three main classes of relaying protocols and analyze their performances. The first class, called **repetition-based cooperation**, is based on the concept of time orthogonality, where relays use non-overlapping sub-intervals to forward a source packet (Fig. 4a, 5b). The second class, called **space-time coded cooperation**, is based on the concept of code orthogonality, where relays can concurrently forward source traffic using an intelligent transmit-diversity technique, known as STBC (Fig. 4b, 5c). The third class, known as **opportunistic relaying**, nominates the best candidate from a pool of relays to perform the relaying task (Fig. 4c). Before delving into modeling protocols from each class, some important remarks are in order:

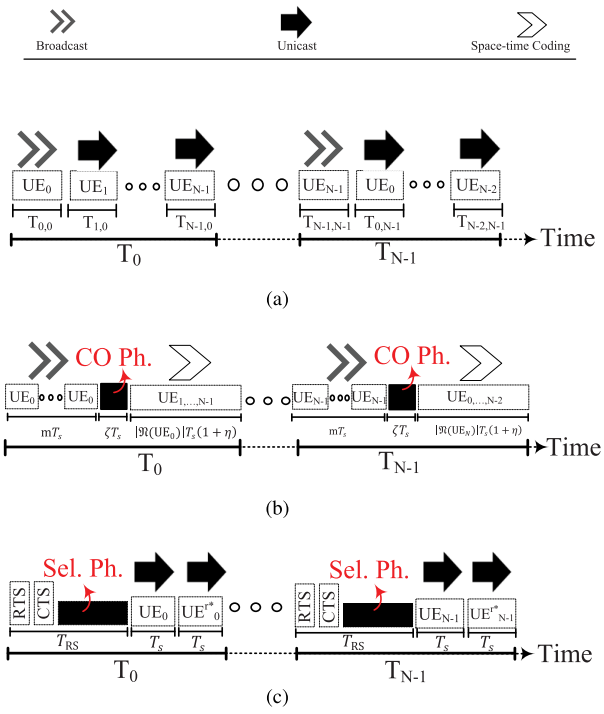

FIGURE 4. Frame structure of three different cooperative paradigms. During slots BC, UC, and STT, information is broadcast, unicast, and coded by space-time transmission, respectively. No payload is communicated during coordination (CO) and relay selection (Sel.) intervals. (a) Repetition-based coding. (b) Space-time coding. (c) Opportunistic-relaying.

Remark 1 (Queue Types in the Model): Two types of queues are utilized in the presented models, i.e., *data* and *probing* queues. Although both types hold payload packets and have the same serving discipline, data queues $\mathrm{UE}_0, \ldots, \mathrm{UE}_{N-1}$ are deemed integral to the modeling task, whereas probing queues (hereinafter, dubbed S_0 and F_0) are auxiliary entities that are introduced to the models for performance monitoring. Moreover, since PS queuing discipline more closely resembles the way traffic of different classes are treated in cooperative networks, it is utilized to model data queues in this paper.

Remark 2 (Model-Protocol Adaptation): The stochasticity of a wireless channel can be well captured by the transition probability matrix (TPM) in the queueing network. With this in mind, protocol details can be reflected onto the following three elements: **(i)** routing probabilities p_i^j (associated with transition from $\mathcal{Q}_i$ to $\mathcal{Q}_j$)[4] **(ii)** service rates $\mu_{i,r}$ (associated with UE_i indicating how fast $\mathcal{C}_r$ data packets can be transmitted), and, if necessary, **(iii)** auxiliary (probing) queues.

Remark 3 (Multipath Fading): The lowest level of abstraction in the proposed models is the outage probability $P_{i,j}^{\mathrm{Out}} = \Pr(\gamma_{i,j} < \gamma^{\mathrm{Out}})$ between UE_i and UE_j, where $\gamma_{i,j}$ is the instantaneous SNR and γ^{Out} is the receiver sensitivity. While such choice was made to avoid unnecessary involvement with system-level parameters, all channel degradation (e.g. small-scale fading, shadowing, and large-scale path loss) manifest themselves in $P_{i,j}^{\mathrm{Out}}$ and there is no restriction on the channel model.

Remark 4 (Time-Selective Fading): The time-selective nature of wireless channel (a.k.a Doppler fading) shall be reflected as a time-dependent $\mathbf{P}^{\mathrm{RPM}}(t)$ in the equivalent model. Moreover, time variability forces the allocation mechanism to allocate resources (adaptively to maximize the total capacity, which also results in time-dependent service rates $\mu_{i,r}(t)$. Given that (for quasi-static channels) such variations happen *discretely*, say every $t_0$sec, BCMP model may be applicable but the SSS is stirred up every time channel changes. Once the latter happens, it takes a while until the transient behavior of the CTMC is vanished and the new SSS is reached. For a BCMP model to remain valid, the channel coherence time t_0 should be larger than the convergence time τ of the model, i.e. $\tau \ll t_0$. This is the case because the CTMC converges exponentially fast to its SSS whereas wireless channel varies slowly in indoor/outdoor environment.

Remark 5 (Combination): Since cooperation depends not only on the way information is disseminated by relays but also on how it is combined at the destination, probabilities associated with a model may vary. In this paper, we adopt selection combining (SC) scheme [23] where destination selects only the received signal whose SNR ($\gamma_{j,N}$) is the highest. In such situation, the cooperation failure probability becomes $p_0^{\mathrm{F}_0} = \Pr(\max(\gamma_{0,N}, \ldots, \gamma_{N-1,N}) < \gamma^{\mathrm{Out}}) \propto \prod_{j=1}^{N}(P_{0,j}^{\mathrm{Out}})$. An extension to other combining methods is to be investigated.

A. REPETITION-BASED COOPERATION

Repetition-based cooperation is based on the concept of *time orthogonality*. As depicted in Fig. 4a, the entire cooperation process is comprised of two consecutive phases: Broadcast (BC) and Unicast. During the BC time slot, traffic source broadcasts the packet. Thereafter, during unicast phase, this transmission, which is overheard by some cooperators,

[4]Since payload packets do not change their types/classes as they traverse through the wireless network (i.e. $s = r$), subscript $_r$ and superscript ^{s} are eliminated from $p_{i,r}^{j,s}$, hereinafter.

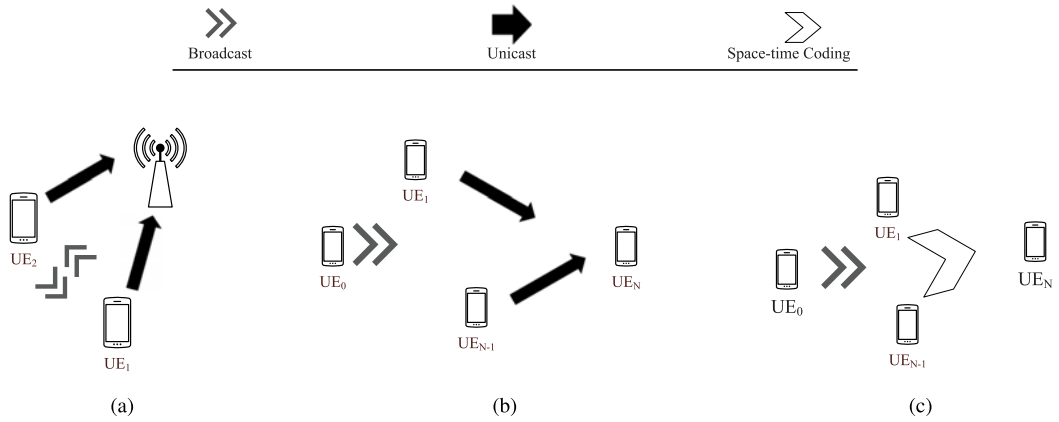

FIGURE 5. Cooperation in wireless networks. (a) $N = 2$ UEs cooperate in uplink. (b) Repetition-based coding scenario. (c) Space-time coding scenario.

is relayed to the destination, according to some predefined strategy.

Let's represent the length of the j^{th} slot (sub-interval) in the i^{th} interval by $T_{j,i}$ and the number of R-UEs by $N-1$. As illustrated in Fig. 4a, a transmission from the i^{th} UE takes $T_i = \sum_{j=0}^{N-1} T_{j,i}$ seconds to complete. According to Fig. 5b, N UEs help each other out in forwarding one another's transmissions. Therefore, each UE transmits (own packets + others) in exactly $1/N$ of the total channel's degree of freedom (DoF). Also, in both non-cooperative and cooperative paradigms, the fraction utilized for transmitting a given UE's traffic is equal to $1/N$ of the total channel DoF.

Several famous protocols in this class are amplify-forward (AF), decode-forward (DF), selection relaying (SR), and incremental relaying (IR). The rest of this subsection deals with establishing BCMP models for each of these protocols.

1) AMPLIFY-FORWARD (AF) / DECODE-FORWARD (DF)

In AF protocol, the transmission process is completed in several time slots. In the first slot, a packet is transmitted by the source UE (S-UE), which might be overheard by a number of relaying UEs (R-UEs) and received by the destination UE_N. Next, R-UEs take turn to amplify the analog received signal and retransmit it to the destination. DF is similar in operation to AF except that, in the former, R-UEs will have to sample, demodulate and decode the signal first. Thereby, only if an R-UE is able to correctly decode the overheard signal, it will re-encode/re-modulate/retransmit it in its corresponding time slot. Fig. 6a and 6b are the corresponding BCMP queue models for the cooperative clusters in Fig. 5b when AF and DF protocols are employed, respectively. For tidiness of illustrations, *only transitions related to UE_0 being the source are drawn*. The eliminated transitions associated with other UEs being the source can simply be deduced from the following derivation logic.

Transition Prob.: To have a correct model, the total probability space for $\mathcal{Q}_i$, which contains all the events that can happen within T_i, is partitioned into N mutually exclusive and collectively exhaustive (MECE) probabilistic subsets $\{E_0, \ldots, E_{N-1}\}$, where E_k only encompasses events that can take place within the sub-interval (slot) $T_{k,i}$ (see Fig. 4a). Noting that $\Pr(E_k) = 1/N$ and $\Pr(E_k|E_i) = 1/(N-1)$, $i \neq k$, the transition probability p_i^j for the AF model is obtained by

$$\begin{aligned} p_i^j &= \sum_{k=0}^{N-1} \Pr\left(\{\text{R-UE}_j \text{ relays for UE}_i\} \cap E_k | E_i\right) \\ &= \Pr\left(\{\text{R-UE}_j \text{ relays for UE}_i\} \cap E_j | E_i\right) \\ &= \frac{1}{(N-1)\mathcal{A}_i^{\text{AF}}} \Pr\left(\{\text{R-UE}_j \text{ relays for UE}_i\}\right) \\ &= \frac{1}{(N-1)\mathcal{A}_i^{\text{AF}}}, \end{aligned} \tag{15}$$

where $\mathcal{A}_i^{\text{AF}}$ is the normalization factor. This factor is introduced for the following reason: With $N-1$ R-UEs, 2^{N-1} combinatorial cases can be imagined, whereof only $N-1$ corresponds to transitions to only one R-UE at a time and the remaining $2^{N-1} - N + 1$ intersecting cases are trivial (have low probability). Therefore, for the sake of retaining the model's simplicity, these trivial cases are eliminated which necessitates the inclusion of the non-unit normalization factor $\mathcal{A}_i^{\text{AF}}$ to satisfy the second probability axiom, that is, $\{\mathcal{A}_{i=0}^{\text{AF}} | \sum_{j=1}^{N-1} p_0^j + p_0^{\text{F}_0} + p_0^{\text{S}_0} = 1\}$. In case of DF where not all R-UEs are able to decode

$$p_i^j = \frac{1}{N-1} \Pr\left(\{\text{R-UE}_j \text{ relays for UE}_i\}\right) = \frac{\overline{P_{i,j}^{\text{BER}}}}{(N-1)\,\mathcal{A}_i^{\text{DF}}}, \tag{16}$$

where $P_{i,j}^{\text{BER}} = \Pr(\text{BER}_{i,j} > \text{BER}^{\text{Th}})$ is the probability that the bit error rate (BER) on the communication link from UE_i to UE_j is larger than the decoding threshold BER^{Th} and

