On the Benefits of Edge Caching for MIMO Interference Alignment

Matha Deghel[⋆,◇], Ejder Baştuğ[◇], Mohamad Assaad[⋆] and Merouane Debbah[◇,†]

[⋆]Laboratoire de Signaux et Systèmes (L2S, UMR8506) CentraleSupélec-CNRS-Université Paris-Sud, Gif-sur-Yvette, France
[◇]Large Networks and Systems Group (LANEAS), CentraleSupélec, Gif-sur-Yvette, France
[†]Mathematical and Algorithmic Sciences Lab, Huawei France R&D, Paris, France
{matha.deghel, ejder.bastug, mohamad.assaad, merouane.debbah}@centralesupelec.fr

***Abstract*—In this contribution, we jointly investigate the benefits of caching and interference alignment (IA) in multiple-input multiple-output (MIMO) interference channel under limited backhaul capacity. In particular, total average transmission rate is derived as a function of various system parameters such as backhaul link capacity, cache size, number of active transmitter-receiver pairs as well as the quantization bits for channel state information (CSI). Given the fact that base stations are equipped both with caching and IA capabilities and have knowledge of content popularity profile, we then characterize an *operational regime* where the caching is beneficial. Subsequently, we find the optimal number of transmitter-receiver pairs that maximizes the total average transmission rate. When the popularity profile of requested contents falls into the operational regime, it turns out that caching substantially improves the throughput as it mitigates the backhaul usage and allows IA methods to take benefit of such limited backhaul.**

***Index Terms*—edge caching, interference alignment, limited backhaul, wireless networks, 5G cellular networks**

I. Introduction

The current mobile cellular networks are evolving towards $5G$ wireless networks, aiming to sustain the huge rise of connected devices and data-hungry application of mobile users. Among the possible solutions [1], proactively caching users' contents at the network edge is shown to achieve significant gains in terms of users' satisfaction and offloading gains [2]. Specifically, the idea of caching is to smartly move the users' contents close to mobile users, yielding less access delays to the contents and reducing the backhaul usage. In the same context, one of the key issue in wireless communication systems is the interference which is caused by the large number of simultaneous transmissions on the same channel, resulting into severe performance degradations unless treated properly. In this regard, interference alignment (IA) is introduced as an efficient interference management method and is shown to result in higher throughputs compared to conventional interference-agnostic methods.

In the context of cellular networks, caching was recently studied by different research groups, both in terms of gains and approximation algorithms [3]–[12]. On the other hand, IA was initially introduced in [13], and is shown to achieve maximum multiplexing gain in multiple-input multiple-output (MIMO) channels [14] under the assumption that all the transmitters have perfect global channel state information (CSI). In frequency-division duplex (FDD) systems, the imperfect case with CSI quantization process for single-antenna receivers [15], and multiple-antenna receivers [16], [17] are studied, showing that the degree-of-freedom (DoF) can be achieved at high signal-to-noise ratio (SNR) regime by using a specific quantization scheme with optimal number of feedback bits. The IA methods that exploit channel reciprocity in time-division duplex (TDD) systems are studied (see [18]–[21] for instance), assuming that the CSI acquisition cost is independent of the transmission rate and is linear in the number of probed receivers. In fact, most of aforementioned IA methods rely on CSI exchange over the backhaul links and do not consider the implications of data traffic on the limited backhaul links and exchange process. From these observations, one can bring caching into the scenario as a way of creating opportunities for CSI exchange over the backhaul. In other words, IA methods could have higher throughputs as the amount of data traffic over the backhaul is substantially reduced, since this reduction results in a saved capacity which can be used for the CSI sharing process.

Based on the motivations above, the main contribution of this work is to jointly analyze the benefits of caching and IA methods under the limited backhaul. In particular, given the fact that users' content requests follow a certain popularity profile (i.e., few contents might be highly popular than the rest or all might have similar popularities), we aim to find an operational regime where the caching is beneficial to IA methods in terms of throughput. To show this, we first derive the expressions for average throughput, then characterize this regime based on the shape of content popularity profile. Finally, we maximize the total average throughput as a key metric of interest. In a similar vein, the work in [11] has jointly studied the caching and power control problem for opportunistic cooperative MIMO. Therein, closed form expressions for power control are derived based on approximated Bellman equation and convex stochastic caching problem is solved via a stochastic subgradient algorithm. The proposed scheme is

This research has been supported by the ERC Starting Grant 305123 MORE (Advanced Mathematical Tools for Complex Network Engineering) and the project BESTCOM.

shown to be asymptotically optimal in the high SNR regime. Another joint solution for cooperative MIMO was introduced in [12], where both caching control and the optimal MIMO precoder in transmit power minimization are investigated.

The rest of this paper is structured as follows. Our system model is given in Section II, including the details of the MIMO interference channel model, IA and caching capabilities at the transmitters with limited backhaul. In Section III, the expressions for average transmission rate are derived as the main performance metrics. Based on these expressions, an operational caching regime that meets certain quality-of-service (QoS) criteria is provided by relying on content popularity profile. Then, an optimization problem for maximizing the average transmission rate is formulated, where the number of active transmitter-receiver pairs is optimized subject to the backhaul capacity constraints. Section IV is dedicated to numerical results and relevant discussions. We finally conclude and draw our future directions in Section V.

Notation: Boldface uppercase symbols (i.e., $\mathbf{B}$) represent matrices whereas lowercases (i.e., $\mathbf{b}$) are used for vectors. The symbol $\mathbf{I}$ denotes square identity matrix. $(.)^*$ denotes the conjugate transpose. $|.|$ indicates the absolute value and $||.||$ is used for the norm of second degree. Lastly, $\mathcal{CN}(\mathbf{b}, \mathbf{B})$ corresponds to a complex Gaussian random vector with mean $\mathbf{b}$ and covariance matrix $\mathbf{B}$.

II. System Model

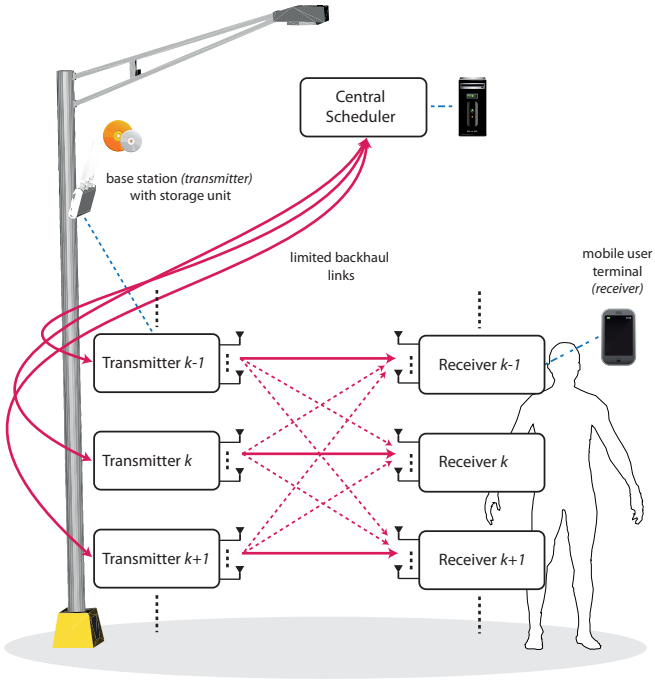

Figure 1: A sketch of L-User MIMO interference network.

We consider a MIMO interference channel with L transmitter-receiver pairs, as illustrated in Fig. 1. For simplicity, we assume a homogeneous network where all transmitters (base stations) are equipped with N_t antennas and all receivers (users) with N_r antennas. The number of independent data streams from transmitter k to its paired receiver k is denoted by d_k, with $d_k \leq \min(N_\mathrm{t}, N_\mathrm{r})$.