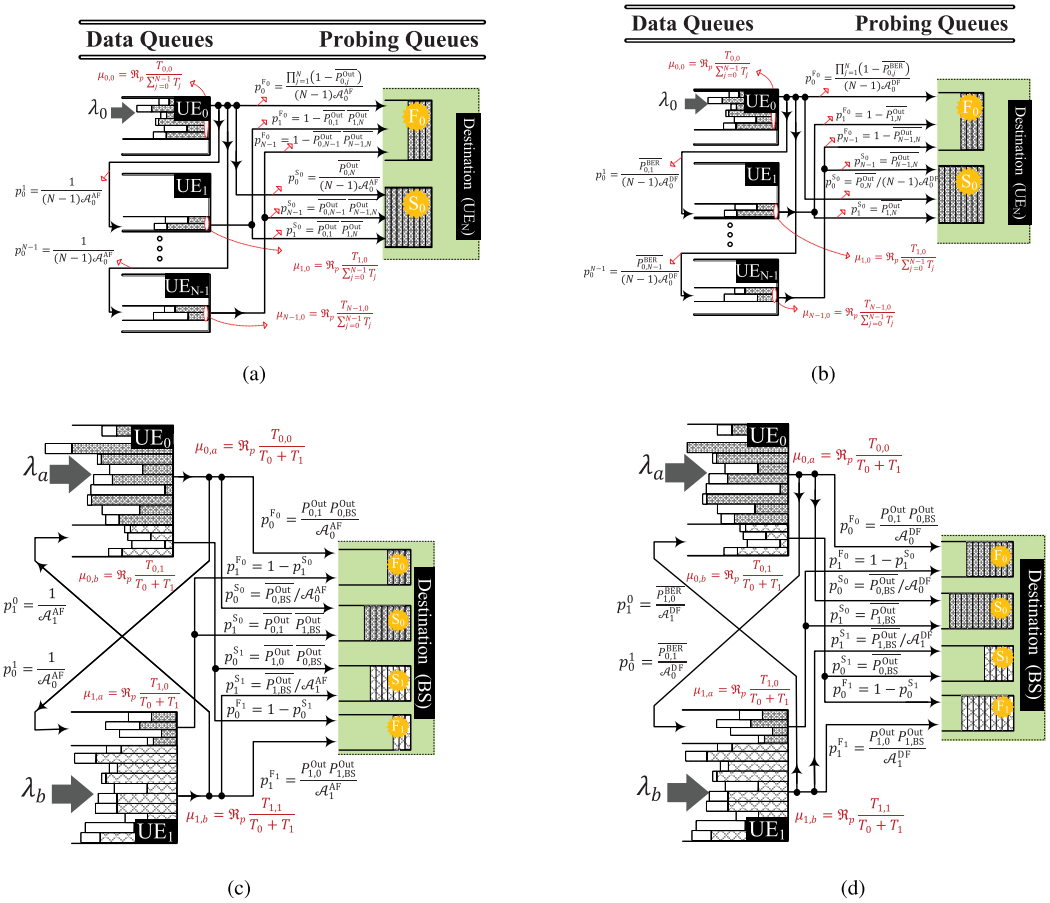

FIGURE 6. Corresponding BCMP models for cooperating clusters in Fig. 5a, 5b. For the sake of tidiness, in general cases "a" and "b", only transitions emanating from UE_0 are drawn. (a) Partial BCMP model for the cooperating cluster in Fig. 5b with AF relaying. (b) Partial BCMP model for the cooperating cluster in Fig. 5b with DF relaying. (c) Full BCMP model for the cooperating cluster in Fig. 5a ($N = 2$) with AF relaying. (d) Full BCMP model for the cooperating cluster in Fig. 5a ($N = 2$) with DF relaying.

$\overline{P_{i,j}^{\text{BER}}} = 1 - P_{i,j}^{\text{BER}}$ is its complementary probability.[5] Since BER is monotonically decreasing function of signal-to-noise ratio (SNR), i.e. $\partial \text{BER}_{i,j}/\partial \gamma_{i,j} < 0$, it is valid to say

$$\underbrace{\Pr(\text{BER}_{i,j} < \text{BER}^{\text{Th}})}_{\overline{P_{i,j}^{\text{BER}}}} = \underbrace{\Pr(\gamma_{i,j} > \gamma^{\text{Out}})}_{\overline{P_{i,j}^{\text{Out}}}} - \overbrace{\Delta^{\text{DF}}}^{>0}. \quad (17)$$

where $P_{i,j}^{\text{Out}}$ is the outage probability between UE_i and UE_j. Once again, due the exclusion of the trivial cases, a normalization factor is introduced which should satisfy $\{\mathcal{A}_{i=0}^{\text{DF}} | \sum_{j=1}^{N-1} p_0^j + p_0^{\text{F}_0} + p_0^{\text{S}_0} = 1\}$.

The probabilities associated with the remaining transitions are attained as follows for the AF/DF BCMP models in Fig. 6[6]

[5]Note that when the channel is dispersive both $\gamma_{i,j}$ and $\text{BER}_{i,j}$ are stochastic, thus, probability of BER (which is, itself, a probability) is meaningful.

[6]Later on, the same logic can be followed to obtain these exit probabilities for SR and IR BCMP models in Fig. 7-8.

(i) A transmission from traffic source UE_0 (in broadcast sub-interval) may be directly decoded by the destination. Such event is modeled by a transition to the auxiliary queue S_0 whose probability $p_0^{\text{S}_0} \propto \overline{P_{0,N}^{\text{Out}}}$ is the strength of the direct channel. **(ii)** On the other hand, the cooperation failure probability $p_0^{\text{F}_0}$ associated with transiting from UE_0 to the auxiliary failure queue F_0 varies depending on the protocol. For instance, in case of AF protocol, $p_0^{\text{F}_0} \propto \prod_{j=1}^{N}(1 - \overline{P_{0,j}^{\text{Out}}})$ as noted in Fig. 6a. This choice is to emphasize that none of R-UEs forwarding for UE_0 (denoted by $\mathfrak{R}(\text{UE}_0)$) nor the destination UE_N (denoted by $\mathfrak{D}(\text{UE}_0)$) could successfully receive the packet (due to fading) after the broadcasting sub-interval $T_{0,0}$. Having explained the logic behind the derivations of the above-mentioned probabilities in Fig. 6, probabilities associated with other transitions are not hard to derive.

Service Rates: Once the transition probabilities are known, the last modeling step is to characterize service rates $\mu_{i,r}$. Let's concentrate on interval T_0, where UE_0 is the S-UE. Representing transmission rate by $\mathfrak{R}_p$pk/s,

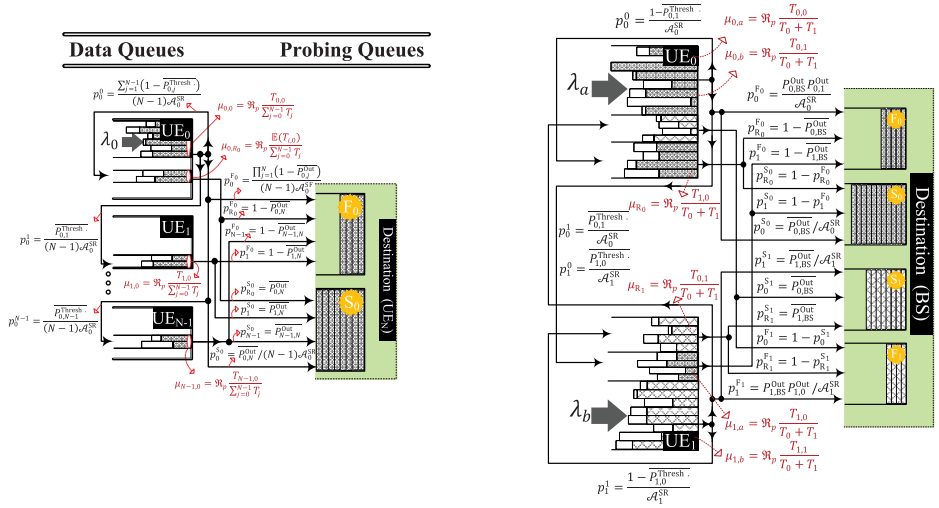

FIGURE 7. The complete and partial queue models for the configurations in Fig. 5a, 5b with SR relaying. (a) BCMP model for the cooperating cluster in Fig. 5b with SR relaying. (b) BCMP model for the cooperating cluster in Fig. 5a ($N = 2$) with SR relaying.

the total of $\Re_p T_{i,0}$ packets are relayed by UE$_i$. Since this is repeated every $\sum_{j=0}^{N-1} T_j$ sec., then UE$_i$'s service rate would be $\Re_p T_{i,0} / \sum_{j=0}^{N-1} T_j$ pk/s. For the sake of clarity, all the service rates are shown by red color inside the BCMP models in this paper.

2) SELECTION-RELAYING (SR)

SR protocol [5] makes more efficient use of resources by requiring R-UE$_j$ to forward S-UE's traffic only if the source-relay channel quality is better than a target threshold $\gamma^{\text{Thresh.}}$, otherwise, the source retransmits the packet in the time slot dedicated to R-UE$_j$. This way a trade-off is created between time-diversity and spatial-diversity.

Transition Prob.: In SR's corresponding BCMP model, the probabilities associated with transitions from UE$_0$ to relaying set $\mathfrak{N}$ (UE$_0$) are proportional to their corresponding channel qualities, that is to say $p_0^j = \overline{P_{0,j}^{\text{Thresh.}}}/((N-1)\mathcal{A}_0^{\text{SR}})$. If neither relays $\mathfrak{N}$(UE$_0$) not the destination $\mathfrak{D}$(UE$_0$) could decode the packet, the latter is considered garbled, joining queue F$_0$ with probability $p_0^{\text{F}_0} = \prod_{j=1}^{N}(1-\overline{P_{0,j}^{\text{Out}}})/((N-1)\mathcal{A}_0^{\text{SR}})$. This model is shown in Fig. 7a.

Note that the choice of the SNR threshold is such that $\gamma^{\text{Thresh.}} > \gamma^{\text{Out}}$, thus,

$$\underbrace{\Pr\left(\gamma_{i,j} > \gamma^{\text{Thresh.}}\right)}_{\overline{P_{i,j}^{\text{Thresh.}}}} = \underbrace{\Pr\left(\gamma_{i,j} > \gamma^{\text{Out}}\right)}_{\overline{P_{i,j}^{\text{Out}}}} - \overbrace{\Delta^{\text{SR}}}^{>0}, \qquad (18)$$

Similar to the logic applied in (15)-(16) for AF and DF, factor $N-1$ in the denominator of p_0^j is to account for the fact that R-UE$_j$ can only relay for UE$_0$ in $1/(N-1)$ fraction of the event space. Since the S-UE has to retransmit a packet when relaying channel quality is poor, there is a transition from UE$_0$ back to itself denoted with probability p_0^0. Because the source-relays wireless channels are statistically independent and a data packet may be retransmitted by the S-UE during sub-intervals $T_{i,0}$, $p_0^0 = \sum_{j=1}^{N-1}\left(1-\overline{P_{0,j}^{\text{Thresh.}}}\right)/((N-1)\mathcal{A}_0^{\text{SR}})$. The normalization factor is obtained by solving $\{\mathcal{A}_0^{\text{SR}} : p_0^{\text{F}_0} + p_0^{\text{S}_0} + \sum_{j=0}^{N-1} p_0^j = 1\}$.

Service Rates: The characterization of service rates remains similar to AF and DF protocols. Based on the above logic, the BCMP queue models for a SR-based cooperating cluster with unknown N and $N = 2$ UEs are shown in Fig. 7a and 7b, respectively. In the former case, only transitions related to UE$_0$ being the source are drawn. One should note that the SR protocol achieves full diversity order for the fact that both source-relay and source-destination channels should be corrupted for the destination UE not to be able to decode the packet.

3) INCREMENTAL-RELAYING (IR)

To further improve the spectral efficiency of cooperation, an IR protocol is proposed in [5] that directly includes the destination in making the decision that whether relaying is needed or not. The IR mechanism works as follows: In broadcasting a packet during $T_{0,0}$ (Fig. 4a), if $\mathfrak{D}$(UE$_0$) can decode it, no R-UE is required to forward that packet ($\mathfrak{N}$(UE$_0$) $= \emptyset$) saving power and spectrum resources. On the contrary, if the $\mathfrak{D}$(UE$_0$) was not able to decode the packet, relaying phase initiates whereupon $\mathfrak{N}$ (UE$_0$) are directly solicited by $\mathfrak{D}$(UE$_0$) to undertake the forwarding task. Such solicitation can be handled by transmitting a limited feedback (a single bit suffices) in the reverse path.

Depending on how packet forwarding is undertaken by R-UEs, two types of IR exist. In type-I, R-UEs $\in \mathfrak{N}$ (UE$_0$) are instructed to take turn and forward the packet during the remaining sub-intervals $T_{1,0}, \ldots, T_{N-1,0}$. In type-II,

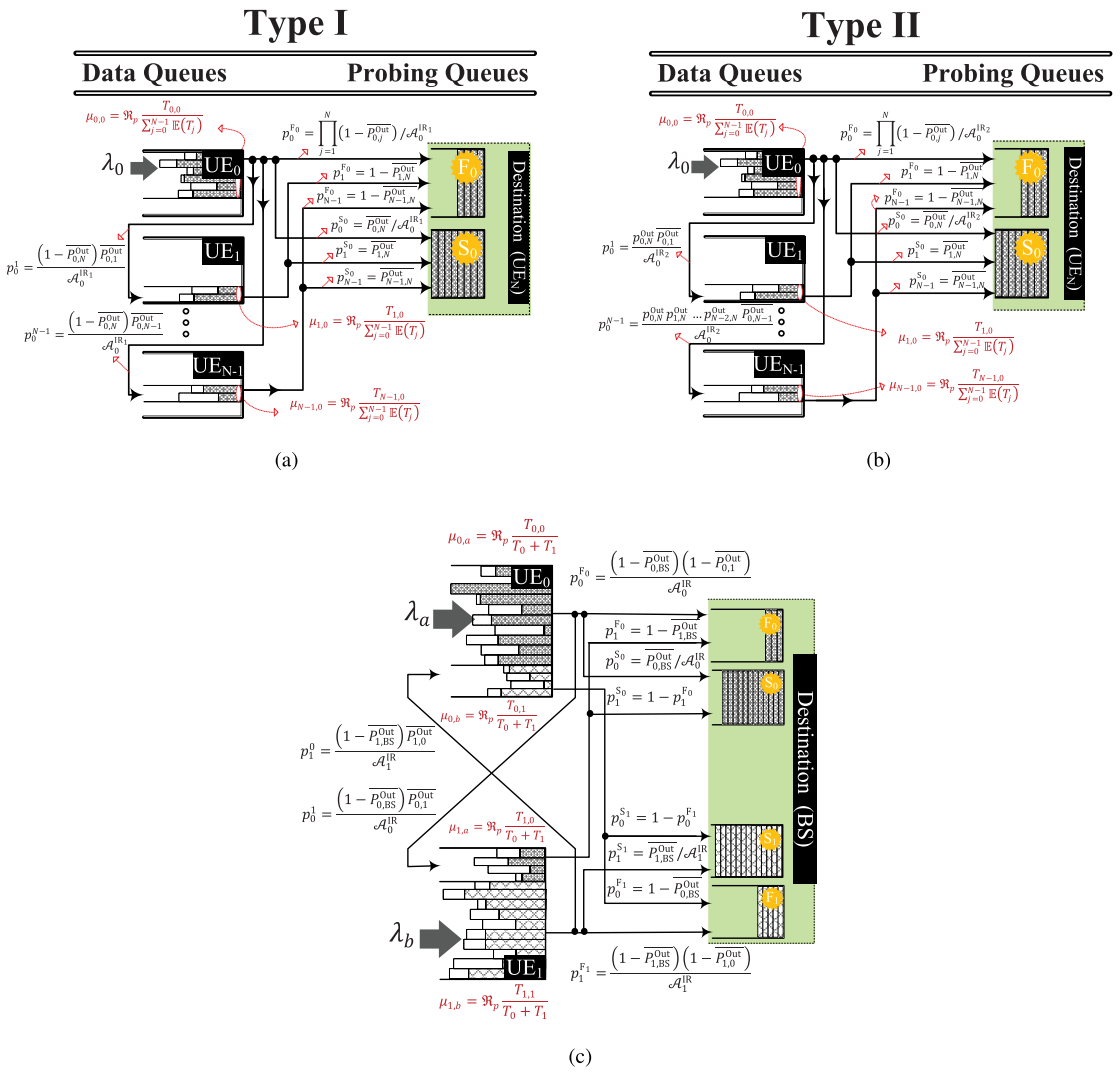

FIGURE 8. The complete and partial queue models for the configurations in Fig. 5a and Fig. 5b with two flavors of IR relaying, respectively. (a) BCMP model for the cooperating cluster in Fig. 5b with type-I IR relaying. (b) BCMP model for the cooperating cluster in Fig. 5b with type-II IR relaying. (c) BCMP model for the cooperating cluster in Fig. 5a with IR relaying (same for both flavors).

resource utilization is further enhanced by interrupting the forwarding phase immediately after $\mathfrak{D}(\mathrm{UE}_0)$ correctly recovers the packet. This saving comes at the price of requiring $\mathfrak{D}(\mathrm{UE}_0)$ to signal its success/failure in decoding the packet at the beginning of every sub-interval. This contributes to $N-1$ times larger feedback overhead compared to Type-I IR protocol.

Another issue about type-II IR protocol is that forwarding order of R-UEs becomes important. Intuitively, R-UEs with stronger source-relay and relay-destination channels are expected to be ranked upper in the forwarding list and it is $\mathfrak{D}(\mathrm{UE}_0)$ that will have to rank R-UEs $\in \mathfrak{N}(\mathrm{UE}_0)$. Figure 8 illustrates the BCMP queue model for both flavors of the IR protocol.

The logic behind the transition probabilities in this figure can be readily understood from our discussions in the previous subsections and the above protocol description. Next we tackle each type, separately.

a: TYPE I-IR

For type-I IR, a packet is forwarded by R-UE$_j$ if it is correctly received by this relay (occurring with probability $\overline{P^{\mathrm{Out}}_{0,j}}$) but not by the destination (occurring with probability $1-\overline{P^{\mathrm{Out}}_{0,N}}$), hence, $p_0^j = (1-\overline{P^{\mathrm{Out}}_{0,N}})\overline{P^{\mathrm{Out}}_{0,j}}/\mathcal{A}_0^{\mathrm{IR}_1}$. If none of the R-UEs $\in \mathfrak{N}(\mathrm{UE}_0)$ nor $\mathfrak{D}(\mathrm{UE}_0)$ could decode the packet, the latter fails and traverses through $\mathrm{UE}_0 \rightarrow \mathrm{F}_0$. This event occurs with probability $p_0^{\mathrm{F}_0} = \prod_{j=1}^{N}(1-\overline{P^{\mathrm{Out}}_{0,j}})/\mathcal{A}_0^{\mathrm{IR}_1}$. Another difference between the model for IR protocol and those of other protocols is that, in this case, there is a possibility that a packet directly traverses from $\mathrm{UE}_0 \rightarrow \mathrm{S}_0$ which occurs with $p_0^{\mathrm{S}_0} = \overline{P^{\mathrm{Out}}_{0,N}}/\mathcal{A}_0^{\mathrm{IR}_1}$. Similar to DF protocol, the high number of overlapping events obliges us to neglect some trivial cases through the normalization factor $\{\mathcal{A}_0^{\mathrm{IR}_1}|p_0^{\mathrm{F}_0} + p_0^{\mathrm{S}_0} + \sum_{j=1}^{N-1} p_0^j = 1\}$.

Service Rates: To derive the service rates, one should note that the interval lengths are no longer constant, as was the case in DF and AF protocols. Instead, intervals $\mathbf{T}_i$ are RVs for IR protocol. For instance, when UE_0 is the S-UE, we have

$$\mathbf{T}_0 = \begin{cases} 0, & q_0 = p_0^{F_0} \\ T_{0,0}, & q_1 = p_0^{S_0} \\ \sum_{j=0}^{N-1} T_{j,0}, & q_2 = 1 - p_0^{S_0} - p_0^{F_0}. \end{cases} \tag{19}$$

Hence, $\mathbb{E}\langle T_0\rangle = \sum_{j=0}^{N-1} T_{j,0} - \overline{p_{j-1,N}^{Out}} \sum_{j=1}^{N-1} T_{j,0}$ is to be used in the model. As a result,

$$\mu_{j,0} = \mathfrak{R}_p \frac{T_{j,0}}{\sum_{j=0}^{N-1} \mathbb{E}\langle \mathbf{T}_j\rangle}, \quad \forall j \in \{0,\ldots,N-1\}. \tag{20}$$

The fact that IR improves the spectrum efficiency and data rate is obvious from (19)-(20). The generic BCMP model for type-I IR is shown in Fig. 8a. Figure 8c is the complete model with $N = 2$ UEs.

b: TYPE II-IR

According to the brief description of type-II IR given earlier, a packet is forwarded by the relay R-UE_j if three conditions are met: **(i)** it is correctly received by this relay (occurring with probability $\overline{P_{0,j}^{Out}}$), **(ii)** yet not received by the destination (occurring with probability $1 - \overline{P_{0,N}^{Out}}$), **(iii)** and none of the preceding R-UEs could deliver the packet (occurring with probability $p_{0,N}^{Out} p_{1,N}^{Out} \ldots$). As such,

$$p_0^j = \frac{p_{0,N}^{Out} p_{1,N}^{Out} \cdots p_{j-1,N}^{Out} \overline{P_{0,j}^{Out}}}{\mathcal{A}_0^{IR_2}}, \tag{21}$$

where $\{\mathcal{A}_0^{IR_2} | p_0^{F_0} + p_0^{S_0} + \sum_{j=1}^{N-1} p_0^j = 1\}$. If none of the R-UEs $\in \mathfrak{N}(UE_0)$ nor the destination could decode the packet, then the packet is dropped ($UE_0 \rightarrow F_0$). The latter event happens with probability $p_0^{F_0} = \prod_{j=1}^{N} (1 - \overline{P_{0,j}^{Out}})/\mathcal{A}_0^{IR_2}$. Similar to type-I IR, the packet can be directly received by the destination without needing R-UEs to cooperate ($UE_0 \rightarrow S_0$). The latter event's likelihood is $p_0^{S_0} = \overline{P_{0,N}^{Out}}/\mathcal{A}_0^{IR_2}$.

Service Rates Likewise type-I, intervals $\mathbf{T}_i$ are RVs and are decomposed as

$$\mathbf{T}_0 = \begin{cases} 0, & q_0 = p_0^{F_0} \\ T_{0,0}, & q_1 = p_0^{S_0} \\ T_{0,0} + T_{0,1}, & q_2 = p_0^1 \\ \vdots \quad , & \quad \vdots \\ \sum_{j=0}^{N-1} T_{j,0}, & q_N = p_0^{N-1}, \end{cases} \tag{22}$$

whose probabilistic average is $\mathbb{E}\langle \mathbf{T}_0\rangle = \sum_{i=0}^{N-1} \sum_{j=0}^{i} T_{j,0}\, q_{i+1}$. Once all $\mathbb{E}\langle \mathbf{T}_i\rangle$ are derived, (20) is used to find service rates. Figure 8b incorporates all the details discussed for type-II IR. Note that, in this figure, it is implicitly assumed that UE_1 ranked highest and UE_{N-1} is ranked lowest among relays $\mathfrak{N}(UE_0)$. Because there is no difference between type-I and II IR when $N = 2$, Fig. 8c is representative of both cases.

B. SPACE-TIME CODED COOPERATION (STCC)

In this section, we introduce a relaying protocol, named space-time coded cooperation (STCC), based on the orthogonal STBC design. Similar to the repetition-based coding, STCC has both broadcast (BC) and relaying phases (see Fig. 5c). Yet there is a fundamental difference between them. First, despite repetition-based coding, where only one data packet of S-UE_0 is delivered within T_0, more than one data packets of S-UE_0 can be delivered using STCC within the same T_0. Second, in contrast to the repetition-based cooperation where relays $\mathfrak{N}(UE_0)$ use orthogonal sub-intervals to forward a packet, STCC empowers them to simultaneously transmit distinct packets in a block on the same frequency. This is possible using STBC codes [24], [25]. Although STBC was invented to provide transmit-diversity for MIMO channels, its applicability was expanded to cooperative scenarios where R-UEs are used as an S-UE's virtual antennas. Among different variants of STBC, those introduced based on the theory of orthogonal design [25] and quasi-orthogonal design [26] have gained popularity.