Given this MIMO interference channel model, the received signal at user k can be written as

$$\mathbf{y}_k = \sum_{i=1}^{L} \sqrt{\frac{\zeta_{ki} P}{d_i}} \mathbf{H}_{ki} \sum_{j=1}^{d_i} \mathbf{v}_i^j x_i^j + \mathbf{z}_k \tag{1}$$

where $\mathbf{y}_k$ is the $N_\mathrm{r} \times 1$ received signal vector, $\mathbf{H}_{ki}$ is the $N_\mathrm{r} \times N_\mathrm{t}$ channel matrix between transmitter i and receiver k with i.i.d. $\mathcal{CN}(0,1)$ elements, ζ_{ki} represents the path loss of channel $\mathbf{H}_{ki}$, P is the total power at each transmitter equally allocated among its streams, x_i^j denotes the j-th data stream from transmitter i, $\mathbf{v}_i^j \in \mathbb{C}^{N_\mathrm{t} \times 1}$ is the corresponding precoding vector of unit norm and $\mathbf{z}_k$ is a vector of i.i.d. complex Gaussian noise with covariance matrix $\sigma^2 \mathbf{I}_{N_\mathrm{r}}$. We denote by α_{ki} the fraction $\frac{\zeta_{ki} P}{d_i}$, for all k, i in $\{1, ..., L\}$.

A. Interference Alignment

IA is a linear precoding technique which can be adopted for the MIMO interference channel. While this technique is commonly used with multiple receiver design, for the sake of simplicity we restrict ourselves to a per-stream zero-forcing receiver. Specifically, let receiver k use the *combiner* vector $\mathbf{u}_k^m \in \mathbb{C}^{N_\mathrm{r} \times 1}$ of unit norm to detect the m-th stream from transmitter k, such as

$$\hat{x}_k^m = (\mathbf{u}_k^m)^* \mathbf{y}_k$$
$$= \overbrace{\sqrt{\alpha_{kk}} (\mathbf{u}_k^m)^* \mathbf{H}_{kk} \mathbf{v}_k^m x_k^m}^{\text{desired signal}} + \overbrace{\sqrt{\alpha_{kk}} \sum_{\substack{j=1 \\ j \neq m}}^{d_k} (\mathbf{u}_k^m)^* \mathbf{H}_{kk} \mathbf{v}_k^j x_k^j}^{\text{inter-stream interference (ISI)}}$$
$$+ \overbrace{\sum_{\substack{i=1 \\ i \neq k}}^{L} \sqrt{\alpha_{ki}} \sum_{j=1}^{d_i} (\mathbf{u}_k^m)^* \mathbf{H}_{ki} \mathbf{v}_i^j x_i^j}^{\text{inter-user interference (IUI)}} + \overbrace{(\mathbf{u}_k^m)^* \mathbf{z}_k}^{\text{noise}}. \tag{2}$$

As observed from (2), two sources of interference affect the detection of the stream at the receiver, namely i) the ISI and ii) the IUI. The IA technique is used to manage this problem by designing the set of precoder and combiner vectors such that

$$(\mathbf{u}_k^m)^* \mathbf{H}_{ki} \mathbf{v}_i^j = 0, \qquad \forall (k,m) \neq (i,j). \tag{3}$$

The *perfect interference alignment* is achieved if the above conditions hold. In other words, supposing that perfect global CSI is available at all the transmitters and each receiver consequently obtains a perfect version of the combiner vector designed at its corresponding transmitter, IUI and ISI can be canceled completely at the receivers. It turns out that obtaining the perfect global CSI at the transmitters is not a straightforward task in practice due to the limited backhaul. The CSI sharing mechanism over the limited backhaul is detailed in the following.