With this introduction, and as shown in Fig. 4b, STCC works in three phases to transmit $\mathtt{m}$ symbols $\mathcal{S}_1, \ldots, \mathcal{S}_{\mathtt{m}}$ of S-UE: **(i)** During the BC phase, symbols are broadcast one at a time taking $\mathtt{m}T_s$ sub-intervals to complete. **(ii)** Then, during coordination (CO) phase those R-UEs which decoded the whole block of $\mathtt{m}$ symbols (decoding set) declare their readiness for cooperation by sending short acknowledgments back to UE_0.[7] Once UE_0 knows $\mathfrak{N}(UE_0)$, it distributes appropriate space-time codes among them by specifying which symbol is to be forwarded in which of the $\mathtt{n} \approx |\mathfrak{N}(UE_0)|(1+\eta)$ time slots during the third phase. At the end of the CO phase, the transmission matrix $\mathbf{T}_{|\mathfrak{N}(UE_0)|\times \mathtt{n}}$ (whose element $[j,k]$ is the symbol that is to be transmitted by R-UE_j during the k^{th} time slot) is completely distributed. **(iii)** Finally, the space-time transmission (STT) phase is initiated as decided in the previous phase. It is important to note that without a feedback from R-UEs to S-UE during the CO phase, S-UE schedules all R-UEs to forward whereas some of them may not have entirely received the whole block. Consequently, those R-UEs may not be able to forward in the ensuing phase, a predicament that destroys the orthogonality of the columns of matrix $\mathbf{T}_{|\mathfrak{N}(UE_0)|\times \mathtt{n}}$ resulting in the destination not being able to recover some or all symbols. These three phases are illustrated in Fig. 4b. The structure of the BCMP model for STCC remains similar to the repetition-based coding with transition probabilities and service rates adapted to the protocol specifications.

Transition Prob.: A data packet from UE_0 is forwarded by R-UE_j, if it can be decoded by this relay, which happens

[7] This can be done using a random access protocol or prespecified ordering controlled by the UE_0 in the previous phase. Note that assigning S-UE as the coordinator is the most viable choice as the latter is within the hearing range of R-UEs in both BC and CO phases.

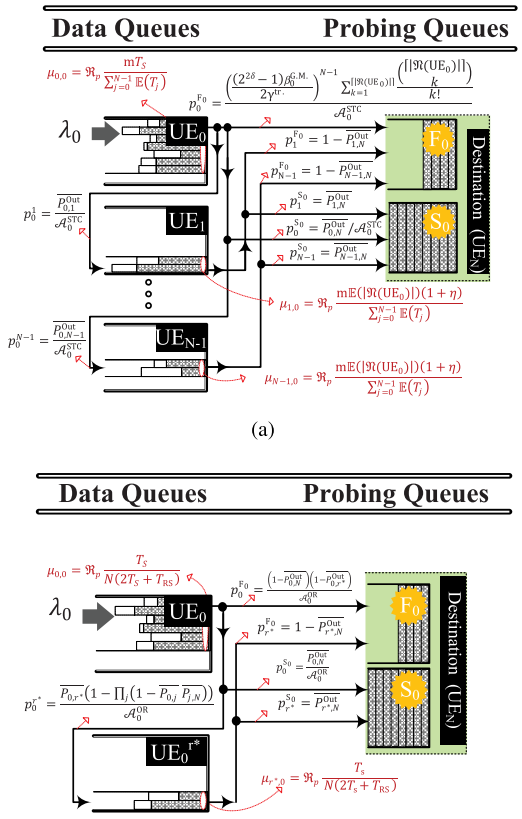

FIGURE 9. **The partial queue models for STCC and opportunistic relaying. (a) BCMP model for the cooperating cluster in Fig. 5c with space-time relaying. (b) BCMP model for the cooperating cluster in Fig. 5a with opportunistic-relaying.**

with probability $p_0^j = \overline{P_{0,j}^{\text{Out}}}/\mathcal{A}_0^{\text{STC}}$. The communications will be unsuccessful ($\text{UE}_0 \rightarrow \text{F}_0$) if $\mathfrak{D}(\text{UE}_0)$ cannot decode the packet at the end of the CO interval (see Fig. 9a). An asymptotic bound for the likelihood of this event is derived in [7, eqs. (22) and (25)] which after normalization by $\mathcal{A}_0^{\text{STC}}$ yields

$$p_0^{\text{F}_0} \lesssim \frac{\left(\frac{(2^{2\delta}-1)(N-1)\beta_0^{\text{G.M.}}}{2\gamma^{\text{tr.}}}\right)^{N-1} \sum_{k=1}^{\lceil \mathbb{E}\langle|\mathfrak{N}(\text{UE}_0)|\rangle\rceil}}{\mathcal{A}_0^{\text{STC}}} \times \frac{\binom{\lceil \mathbb{E}\langle|\mathfrak{N}(\text{UE}_0)|\rangle\rceil}{k}}{k!} \tag{23}$$

where $\{\mathcal{A}_0^{\text{STC}}|p_0^{\text{F}_0} + p_0^{\text{S}_0} + \sum_{j=1}^{N-1} p_0^j = 1\}$ and δ, $\gamma^{\text{tr.}}$, $\mathbb{E}\langle|\mathfrak{N}(\text{UE}_0)|\rangle$ are the spectral efficiency, transmit SNR, average cardinality of decoding set $\mathfrak{N}(\text{UE}_0)$, respectively. Operator $\lceil\cdot\rceil$ represents the ceiling function. Also, $\beta_i^{\text{G.M.}} = (\prod_{j=0}^{N-1} \check{\beta}_{i,j})^{1/(N-1)}$ is the geometric mean (G.M.) of the metrics $\check{\beta}_{i,j} = \max\left(|\beta_{i,j}|, |\beta_{j,N}|\right)$.

Service Rates: To calculate the service rates of the STCC model's queues in Fig. 9a, one should pay attention to the subtle point that, in STCC, $\mathtt{m}$ distinct symbols are communicated during T_i whereas, in repetition-based coding, only one symbol is transmitted per T_i. Therefore, **a factor $\mathtt{m}$ is to be multiplied in the service rate**. Also, given the fact that $\Pr(\text{UE}_j \in \mathfrak{N}(\text{UE}_0)) = \overline{P_{0,j}^{\text{Out}}}$, it is safe to say that

$$\mathbb{E}\langle|\mathfrak{N}(\text{UE}_0)|\rangle = \sum_{j=1}^{N-1} \overline{P_{0,j}^{\text{Out}}}. \tag{24}$$

Neglecting the length of CO phase ($\zeta = 0$), then $\mathbb{E}\langle\mathbf{T}_i\rangle = (\mathbb{E}\langle|\mathfrak{N}(\text{UE}_i)|\rangle\,(1+\eta) + \mathtt{m})\,T_s$ (see Fig. 4b) which is used in (25) to obtain the service rate of S-UE and R-UEs as shown in Fig. 9a.

$$\mu_{0,0} = \mathfrak{R}_p \frac{\mathtt{m}T_s}{\sum_{j=0}^{N-1} \mathbb{E}\langle\mathbf{T}_j\rangle}, \mu_{i,0} = \mathfrak{R}_p \frac{\mathtt{m}\mathbb{E}\langle|\mathfrak{N}(\text{UE}_0)|\rangle\,(1+\eta)}{\sum_{j=0}^{N-1} \mathbb{E}\langle\mathbf{T}_j\rangle} \tag{25}$$

Comparing the spectral efficiency between repetition-based and space-time codings, one can easily see that sending $\mathtt{m}$ symbols takes $(\mathbb{E}\langle|\mathfrak{N}(\text{UE}_i)|\rangle\,(1+\eta) + \mathtt{m})\,T_s$ with the latter and $\mathtt{m}NT_s$ with the former. Therefore, even in the worst case (in terms of rate) that $\mathbb{E}\langle|\mathfrak{N}(\text{UE}_i)|\rangle = N-1$, STT is more resource-efficient than its rival so long as $(1+\eta) < \mathtt{m}$. This is the case almost all the times.

C. OPPORTUNISTIC-RELATING (OR)

Synchronizing R-UEs in repetition-based and STCC is difficult and costy. In both approaches, the spectrum resource is wasted to improve diversity gain. In fact, while STCC is effective in low SNR regimes, it unnecessarily wastes resources at higher SNRs. Motivated by these facts, [21] proposes a distributed protocol that selects the best R-UE among a number of potential candidates. As illustrated in Fig. 4c, OR protocol works as in four stages: **(i)** S-UE (UE_0) sends a ready-to-send (RTS) control packet to its destination UE_N. This transmission is decoded by a set of R-UEs, namely $\mathfrak{N}^{\text{I}}(\text{UE}_0)$. These relays use this transmission to estimate their channel gains $|\beta_{0,j}|$ to the source. **(ii)** Next, destination UE_N acknowledges the receipt of the RTS packet by sending the clear-to-send (CTS) control packet. This packet is again decoded by another set of R-UEs, named $\mathfrak{N}^{\text{II}}(\text{UE}_0)$, and is used by this set to estimate the channel gains $|\beta_{N,j}|$. Since the system is time-duplex, reciprocity is a legitimate assumption, hence, $|\beta_{j,N}| = |\beta_{N,j}|$. The assumption is that the channel is slowly varying, thus, $|\beta_{i,j}|$ remains constant for the period of cooperation. **(iii)** Upon receiving CTS, each R-$\text{UE}_j \in \mathfrak{N}^{\text{I}}(\text{UE}_0) \cap \mathfrak{N}^{\text{II}}(\text{UE}_0)$ initiates a backoff timer whose length is proportional to $\mathfrak{T}_j = 1/\hat{\beta}_j$ where $\hat{\beta}_j = g(|\beta_{0,j}|, |\beta_{j,N}|) = \min\left(|\beta_{0,j}|, |\beta_{j,N}|\right)$ or, alternatively, $\hat{\beta}_j = g(|\beta_{0,j}|, |\beta_{j,N}|) = 2/(|\beta_{0,j}|^{-1} + |\beta_{j,N}|^{-1})$. Both formulas behave similarly, with the former being the limiting case of the latter. **(iv)** The relay whose counter expires first (call it R-UE_0^{r*}) has the highest $\hat{\beta}^* = \max \hat{\beta}_j$, hence, becomes the ultimate relay by sending an immediate short flag message that is decodable by all.

Transition Prob.: In modeling this protocol, a transmission is deemed unsuccessful ($UE_0 \rightarrow F_0$) if neither the destination nor any relays can decode the packet, hence, making the probability of such event

$$p_0^{F_0} = \frac{\left(1-\overline{P_{0,N}^{Out}}\right)\left(1-\overline{P_{0,r^*}^{Out}}\right)}{\mathcal{A}_0^{OR}}, \tag{26}$$

where $\{\mathcal{A}_0^{OR} | p_0^{F_0} + p_0^{S_0} + p_0^{r^*} = 1\}$. The other quantity of interest is the probability $p_0^{r^*}$ associated with transition $UE_0 \rightarrow R\text{-}UE_0^{r^*}$, which is derived as

$$p_0^{r^*} = \frac{\Pr\left(UE_0^{r^*} \text{ relays}\right)}{\mathcal{A}_0^{OR}} = \frac{\Pr\left(UE_0^{r^*} \text{ receives}\right)\Pr(UE_0^{r^*})}{\mathcal{A}_0^{OR}}$$

$$= \frac{\overline{P_{0,r^*}^{Out}}\,\Pr\left(\hat{\beta}^* > \gamma\right)}{\mathcal{A}_0^{OR}}. \tag{27}$$

Finally, the probabilities associated with transitions $UE_0^{r^*} \rightarrow F_0$ and $UE_0^{r^*} \rightarrow S_0$ are $p_{r^*}^{F_0} = 1 - \overline{P_{r^*,N}^{Out}}$ and $p_{r^*}^{S_0} = \overline{P_{r^*,N}^{Out}}$, respectively. In order to derive $\overline{P_{0,r^*}^{Out}}$ and $\overline{P_{r^*,N}^{Out}}$ as needed in (26)-(27), it should be noted that $UE_0^{r^*}$ is the best one in a pool, hence, its statistic is an ordered one. Thereby,

$$\begin{aligned}
&\Pr(\hat{\beta}^* > \gamma) \\
&= 1 - \Pr(\hat{\beta}^* < \gamma) \\
&= 1 - \Pr\left(\max_{j \in \mathfrak{N}^{I} \cap \mathfrak{N}^{II}} \min\left(|\beta_{0,j}| < \gamma, |\beta_{j,N}| < \gamma\right)\right) \\
&= 1 - \prod_{j \in \mathfrak{N}^{I} \cap \mathfrak{N}^{II}} \left(1 - \Pr\left(|\beta_{0,j}| > \gamma\right)\Pr\left(|\beta_{j,N}| > \gamma\right)\right) \\
&= 1 - \prod_{j \in \mathfrak{N}^{I} \cap \mathfrak{N}^{II}} \left(1 - \overline{P_{0,j}^{Out}}\,\overline{P_{j,N}^{Out}}\right)
\end{aligned} \tag{28}$$

Because $\overline{P_{0,r^*}^{Out}}$ and $\overline{P_{r^*,N}^{Out}}$ corresponding to $\Pr(\hat{\beta}^* < x)$ are also needed individually for full specification of the BCMP model in Fig. 9b, the last stage is to go the reverse way by calculating $g^{-1}(\hat{\beta}^*)$. In fact, this is as far as one can proceed analytically. Assuming that $UE_0^{r^*}$ is at the same distance from source and destination yields

$$\overline{P_{0,r^*}^{Out}} = \overline{P_{r^*,N}^{Out}} = \sqrt{1 - \Pr\left(\hat{\beta}^* < \gamma\right)}. \tag{29}$$

Following the same logic as before,

$$\mu_{0,0} = \mu_{r^*,0} = \frac{\mathfrak{R}_p T_s}{N\left(2T_s + T_{RS}\right)}. \tag{30}$$

where T_{RS} is the length of relay selection process and the fixed slot length T_s is the time it takes to send a packet.