B. CSIT Sharing Over Limited Capacity Backhaul Links

As alluded earlier, global CSI is required at each transmitting node in order to design the IA vectors that satisfy (3). As shown in Fig. 1, we suppose that all the transmitters are connected to a central node via their limited backhaul links, which serves as: (i) a way for connecting transmitters to each other and (ii) a mean to link the system to the Internet for data transfer. We assume a TDD transmission strategy where the users send their training sequences, allowing each transmitter to estimate its *local* CSI, meaning that the i-th transmitter estimates perfectly the channels $\mathbf{H}_{ki}$, $k = 1, ..., L$. However, the local CSI (excluding the direct links) of other transmitters are obtained via backhaul links of limited capacity.

In this contribution, we suppose that the backhaul is error-free and has a fixed capacity of C. The capacity of each link from a transmitter to the central node is then given by $C_k = \frac{C}{L}$, as a function of the number of active transmitter-receiver pairs. Note that k refers to pair k, where $k = 1, ..., L$. Denoting C_{c} as the capacity reserved for CSI sharing and C_{d} as the part dedicated to data transfer, the capacity of each link can be also written as $C_k = C_{k\mathrm{c}} + C_{k\mathrm{d}}$. We assume that $C_{k\mathrm{c}} = \frac{C_{\mathrm{c}}}{L}$ and $C_{k\mathrm{d}} = \frac{C_{\mathrm{d}}}{L}$. In such limited backhaul conditions, a codebook-based quantization technique needs to be adopted to reduce the huge amount of information exchange used for CSI sharing, which we detail as follows. Let $\mathbf{h}_{ki}$ denote the vectorization of the channel matrix $\mathbf{H}_{ki}$. Then, for all $k \neq i$, transmitter i selects the index n_o which corresponds to the optimal codeword in a predetermined codebook $\mathcal{CB} = \left[\hat{\mathbf{h}}_{ki}^{1}, ..., \hat{\mathbf{h}}_{ki}^{2^B}\right]$ according to

$$n_o = \underset{1 \leq n \leq 2^B}{\arg\max} \left| \tilde{\mathbf{h}}_{ki}^{*} \hat{\mathbf{h}}_{ki}^{n} \right|^2, \tag{4}$$

in which B is the number of bits used to quantize $\mathbf{H}_{ki}$ and $\tilde{\mathbf{h}}_{ki} = \frac{\mathbf{h}_{ki}}{\|\mathbf{h}_{ki}\|}$ is the channel direction vector.

After quantizing all the matrices of its local CSI, we assume that transmitter i sends the corresponding optimal indexes to all other transmitters which share the same codebook, allowing these transmitters to reconstruct the quantized local knowledge of transmitter i. Let us now define the quantization error as $e_{ki} = 1 - \frac{\left|\hat{\mathbf{h}}_{ki}^{*} \mathbf{h}_{ki}\right|^2}{\|\mathbf{h}_{ki}\|^2}$ and adopt the same model in [22], [23] which relies on the theory of quantization cell approximation. The cumulative distribution function (CDF) of e_{ki} is then given by

$$\Pr(e_{ki} \leq \varepsilon) = \begin{cases} 2^B \varepsilon^Q, & 0 \leq \varepsilon \leq 2^{-\frac{B}{Q}} \\ 1, & \varepsilon > 2^{-\frac{B}{Q}} \end{cases} \tag{5}$$

where $Q = N_{\mathrm{t}} N_{\mathrm{r}} - 1$.

Recall that we consider a finite capacity backhaul in which we perform a quantization scheme to reduce the CSI sharing cost. Since these limited capacity backhaul links are also used for actual data transfer, one additional way to allocate more capacity for CSI sharing is to decrease this data transfer. This is generally accomplished by means of caching in which we describe in the following.

C. Cache-enabled Transmitters

Several studies have shown that certain types of content are relatively more requested than others such as viral videos with millions of views, share of popular people in social media, well-known news and blog pages. Indeed, accessing the same information by many users is one of the major reasons for network congestion and latency increase. Let us assume that each transmitter is associated with a storage unit (cache) which stores the content with respect to a certain popularity profile.