V. PEFORMANCE METRICES AND CLOSED-FORMS

According to the BCMP models of the previous section, each class originates from one and only one S-UE whereas packets of different classes can be relayed by R-UEs according to a RPM that is different for each protocol. Also, in practice, packets transiting from a S-UE to a R-UE do not change type ($p_{i,r}^{j,s} \neq 0$ iff $r = s$), implying the traffic originating from UE_i remains isolated from that of UE_j. Subsequently, class and queue indices are indistinguishable and a packet from $\mathcal{C}_r$ also means a packet from $\mathcal{Q}_r$, thus, may alternatively be used in the following. Also, for the sake of clarity, when $N = R = 2$, traffic class is indexed by $\{a, b\}$.[8]

A. THROUGHPUT AND FAIRNESS

Let's momentarily assume the case where each S-UE is aided by one R-UE ($N = R = 2$). We define the **throughput** metric Γ as the fraction of packets accumulated in the auxiliary S_0 queue to the total number of packets accumulated in S_0 and F_0. Equivalently, one can define Γ_i (of UE_i) as

$$\Gamma_i = \frac{\Lambda_{S_i}}{\Lambda_{S_i} + \Lambda_{F_i}}, \tag{31}$$

where Λ_{S_i} and Λ_{F_i} are the gross arrival rates to queues S_i and F_i, respectively. These rates are obtained by writing the TBE as in (3). For the AF protocol with $N = 2$ in Fig. 6a, it can be easily verified that $\Lambda_{S_0} = p_1^{S_0}\Lambda_{1,a} + p_0^{S_0}\Lambda_{0,a}$, $\Lambda_{F_0} = p_1^{F_0}\Lambda_{1,a} + p_0^{F_0}\Lambda_{0,a}$, $\Lambda_{0,a} = \lambda_a$, and $\Lambda_{1,a} = p_0^1\Lambda_{0,a}$. Given that transition probabilities $p_0^{F_0}$, $p_0^{S_0}$, p_0^1, and p_1^0 are distinct for different protocols, the achievable throughput would vary. Table 2 provides the closed-form expressions of throughput for the cooperative protocols introduced in Section IV.

The notion of **fairness** is quantifiable by the contribution of a cooperating UE to the collective good of a cooperating cluster minus the contribution of that cluster to that same UE's good. This can mathematically be written as $Cr_i = \sum_{\substack{r=0 \\ r \neq i}}^{N-1} \Lambda_{i,r} - \sum_{\substack{j=0 \\ j \neq i}}^{N-1} \Lambda_{j,i}$. With this definition in mind, we define a collaboration as being fair when $Cr_i \propto Cr_j$, $\forall i \neq j$. Noting that $\Lambda_{i,j} = \Lambda_{j,i}, \forall i, j$, implies $Cr_i = Cr_j$, then functions $\kappa_{i,j}^{I} = \min(\Lambda_{i,j}, \Lambda_{j,i})$ and $\kappa_{i,j}^{II} = 2\Lambda_{i,j}\Lambda_{j,i}/(\Lambda_{i,j} + \Lambda_{j,i})$, that are maximized at $\Lambda_{i,j} = \Lambda_{j,i}$, can both be used to quantify pairwise fairness. Adopting $\kappa_{i,j}^{I}$ in this case study, Table 2 presents the closed-form expressions of the metric of fairness for different cooperative protocols ($N \geq 2$).

B. QUEUE LENGTH DISTRIBUTION

Let's start with the simple case of two queues, $N = 2$, and generalize the results to the arbitrary N. Assume that the arrival process to UE_i is independent of its state $\mathbf{S}_i = (\mathbf{n}_{i,a}, \mathbf{n}_{i,b})$, $\mathbf{n}_{i,r}$ being the number of $\mathcal{Q}_r$ packets in $\mathcal{Q}_i$. Therefore, based on (7), the multiplicative factor appearing in the product-form solutions (8) is simplified as $D\left(\mathcal{M}\left(\mathsf{S}\right)\right) =$

[8] Contrary to the notation system so far, hereinafter, a random-scalar quantity may be represented in regular-uppercase font.

TABLE 2. Throughput (Γ) and Fairness (κ) expressions for different cooperative protocols.

Protocol	Γ_0 (Throughput Metric)	$\kappa_{0,1}$ (Fairness Metric)
AF ($N=2$)	$\frac{\overline{P^{\mathrm{Out}}_{0,\mathrm{BS}}}+\overline{P^{\mathrm{Out}}_{0,1}}\,\overline{P^{\mathrm{Out}}_{1,\mathrm{BS}}}}{1+\overline{P^{\mathrm{Out}}_{0,\mathrm{BS}}}+\left(1-\overline{P^{\mathrm{Out}}_{0,\mathrm{BS}}}\right)\left(1-\overline{P^{\mathrm{Out}}_{0,1}}\right)}$	$\frac{1}{2-0.5\left(\overline{P^{\mathrm{Out}}_{0,1}}\left(1-\overline{P^{\mathrm{Out}}_{0,\mathrm{BS}}}\right)+\overline{P^{\mathrm{Out}}_{1,0}}\left(1-\overline{P^{\mathrm{Out}}_{1,\mathrm{BS}}}\right)\right)}$
DF ($N=2$)	$\frac{\overline{P^{\mathrm{Out}}_{0,\mathrm{BS}}}+\overline{P^{\mathrm{BER}}_{0,1}}\,\overline{P^{\mathrm{Out}}_{1,\mathrm{BS}}}}{\overline{P^{\mathrm{BER}}_{0,1}}+\overline{P^{\mathrm{Out}}_{0,\mathrm{BS}}}+\left(1-\overline{P^{\mathrm{Out}}_{0,\mathrm{BS}}}\right)\left(1-\overline{P^{\mathrm{Out}}_{0,1}}\right)}$	$\frac{1}{1+\frac{\left(1-\overline{P^{\mathrm{Out}}_{1,0}}\left(1-\overline{P^{\mathrm{Out}}_{1,\mathrm{BS}}}\right)\right)}{2\overline{P^{\mathrm{BER}}_{1,0}}}+\frac{\left(1-\overline{P^{\mathrm{Out}}_{0,1}}\left(1-\overline{P^{\mathrm{Out}}_{0,\mathrm{BS}}}\right)\right)}{2\overline{P^{\mathrm{BER}}_{0,1}}}}$
SR ($N=2$)	$\frac{\overline{P^{\mathrm{Out}}_{0,\mathrm{BS}}}\left(2-\overline{P^{\mathrm{Thresh.}}_{0,1}}\right)+\overline{P^{\mathrm{Out}}_{1,\mathrm{BS}}}\,\overline{P^{\mathrm{Thresh.}}_{0,1}}}{1+\overline{P^{\mathrm{Out}}_{0,\mathrm{BS}}}+\left(1-\overline{P^{\mathrm{Out}}_{0,\mathrm{BS}}}\right)\left(1-\overline{P^{\mathrm{Out}}_{0,1}}\right)}$	$\frac{1}{\frac{1-0.5\overline{P^{\mathrm{Out}}_{1,0}}\left(1-\overline{P^{\mathrm{Out}}_{1,\mathrm{BS}}}\right)}{\overline{P^{\mathrm{Thresh.}}_{1,0}}}+\frac{1-0.5\overline{P^{\mathrm{Out}}_{0,1}}\left(1-\overline{P^{\mathrm{Out}}_{0,\mathrm{BS}}}\right)}{\overline{P^{\mathrm{Thresh.}}_{0,1}}}}$
IR ($N=2$)	$\overline{P^{\mathrm{Out}}_{0,\mathrm{BS}}}+\overline{P^{\mathrm{Out}}_{1,\mathrm{BS}}}\,\overline{P^{\mathrm{Out}}_{0,1}}\left(1-\overline{P^{\mathrm{Out}}_{0,\mathrm{BS}}}\right)$	$\frac{1}{\frac{1}{2\left(1-\overline{P^{\mathrm{Out}}_{1,\mathrm{BS}}}\right)\overline{P^{\mathrm{Out}}_{1,0}}}+\frac{1}{2\left(1-\overline{P^{\mathrm{Out}}_{0,\mathrm{BS}}}\right)\overline{P^{\mathrm{Out}}_{0,1}}}}$
STT ($N\geq 2$)	$\frac{\sum_{i=1}^{N-1}\left(\overline{P^{\mathrm{Out}}_{0,i}}\,\overline{P^{\mathrm{Out}}_{i,N}}\right)+\overline{P^{\mathrm{Out}}_{0,N}}}{\sum_{i=1}^{N-1}\overline{P^{\mathrm{Out}}_{0,i}}+\overline{P^{\mathrm{Out}}_{0,N}}+\mathcal{A}_0^{\mathrm{STC}}p_0^{\mathrm{F}0}}$ ($p_0^{\mathrm{F}0}$ given by (23))	$\frac{2\overline{P^{\mathrm{Out}}_{i,j}}\,\overline{P^{\mathrm{Out}}_{j,i}}}{\mathcal{A}_j\overline{P^{\mathrm{Out}}_{i,j}}+\mathcal{A}_i\overline{P^{\mathrm{Out}}_{j,i}}}$ $\{\mathcal{A}_k \mid p_k^{\mathrm{F}}+p_k^{\mathrm{S}}+\sum_{\substack{j=0\\ j\neq k}}^{N-1}p_k^j=1\}$
OR ($N\geq 2$)	$\frac{\overline{P^{\mathrm{Out}}_{0,N}}+\overline{P^{\mathrm{Out}}_{0,r^*}}\,\overline{P^{\mathrm{Out}}_{r^*,N}}\left(1-\prod_{j\in\mathfrak{N}^{\mathrm{I}}\cap\mathfrak{N}^{\mathrm{II}}}\overline{P^{\mathrm{Out}}_{0,j}}\,\overline{P^{\mathrm{Out}}_{j,N}}\right)}{1+\overline{P^{\mathrm{Out}}_{0,r^*}}\left(\overline{P^{\mathrm{Out}}_{0,N}}-\prod_{j\in\mathfrak{N}^{\mathrm{I}}\cap\mathfrak{N}^{\mathrm{II}}}\overline{P^{\mathrm{Out}}_{0,j}}\,\overline{P^{\mathrm{Out}}_{j,N}}\right)}$	Not Applicable

$\lambda_a^{(n_{0,a}+n_{1,a})}\cdot\lambda_b^{(n_{0,b}+n_{1,b})}$. Assuming that the queuing discipline is PS, the product-form solution (8) becomes,

$$
\begin{aligned}
&f_{\mathbf{S}_1,\mathbf{S}_2}(\mathsf{S}_1,\mathsf{S}_2)\\
&\quad=\mathtt{k}_2\lambda_a^{(n_{0,a}+n_{1,a})}\lambda_b^{(n_{0,b}+n_{1,b})}\\
&\qquad\times\binom{n_{0,a}+n_{0,b}}{n_{0,a}}\binom{n_{1,a}+n_{1,b}}{n_{a,b}}\\
&\qquad\times\left(\frac{\mathsf{e}_{0,a}}{\mu_{0,a}}\right)^{n_{0,a}}\left(\frac{\mathsf{e}_{0,b}}{\mu_{0,b}}\right)^{n_{0,b}}\left(\frac{\mathsf{e}_{1,a}}{\mu_{1,a}}\right)^{n_{1,a}}\left(\frac{\mathsf{e}_{1,b}}{\mu_{1,b}}\right)^{n_{1,b}},
\end{aligned}
\tag{32}
$$

where $\binom{a}{b}$ are the binomial coefficients. Denoting $\rho_{i,r}=\Lambda_{i,r}/\mu_{i,r}$, where $\Lambda_{i,r}=\mathsf{e}_{i,r}\lambda_{i,r}$, and leveraging the fact that $\sum_{\mathsf{S}_1,\mathsf{S}_2} f_{\mathbf{S}_1,\mathbf{S}_2}(\mathsf{S}_1,\mathsf{S}_2)=1$ to find the constant $\mathtt{k}_2$, the joint PMF of the number of packets in both queues becomes,

$$
\begin{aligned}
f_{\mathbf{S}_1,\ldots,\mathbf{S}_{N(=2)}}&=\prod_{i=0}^{N-1(=1)}\prod_{r=0}^{R-1(=1)}P_{i,r}(n_{i,r}),\\
P_{i,r}(n_{i,r})&=\left(1-\Omega_{i,r}\right)\left(\Omega_{i,r}\right)^{n_{i,r}},\\
\Omega_{i,r}&=\frac{\rho_{i,r}}{1-\sum_{s\neq r}\rho_{i,s}}.
\end{aligned}
\tag{33}
$$

The joint PMF in (33), which remains valid for arbitrary $N=R$, demonstrates two important points: (a) the number of $\mathcal{C}_r$ packets in $\mathcal{Q}_i$ (i.e. RV $n_{i,r}$) is geometrically distributed and (b) the joint PMF in (33) is in a complete product-form. The latter indicates that all $n_{i,r}$ are independent of each other. In view of this finding, the mean and variance of $n_{i,r}$ are

$$
\begin{aligned}
\overline{n}_{i,r}&=\mathbb{E}\langle\mathbf{n}_{i,r}\rangle=\frac{\rho_{i,r}}{1-\sum_{s=1}^{R}\rho_{i,s}},\\
\sigma^2_{n_{i,r}}&=\mathbb{E}\langle\left(\mathbf{n}_{i,r}-\overline{n}_{i,r}\right)^2\rangle=\overline{n}_{i,r}+\overline{n}^2_{i,r}.
\end{aligned}
\tag{34}
$$

An important implication of (34) is that the tail of the distribution is not heavy. Also, the coefficient of variation $\mathrm{CoV}_{n_{i,r}}=\sqrt{1+1/\overline{n}_{i,r}}\to 1$ as $\overline{n}_{i,r}\to\infty$. These closed-forms can be of significant importance in understanding the behavior of most wireless networks that satisfy the aforementioned assumptions of the BCMP theory.

C. TOTAL LATENCY

In all the generic models introduced in the previous section, the total latency $\mathcal{D}_{(i)}$ that a packet of $\mathcal{C}_i$ originating from UE_i experiences till it leaves the network can be expressed by (35), where $\mathcal{D}_{i,k}, k\in\{0,\cdots,N-1\}, k\neq i$ is the sojourn delay[9] that packets of $\mathcal{Q}_i$ experience in $\mathcal{Q}_k$ and $\mathcal{D}_{i,i}$ ($\mathcal{D}_{i,-i}$) is the sojourn delay packet experiences at the source due to transmission (retransmission). Also, $q_{i,k}, q_{i,i}$, and $q_{i,-i}$ are the

[9]Sojourn time, that is a.k.a. response time, is the sum of queue waiting time and service time.

normalized probability associated with above cases. Note that $q_{i,-i}+\sum_{k=0}^{N-1} q_{i,k}=1$.

$$\mathcal{D}_{(i)}=\begin{cases}\mathcal{D}_{i,i}, & q_{i,i}=\dfrac{p_i^{\mathrm{S}_i}}{p_i^{\mathrm{S}_i}+\sum\limits_{j=0}^{N-1}p_i^j}\\ \mathcal{D}_{i,i}+\mathcal{D}_{i,-i}, & q_{i,-i}=\dfrac{p_i^{i}}{p_i^{\mathrm{S}_i}+\sum\limits_{j=0}^{N-1}p_i^j}\\ \mathcal{D}_{i,i}+\mathcal{D}_{i,k} & \\ \quad k\in\{0,\cdots,N-1\},k\neq i & , \\ & q_{i,k}=\dfrac{p_i^{k}}{p_i^{\mathrm{S}_i}+\sum\limits_{j=0}^{N-1}p_i^j}\end{cases} \tag{35}$$

Using (35), and exploiting the fact that $x \perp\!\!\!\perp y \rightarrow f_{x+y}=f_x * f_y$, the PDF of the sojourn delay that packets of $\mathcal{C}_i$ experience is given by

$$f_{\mathcal{D}_{(i)}}(t)=q_{i,i}f_{\mathcal{D}_{i,i}}(t)+q_{i,-i}\left(f_{\mathcal{D}_{i,i}}(t)*f_{\mathcal{D}_{i,-i}}(t)\right)+\sum_{\substack{k=0\\k\neq i}}^{N-1}q_{i,k}\left(f_{\mathcal{D}_{i,i}}(t)*f_{\mathcal{D}_{i,k}}(t)\right). \tag{36}$$

It is to be noted that $q_{i,-i}$ is non-zero only for the SR protocol as $p_i^i\neq 0$ only in the BCMP model of the latter. As it is evident in (36), in calculating the distribution of the total latency, the sojourn time distributions $f_{\mathcal{D}_{i,r}}(t)$ are needed. However, despite the simple form of the queue length distribution for PS queues, finding analytical solutions for the moments of $\mathcal{D}_{i,r}$ is much more difficult, even in the basic M/M/1 case. Until recently, there was little known about $f_{\mathcal{D}_{i,r}}(t)$.[10] Nevertheless, there are known results for the LT of $f_{\mathcal{D}_{i,r}}(t)$ in M/G/1/PS queue, denoted by $\mathcal{L}_{\mathcal{D}_{i,r}}(s)$. The latter can be written in terms of the LT of the sojourn time conditioned on a tagged packet needing service requirement $\mathrm{T}_{i,r}=t$ as

$$\mathcal{L}_{\mathcal{D}_{i,r}}(s)=\mathbb{E}\langle\exp(-s\mathcal{D}_{i,r})\rangle=\int_{t=0}^{\infty}\mathcal{L}_{\mathcal{D}_{i,r}}(s|\mathrm{T}_{i,r}=t)f_{\mathrm{T}_{i,r}}(t)\mathrm{d}t, \tag{37}$$

It was shown in [28] that the conditional LT $\mathcal{L}_{\mathcal{D}_{i,r}}(s|\mathrm{T}_{i,r}=t)$ for M/G/1/PS queue is given by

$$\mathcal{L}_{\mathcal{D}_{i,r}}(s|\mathrm{T}_{i,r}=t)=\frac{1-\Omega_{i,r}}{(1-\Omega_{i,r})\psi_1(s;t)+s\psi_2(s;t)},\quad(\Re(s)\geq 0) \tag{38}$$

whereby

$$\psi_1(0;t)=1 \text{ for } t\geq 0,$$
$$\psi_1(s;t)=\frac{1}{2\pi j}\int_{0-j\infty}^{0+j\infty}\frac{w-\Lambda_{i,r}\left(1-\mathcal{L}_{\mathrm{T}_{i,r}}(w)\right)}{w\left(w-s-\Lambda_{i,r}\left(1-\mathcal{L}_{\mathrm{T}_{i,r}}(w)\right)\right)}\mathrm{e}^{wt}\mathrm{d}w,$$
$$\psi_2(s;t)=\frac{1}{2\pi j}\int_{0-j\infty}^{0+j\infty}\frac{\Omega_{i,r}w-\Lambda_{i,r}\left(1-\mathcal{L}_{\mathrm{T}_{i,r}}(w)\right)}{w^2\left(w-s-\Lambda_{i,r}\left(1-\mathcal{L}_{\mathrm{T}_{i,r}}(w)\right)\right)}\mathrm{e}^{wt}\mathrm{d}w, \tag{39}$$

Note that $\psi_1(s;t)$ and $\psi_2(s;t)$ are the inverse LT of the terms under the integrals in (39). As a common practice, the method of partial fraction by decomposition should first be tried before resorting to contour integration in (39). In characterizing $\mathcal{L}_{\mathcal{D}_{i,r}}(s)$, equations (3) and (33) are used to obtain the gross arrival rates $\Lambda_{i,r}$ and equivalent utilization factors $\Omega_{i,r}$, respectively. We take the LT of $f_{\mathcal{D}_{(i)}}(t)$ in (36) in order to obtain $\mathcal{L}_{\mathcal{D}_{(i)}}(s)$ of the total latency for UE_i as follows

$$\mathcal{L}_{\mathcal{D}_{(i)}}(s)=\mathcal{L}_{\mathcal{D}_{i,i}}(s)\cdot\left(q_{i,i}+q_{i,-i}\mathcal{L}_{\mathcal{D}_{i,-i}}(s)+\sum_{\substack{k=0\\k\neq i}}^{N-1}q_{i,k}\mathcal{L}_{\mathcal{D}_{i,k}}(s)\right). \tag{40}$$

These results can be leveraged to obtain the moments of $\mathcal{D}_{(i)}$ provided that $\mathbb{E}\langle\mathcal{D}_{(r)}^n\rangle=(-1)^n\mathrm{d}^n\mathcal{L}_{\mathcal{D}_{(i)}}(s)/\mathrm{d}s^n|_{s=0}$. For instance, for all cooperative protocols except SR ($p_i^i=0\rightarrow q_{i,-i}=0$), the mean and variance of the total latency $\mathcal{D}_{(i)}$ is given by

$$\overline{\mathcal{D}}_{(i)}=\mathbb{E}\langle\mathcal{D}_{(i)}\rangle=q_{i,i}\overline{\mathcal{D}}_{i,i}+\sum_{\substack{k=0\\k\neq i}}^{N-1}q_{i,k}\left(\overline{\mathcal{D}}_{i,i}+\overline{\mathcal{D}}_{i,k}\right),$$
$$\mathbb{E}\langle{\mathcal{D}_{(i)}}^2\rangle=q_{i,i}\mathbb{E}\langle{\mathcal{D}_{i,i}}^2\rangle+\sum_{\substack{k=0\\k\neq i}}^{N-1}q_{i,k}\left(\mathbb{E}\langle\mathcal{D}_{i,i}^2\rangle+\mathbb{E}\langle\mathcal{D}_{i,k}^2\rangle+2\overline{\mathcal{D}}_{i,i}\overline{\mathcal{D}}_{i,k}\right),$$
$$\sigma^2_{\mathcal{D}_{(i)}}=\mathbb{E}\langle{\mathcal{D}_{(i)}}^2\rangle-\left(\overline{\mathcal{D}}_{(i)}\right)^2, \tag{41}$$

where $q_{i,k}$ are given by (35), and $\overline{\mathcal{D}}_{i,r}$, $\mathbb{E}\langle\mathcal{D}_{i,r}^2\rangle$ are the first two moments[11] of the sojourn delay. There are known closed-form results, due to [29], for these moments obviating the need to work out the complex integrations in (37)-(39):

$$\overline{\mathcal{D}}_{i,r}=\mathbb{E}\langle\mathcal{D}_{i,r}\rangle=\frac{1}{\mu_{i,r}\left(1-\Omega_{i,r}\right)},\quad\text{(M/G/1/PS)}$$
$$\sigma^2_{\mathcal{D}_{i,r}}=\mathbb{E}\langle\left(\mathcal{D}_{i,r}-\overline{\mathcal{D}}_{i,r}\right)^2\rangle=\frac{2\Omega_{i,r}}{\mu_{i,r}\left(1-\Omega_{i,r}\right)^2\left(2-\Omega_{i,r}\right)},\quad\text{(M/M/1/PS)} \tag{42}$$

VI. CASE STUDY AND NUMERICAL ANALYSIS

[10] This is due to the fact that the classic methods of analysis in queuing theory were futile to analyze PS queues. See [27] for a complete survey on the analysis of sojourn time in PS queues. In fact, as of now, there is still no known closed-form expression for $f_{\mathcal{D}_{i,r}}(t)$.

[11] The correctness of $\overline{\mathcal{D}}_{i,r}$ in (42) can also be verified by *Little*'s theorem, which states that $\overline{\mathcal{D}}_{i,r}=\overline{n}_{i,r}/\Lambda_{i,r}$, where $\overline{n}_{i,r}$ is given by (34).