At the transmitters, for ease of analysis, we consider the trivial approach that consists in storing the most popular content, which results from the reasonable fact that a user's request matches with the global popular contents [3]. Indeed, the content popularity can be described by the probability distribution function, given by the following expression

$$f_{\mathrm{pop}}(f, \eta) = \begin{cases} (\eta - 1) f^{-\eta}, & f \geq 1 \\ 0, & f < 1 \end{cases} \tag{6}$$

where f represents a point in the support of the corresponding content, and η stands for a factor that describes the steepness of the popularity distribution curve. Lower values of η corresponds to a uniform behaviour (almost all contents have the same popularities), whereas a high η value would results in a steeper distribution (very few contents are highly popular than the rest). Now, suppose that each transmitter stores the contents up to f_0 (namely cache size) from the distribution in (6). Then, the probability that a content request falls in the range $\Delta = [0, f_0]$, namely *cache hit probability*, can be calculated as

$$\begin{aligned} \Pr_{\mathrm{hit}} &= \int_0^{f_0} f_{\mathrm{pop}}(f, \eta)\, df \\ &= 1 - f_0^{1-\eta}. \end{aligned} \tag{7}$$

Consequently, the probability that a content demand is missing from the cache can be given by $\Pr_{\mathrm{miss}} = 1 - \Pr_{\mathrm{hit}} = f_0^{1-\eta}$. Based on the above model which considers IA and caching capabilities at the transmitters, we next focus on the performance analysis of the system.

III. Performance Analysis

In this section, we derive the expression for the total average transmission rate and characterize an operational regime where caching is beneficial. Then, we provide an optimization problem that maximizes the transmission rate.

A. Average Transmission Rate

As explained in the preceding section, the IA vectors are designed based on the available CSI that results after the transmitting nodes quantize and share their perfect local knowledge between each other. Thus, the IA technique adopted is able to completely suppress the ISI since local CSI is perfectly known, but not the IUI because of the quantization process which leads to imperfect global CSI at the transmitters. Under

such conditions and using the results in [24], the signal-to-interference-plus-noise ratio (SINR) for stream m at receiver k can be expressed as

$$\gamma_k^m = \frac{\alpha_{kk}\left|(\hat{\mathbf{u}}_k^m)^* \mathbf{H}_{kk}\hat{\mathbf{v}}_k^m\right|^2}{\sigma^2 + \sum\limits_{\substack{i=1\\ i\neq k}}^{L}\alpha_{ki}\sum\limits_{j=1}^{d_i}\left|(\hat{\mathbf{u}}_k^m)^* \mathbf{H}_{ki}\hat{\mathbf{v}}_i^j\right|^2} = \frac{\alpha_{kk}\left|(\hat{\mathbf{u}}_k^m)^* \mathbf{H}_{kk}\hat{\mathbf{v}}_k^m\right|^2}{\sigma^2 + \sum\limits_{\substack{i=1\\ i\neq k}}^{L}\alpha_{ki}\left\|\mathbf{h}_{ki}\right\|^2 e_{ki}\sum\limits_{j=1}^{d_i}\left|\mathbf{w}_{ki}^*\mathbf{s}_{k,i}^{m,j}\right|^2}, \quad (8)$$

where $\mathbf{w}_{ki}$ is a unit norm vector isotropically distributed in the null space of $\hat{\mathbf{h}}_{ki}$, $\mathbf{s}_{k,i}^{m,j} = \hat{\mathbf{v}}_i^j \otimes (\hat{\mathbf{u}}_k^m)^*$ ($\otimes$ is the Kronecker product), $\hat{\mathbf{v}}_k^m$ and $\hat{\mathbf{u}}_k^m$ are the precoding and combining vectors, respectively, designed based on the available CSI described in the previous section.