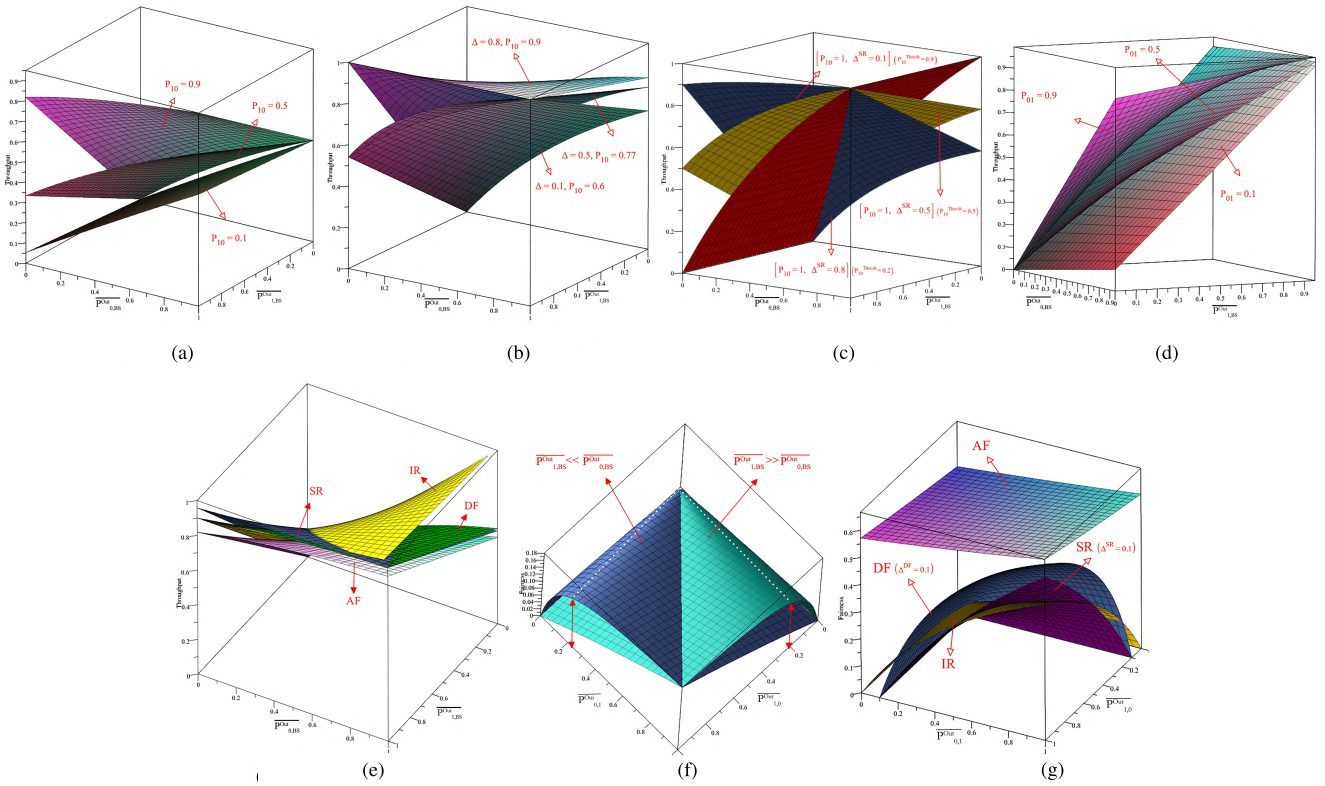

FIGURE 10. (a)-(e): 3D plots of Γ_0 (throughput of UE_0) vs the quality of direct wireless links $0 \rightarrow BS$ and $1 \rightarrow BS$ for different repetition-based cooperative protocols ($N = 2$) when the cooperation links are symmetrical ($\overline{P^{Out}_{1,0}} = \overline{P^{Out}_{0,1}}$). (f)-(g): 3D plots of fairness ($\kappa_{0,1}$) vs the quality of cooperative wireless links $0 \rightarrow 1$ and $1 \rightarrow 0$ for different protocols ($N = 2$). Note that each subfigure in (a)-(d) contains three curves corresponding to $\overline{P^{Out}_{0,1}} \in \{0.1, 0.5, 0.9\}$.

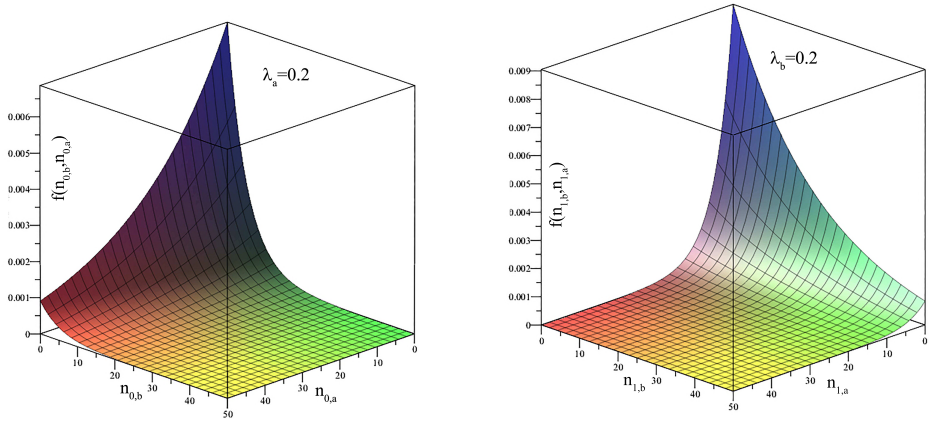

FIGURE 11. Marginal PMF of queue occupancy for the IR protocol with $N = 2$. (a) $f(n_{0,a}, n_{0,b})$ for $\rho_{0,a} = 0.8$, $\rho_{0,b} = 0.16$. (b) $f(n_{1,a}, n_{1,b})$ for $\rho_{1,a} = 0.3$ and $\rho_{1,b} = 0.65$.

A. FAIRNESS AND THROUGHPUT

For the simplicity of exposition, we will study the case in Fig. 5a where two UEs cooperate with each other to send traffic in uplink to the BS. We presume the simple case where the relaying channel has reciprocity, thus, $\overline{P^{Out}_{1,0}} = \overline{P^{Out}_{0,1}}$. Even though in all protocols, the throughput metric Γ_0 increases as the channel strength on links UE_0-BS, UE_1-BS, or UE_0-UE_1 improves, $\partial\Gamma_0/\partial\overline{P^{Out}_{0,BS}} > \partial\Gamma_0/\partial\overline{P^{Out}_{1,BS}}$ and $\partial\Gamma_0/\partial\overline{P^{Out}_{0,BS}} > \partial\Gamma_0/\partial\overline{P^{Out}_{0,1}}$, which signifies the relative importance of the channel quality on the direct link. It is observed that, regardless of the choice of parameters, $\partial^2\Gamma_0/\partial x^2 < 0$, $x \in \{\overline{P^{Out}_{0,BS}}, \overline{P^{Out}_{1,BS}}, \overline{P^{Out}_{0,1}}\}$. This observation implies that the better

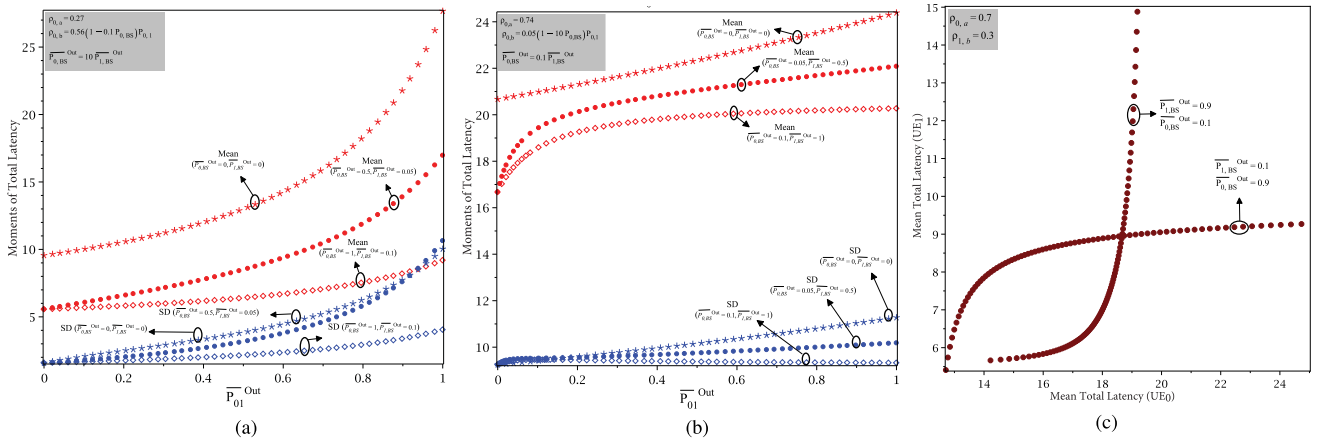

FIGURE 12. The mean and SD of cooperation total latency in a cooperative cluster with $N = 2$ exploiting IR protocol. (a) $[\overline{\mathcal{D}}_{(0)}, \sigma_{\mathcal{D}_{(0)}}]$, $(\overline{P^{\text{Out}}_{0,\text{BS}}} \gg \overline{P^{\text{Out}}_{1,\text{BS}}})$ (b) $[\overline{\mathcal{D}}_{(0)}, \sigma_{\mathcal{D}_{(0)}}]$, $(\overline{P^{\text{Out}}_{0,\text{BS}}} \ll \overline{P^{\text{Out}}_{1,\text{BS}}})$.

the channel quality is, the harder it becomes to ramp up throughput, a fact that confirms that cooperation achieves *higher diversity gain at lower SNR regimes*.

The situation in the SR protocol is dissimilar. Assuming that the relaying channel is ideal ($\overline{P^{\text{Out}}_{0,1}} = 1$) and setting $\Delta^{\text{SR}} \in \{0.1, 0.5, 0.8\}$, according to (18), relaying takes place with probability $\overline{P^{\text{Thresh.}}_{0,1}} \in \{0.9, 0.5, 0.2\}$, respectively. That being the case, and as evident from Fig. 10c, higher throughput is achieved by not forwarding if $\overline{P^{\text{Out}}_{0,\text{BS}}} > \overline{P^{\text{Out}}_{1,\text{BS}}}$ and vice versa.

Fig. 10e compares the four repetition-based cooperative protocols when $\overline{P^{\text{Out}}_{0,1}} = 0.9$. Note that the STT and OR protocols are not included here as they require $N > 2$ invalidating any comparison. According to this figure, all protocols perform similarly in high SNR regimes. Also, regardless of the operation regime IR exhibits the best performance. In fact, as $\overline{P^{\text{Out}}_{1,\text{BS}}} \to 0$, the throughput gain achieved through IR becomes more pronounced from the others. Of course, such improvement is gained at the cost of a limited feedback from the destination (BS) to relay(s) needed in this protocol.

Fig. 10g compares repetition-based protocols in terms of the metric of fairness ($\kappa_{0,1}$) vs the strength of the relaying channels. Here it is assumed that $\overline{P^{\text{Out}}_{1,\text{BS}}} = \overline{P^{\text{Out}}_{0,\text{BS}}} = 0.5$. The AF (which exhibited the lowest throughput) offers the *highest level of fairness* compared to other protocols. When the relaying channel is symmetric, $\text{DF} \succ_\kappa \text{SR} \succ_\kappa \text{IR}$.[12] On the contrary, when the relaying channel is asymmetric, $\text{IR} \succ_\kappa \text{SR}$ and as the level of asymmetry increases, $\text{IR} \succ_\kappa \text{DF} \succ_\kappa \text{SR}$. Due to the above symmetric assumption, the maximum value of fairness is attained on the bisecting line $\overline{P^{\text{Out}}_{1,0}} = \overline{P^{\text{Out}}_{0,1}}$. This is not the case when the direct channels are asymmetrical. This is illustrated in Fig. 10f where the fairness metric is plotted for $\overline{P^{\text{Out}}_{1,\text{BS}}} \gg \overline{P^{\text{Out}}_{0,\text{BS}}}$ and $\overline{P^{\text{Out}}_{1,\text{BS}}} \ll \overline{P^{\text{Out}}_{0,\text{BS}}}$. For example, in the latter case, fairness is higher when $\overline{P^{\text{Out}}_{1,0}} \gg \overline{P^{\text{Out}}_{0,1}}$. This implies that UE_0 shall relay UE_1's traffic if some level of fairness is expected illustrating the *inherent tradeoff between direct and relaying channel qualities*.

[12] $\text{x} \succ_\kappa \text{y}$ is the precedence operator signifying $\kappa(x) > \kappa(y)$.