Using the SINR expression in (8), the *instantaneous rate* for user k can be given by

$$R_k = \sum_{m=1}^{d_k}\log_2\left(1 + \frac{\alpha_{kk}\left|(\hat{\mathbf{u}}_k^m)^* \mathbf{H}_{kk}\hat{\mathbf{v}}_k^m\right|^2}{\sigma^2 + \sum\limits_{\substack{i=1\\ i\neq k}}^{L}\alpha_{ki}\sum\limits_{j=1}^{d_i}\left|(\hat{\mathbf{u}}_k^m)^* \mathbf{H}_{ki}\hat{\mathbf{v}}_i^j\right|^2}\right). \quad (9)$$

We assume that the quantization error plays the role of an additional source of Gaussian noise, regardless of its distribution [25]. Under this assumption, the *average rate* for user k achieved by IA can be written as

$$\bar{R}_k = \sum_{m=1}^{d_k}\mathbb{E}\left[\log_2\left(1 + \frac{\alpha_{kk}\left|(\hat{\mathbf{u}}_k^m)^* \mathbf{H}_{kk}\hat{\mathbf{v}}_k^m\right|^2}{\sigma^2 + \sum\limits_{\substack{i=1\\ i\neq k}}^{L}\alpha_{ki}\sum\limits_{j=1}^{d_i}\mathbb{E}\left[\left|(\hat{\mathbf{u}}_k^m)^* \mathbf{H}_{ki}\hat{\mathbf{v}}_i^j\right|^2\right]}\right)\right] \quad (10)$$

where we note that the outer expectation is only over the direct channel. Therefore, the leakage interference terms $(\hat{\mathbf{u}}_k^m)^* \mathbf{H}_{ki}\hat{\mathbf{v}}_i^j$ are nothing but an independent sources of additive Gaussian noise, irrespective of their actual distribution. The following lemma will be useful for the rest of analysis.

Lemma 1. *The average rate for user k can be written in exponential form as*

$$\bar{R}_k = d_k \log_2(e) e^{\frac{1}{\beta_k}} E_1\left(\frac{1}{\beta_k}\right) \quad (11)$$

where $\beta_k = \frac{P\zeta_{kk}}{d_{kk}\left(\sigma^2 + P2^{1-\frac{B}{Q}}\sum\limits_{i=1,i\neq k}^{L}\zeta_{ki}\right)}$ *and* $E_1(.)$ *is the exponential integral defined as* $E_1(a) = \int\limits_1^\infty t^{-1}e^{-at}dt$.

Proof. The proof is provided in Appendix A. □

Note that the rate metrics we derived so far are related to the wireless downlink transmission achieved by IA, whereas in the following, we shall derive more elaborated rate expressions by taking into account caching and limited backhaul aspects. We shall now define the *instantaneous transmission rate* for user k, such as

$$r_k = \begin{cases} R_k, & f_r \in \Delta \\ C_{\mathrm{kd}}, & f_r \notin \Delta \end{cases} \quad (12)$$

where f_r represents the requested content and Δ is the available catalog in the local cache. The main intuition behind this definition is the following. If the requested content exists in the local cache, the amount of rate given to the user is R_k. On the other hand, if the content does not exist in the local cache, the content is fetched from the Internet via the backhaul, thus the given rate is C_{kd}. We assume that $C_{\mathrm{kd}} < R_k$ always holds. This assumption comes from the motivation that the backhaul link capacity in 5G networks is expected to be a limited factor compared to wireless link capacity, especially in ultra-dense deployment of base stations (BSs) [1]. Given this definition and assumption, we state the following theorem.