B. QUEUE LENGTH DISTRIBUTION

Based on the findings of subsection V-B, the marginal PMFs of the occupancy level of $N = 2$ cooperating UEs (exploiting IR protocol) are depicted in Fig. 11. Obviously, even in this simplest scenario, the joint PMF $f(n_{0,a}, n_{1,a}, n_{0,b}, n_{1,b})$ cannot be visualized in the 3D plane. Therefore, the PMF of each UE is plotted separately. The arrival rates are $\lambda_a = \lambda_b = 0.2$. The service rates $\mu_{i,r}$, $i \in \{0, 1\}, r \in \{a, b\}$ are chosen such that $\rho_{0,a} = 0.8$, $\rho_{0,b} = 0.16$, $\rho_{1,a} = 0.3$, and $\rho_{1,b} = 0.65$. This was made possible through the appropriate choice of subintervals $T_{i,j}$ and base transmission rate $\mathfrak{R}_p$ as shown in BCMP models of Fig. 8. Recall that the BCMP model allows for queues with generic service time distribution for PS queues, yet due to the nice structure behavior of BCMP models, only the average service rates participate in closed-forms of queue length PMF.

C. TOTAL LATENCY

Following the derivations in subsection V-C, Fig. 12 presents the numerical results of the first two moments of the total latency in a cluster with $N = 2$ cooperating UEs when the IR protocol is exploited. Two cases are considered: **(i)** $\overline{P^{\text{Out}}_{0,\text{BS}}} \gg \overline{P^{\text{Out}}_{1,\text{BS}}}$ (Fig. 12a) **(ii)** and $\overline{P^{\text{Out}}_{0,\text{BS}}} \ll \overline{P^{\text{Out}}_{1,\text{BS}}}$ (Fig. 12b). Assuming reciprocity on the relaying channel ($\overline{P^{\text{Out}}_{0,1}} = \overline{P^{\text{Out}}_{1,0}}$), the mean $\overline{\mathcal{D}}_{(0)}$ and standard deviation (SD) $\sigma_{\mathcal{D}_{(0)}}$ of UE_0's total latency are plotted in each case. The first important observation is that $\sigma_{\mathcal{D}_{(0)}} < \overline{\mathcal{D}}_{(0)}$ (implying $\text{CoV}_{\mathcal{D}_{(0)}} < 1$) no matter how busy each queue is and how poor/strong the channels are. This guarantees that, by cooperating with other UEs, the variability remains bounded, thus, QoS will remain predictable. In both scenarios, as $\overline{P^{\text{Out}}_{0,\text{BS}}}$ increases (channel UE_0-BS improves), delay decreases since there is no need for the packets to be

relayed and undergo extra waiting times in the cooperating relay. The fact that $\partial\overline{\mathcal{D}}_{(0)}/\partial\overline{P}^{\text{Out}}_{1,0} > 0$ shall not seem counterintuitive. That is because, according to (35), the metric of delay is only defined for successfully delivered packets. Thereby, as $\overline{P}^{\text{Out}}_{1,0}$ increases, more packets can be successfully transmitted, thus throughput boosts (Fig. 10d), the queues become more occupied, and subsequently, larger delay per packet is experienced. Finally, Fig. 12c illustrates the mean total latency of UE_0 vs UE_1 for the aforementioned two extreme cases where $\overline{P}^{\text{Out}}_{0,\text{BS}} \gg \overline{P}^{\text{Out}}_{1,\text{BS}}$ and $\overline{P}^{\text{Out}}_{0,\text{BS}} \ll \overline{P}^{\text{Out}}_{1,\text{BS}}$. Each point $(\overline{\mathcal{D}}_{(0)}, \overline{\mathcal{D}}_{(1)})$ in this figure corresponds to a given value of $0 < p^1_0 < 1$. This plot demonstrates the scale by which the channel strength disparity reflects into the nonlinear behavior of total latency experienced by terminals.

VII. CONCLUSION

Cooperation is a fundamental concept in materializing the promises of 5G cellular networks. By leveraging the spatial diversity of the wireless channel, cooperation can significantly improve network coverage, transmission rate, and power consumption. Despite the abundance of studies on the capacity analysis of cooperative protocols, a higher-level analysis of cooperation is missing. This research was tasked with the mission to establish such framework. The theory of BCMP networks, from the discipline of queueing theory, was leveraged to build packet-level models that help us understand how the service quality is influenced in the presence of cooperation. The BCMP models were introduced for several renowned cooperative protocols, such as AF, DF, SR, IR, and OR. A rich set of packet-level metrics were defined, including the metric of fairness, throughput, total latency, as well as buffer length. Appropriate closed-form expressions were extracted for these metrics from their corresponding models, which were used to compare the performance of the above-mentioned protocols. The hope is that the proposed modeling enhances our understanding of cooperation in wireless networks in a more tangible manner and accelerates the adoption of this technology into the future generation of cellular networks. As the future research direction, the proposed queueing analysis can also be used to model energy-harvesting, full-duplex, and green communications networks.

REFERENCES

[1] J. Xu, L. Duan, and R. Zhang, "Cost-aware green cellular networks with energy and communication cooperation," *IEEE Commun. Mag.*, vol. 53, no. 5, pp. 257–263, May 2015.

[2] E. Hossain and M. Hasan, "5G cellular: Key enabling technologies and research challenges," *IEEE Instrum. Meas. Mag.*, vol. 18, no. 3, pp. 11–21, Jun. 2015.

[3] X. Tao, X. Xu, and Q. Cui, "An overview of cooperative communications," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 65–71, Jun. 2012.

[4] F. Baskett, K. M. Chandy, R. Muntz, and F. Palacios, "Open, closed, and mixed networks of queues with different classes of customers," *J. ACM*, vol. 22, no. 2, pp. 248–260, Apr. 1975.

[5] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[6] A. Nosratinia, T. E. Hunter, and A. Hedayat, "Cooperative communication in wireless networks," *IEEE Commun. Mag.*, vol. 42, no. 10, pp. 74–80, Oct. 2004.

[7] J. N. Laneman and G. W. Wornell, "Distributed space-time coded protocols for exploiting cooperative diversity in wireless networks," *IEEE. Trans. Inf. Theory*, vol. 49, no. 10, pp. 2415–2425, Oct. 2003.

[8] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity, Part I: System description," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1927–1938, Nov. 2003.

[9] A. Tukmanov, Z. Ding, S. Boussakta, and A. Jamalipour, "On the impact of network geometric models on multicell cooperative communication systems," *IEEE Wireless Commun.*, vol. 20, no. 1, pp. 75–81, Feb. 2013.

[10] D. Niyato, P. Wang, E. Hossain, W. Saad, and Z. Han, "Game theoretic modeling of cooperation among service providers in mobile cloud computing environments," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Paris, France, Apr. 2012, pp. 3128–3133.

[11] B. Cao, J. W. Mark, Q. Zhang, R. Lu, X. Lin, and X. S. Shen, "On optimal communication strategies for cooperative cognitive radio networking," in *Proc. IEEE INFOCOM*, Turin, Italy, Apr. 2013, pp. 1726–1734.

[12] M. Khabazian and S. Aïssa, "Modeling and performance analysis of cooperative communications in cognitive radio networks," in *Proc. IEEE 22nd Int. Symp. (PIMRC)*, Toronto, ON, Canada, Sep. 2011, pp. 598–603.

[13] H. Tran, H.-J. Zepernick, H. Phan, and L. Sibomana, "Performance analysis of a cognitive radio network with a buffered relay," *IEEE Trans. Veh. Technol.*, vol. 64, no. 2, pp. 566–579, Feb. 2015.

[14] O. Simeone, Y. Bar-Ness, and U. Spagnolini, "Stable throughput of cognitive radios with and without relaying capability," *IEEE Trans. Commun.*, vol. 55, no. 12, pp. 2351–2360, Dec. 2007.

[15] I. Suliman and J. Lehtomaki, "Queueing analysis of opportunistic access in cognitive radios," in *Proc. Int. Workshop Cogn. Radio Adv. Spectr. Manage.*, Aalborg, Denmark, May 2009, pp. 153–157.

[16] C. Zhang, X. Wang, and J. Li, "Cooperative cognitive radio with priority queueing analysis," in *Proc. IEEE Int. Conf. Commun. (ICC)*, New York, NY, USA, Jun. 2009, pp. 1–5.

[17] N. Tadayon and S. Aïssa, "Modeling and analysis framework for multi-interface multi-channel cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 935–947, Feb. 2015.

[18] S. Wang, J. Zhang, and L. Tong, "Delay analysis for cognitive radio networks with random access: A fluid queue view," in *Proc. IEEE INFOCOM*, San Diego, CA, USA, Apr. 2010, pp. 1–9.

[19] S. Laourine, S. Chen, and L. Tong, "Queuing analysis in multichannel cognitive spectrum access: A large deviation approach," in *Proc. IEEE INFOCOM*, San Diego, CA, USA, Mar. 2010, pp. 1–9.

[20] N. Tadayon and S. Aïssa, "Modeling and analysis of cognitive radio based IEEE 802.22 wireless regional area networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4363–4375, Sep. 2013.

[21] A. Bletsas, A. Khisti, D. P. Reed, and A. Lippman, "A simple cooperative diversity method based on network path selection," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 659–672, Mar. 2006.

[22] J. Kaufman, "Blocking in a shared resource environment," *IEEE Trans. Commun.*, vol. COM-29, no. 10, pp. 1474–1481, Oct. 1981.

[23] A. Goldsmith, *Wireless Communications*, 1st ed. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[24] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block coding for wireless communications: Performance results," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 3, pp. 451–460, Mar. 1999.

[25] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1456–1467, Jul. 1999.

[26] H. Jafarkhani, "A quasi-orthogonal space-time block code," *IEEE Trans. Commun.*, vol. 49, no. 1, pp. 1–4, Jan. 2001.

[27] S. F. Yashkov, "Mathematical problems in the theory of shared-processor systems," *J. Sov. Math.*, vol. 58, no. 2, pp. 147–191, Jan. 1992.

[28] T. J. Ott, "The sojourn-time distribution in the M/G/1 queue with processor sharing ," *J. Appl. Probab.*, vol. 21, no. 2, pp. 360–378, Jun. 1984.

[29] E. G. Coffman, Jr., R. R. Muntz, and H. Trotter, "Waiting time distributions for processor-sharing systems," *J. ACM*, vol. 17, no. 1, pp. 123–130, Jan. 1970.

NAVID TADAYON (S'15–M'18) received the M.Sc. degree from the University of Massachusetts, Amherst, MA, USA, in 2011, and the Ph.D. degree from the Institut National de la Recherche Scientifique, University of Quebec, Montréal, QC, Canada. From 2016 to 2017, he has been a Post-Doctoral Associate with the Laboratory of Information and Decision Systems, Massachusetts Institute of Technology (MIT). He is currently a Post-Doctoral Fellow at the University of Toronto (UofT). His research interests include modeling and analysis of wireless communication networks as well as designing efficient mechanisms and algorithms, with a particular focus on 5G enabling technologies, such as cognitive radio networks, HetNets, and D2D. His recent research work at MIT and UofT has been focused on designing and building a pervasive and precise indoor localization system using MIMO-OFDM signals.

Dr. Tadayon has been the holder of several prestigious Canadian awards, including the National NSERC Post-Doctoral Fellowship and the FQRNT PhD Merit Scholarship. The outcomes of his investigations have been published as a book and dozens of papers in distinguished transaction journals and flagship conferences.

Georges Kaddoum (M'11) received the bachelor's degree in electrical engineering from the Ecole Nationale Supérieure de Techniques Avancées (ENSTA Bretagne), Brest, France, the M.S. degree in telecommunications and signal processing (circuits, systems, and signal processing) from the Université de Bretagne Occidentale and Telecom Bretagne (ENSTB), Brest, in 2005, and the Ph.D. degree (Hons.) in signal processing and telecommunications from the National Institute of Applied Sciences, University of Toulouse, Toulouse, France, in 2009. Since 2013, he has been an Assistant Professor of electrical engineering with the Ecole de Technologie Supérieure (ÉTS), University of Quebec, Montréal, QC, Canada. In 2014, he was awarded the ÉTS Research Chair in physical-layer security for wireless networks. Since 2010, he has been a Scientific Consultant of space and wireless telecommunications for several U.S. and Canadian companies. He has published over 130 journal and conference papers and holds two pending patents. His current research interests include mobile communication systems, modulations, security, and space communications and navigation. He received the Best Paper Awards at the 2014 IEEE International Conference on Wireless and Mobile Computing, Networking, Communications, with three co-authors, and at the 2017 IEEE International Symposium on Personal Indoor and Mobile Radio Communications, with four co-authors. He was honored the 2015 and 2017 IEEE TRANSACTIONS ON COMMUNICATIONS Exemplary Reviewer Award. He is currently serving as an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and the IEEE COMMUNICATIONS LETTERS.