Theorem 1 (Average Transmission Rate)**.** *The average transmission rate for user k can be given by*

$$\bar{r}_k = d_k \log_2(e) e^{\frac{1}{\beta_k}} E_1\left(\frac{1}{\beta_k}\right)(1 - f_0^{1-\eta_k}) + C_{\mathrm{kd}} f_0^{1-\eta_k}. \quad (13)$$

Proof. We have $\bar{r}_k = \mathbb{E}[R_k]\mathrm{Pr}_{\mathrm{hit}} + C_{\mathrm{kd}}\mathrm{Pr}_{\mathrm{miss}} = \bar{R}_k(1 - f_0^{1-\eta_k}) + C_{\mathrm{kd}} f_0^{1-\eta_k}$. By replacing $\bar{R}_k$ by its expression given in Lemma 1, the result in (13) follows. □

Consequently, the *total average transmission rate* of the system can be found straightforwardly by taking the sum over all the pairs of the expression in (13) as follows

$$\bar{r}_T = \sum_{k=1}^{L}\left(d_k \log_2(e) e^{\frac{1}{\beta_k}} E_1\left(\frac{1}{\beta_k}\right)(1 - f_0^{1-\eta_k}) + C_{\mathrm{kd}} f_0^{1-\eta_k}\right) \quad (14)$$

Remark 1. *The more storage (caching) capacity increases, the more missing probability decreases, and consequently the hitting probability increases. Thus, for a fixed steepness factor η, the support of cached contents (represented by f_0) has an important impact on the total average transmission rate. Similar remarks can be given for the number of active pairs L and the number of bits B.*

B. Operational Caching Regime

The steepness factor η describes how much steep is the popularity distribution function, and it depends on requested contents of the corresponding user. In other words, a high value of η results from the fact that some contents are much more popular than other contents and thus, because the cache contains the most popular contents, the hitting probability will be high. On the other side, a low value of η is due to (more or

less) the same popularity of the requested contents and then the hitting probability can not reach important values. This analysis can be resumed by the following proposition.

Proposition 1. *The average rate for user k (with $k = 1, ..., L$) is an increasing function with respect to its corresponding steepness factor η_k.*

Proof. The first derivative $\frac{d\bar{r}_k}{d\eta_k} = (\bar{R}_k - C_{\mathrm{kd}}) f_0^{1-\eta_k} \ln f_0$. This derivative is positive since we have $\bar{R}_k > C_{\mathrm{kd}}$, and hence the statement of Proposition 1 follows. □

We will now derive two bounds based on the steepness factor η_k of pair k, under different observations and constraints on the average transmission rate:

1) Minimum Guaranteed Transmission Rate: A minimum desired average transmission rate at user k can be expressed using the following inequality $\bar{r}_k \geq p\bar{R}_k$, where $p < 1$ is a QoS factor that dictates how much the actual transmission rate should be achieved. Using this inequality, we can derive a lower bound on η_k as

$$\bar{r}_k = \bar{R}_k(1 - f_0^{1-\eta_k}) + C_{\mathrm{kd}} f_0^{1-\eta_k} \geq p\bar{R}_k, \tag{15}$$

thus results in a steepness factor

$$\eta_k \geq 1 - \frac{\ln\left(\frac{\bar{R}_k(1-p)}{\bar{R}_k - C_{\mathrm{kd}}}\right)}{\ln f_0}. \tag{16}$$

2) Constant Average Rate Variation: One could notice that there exists a regime where the average transmission rate has almost a constant variation in function of η_k. To detect this regime, a simple but effective way is to consider $\frac{d\bar{r}_k}{d\eta_k} < \epsilon$, where ϵ is a parameter that describes how much the first derivative is close to zero. Under this consideration, we can calculate a lower bound on η_k as

$$\frac{d\bar{r}_k}{d\eta_k} = f_0^{1-\eta_k}(\bar{R}_k - C_{\mathrm{kd}}) \ln f_0 < \epsilon, \tag{17}$$

thus gives a steepness factor

$$\eta_k > 1 - \frac{\ln\left(\frac{\epsilon}{(\bar{R}_k - C_{\mathrm{kd}}) \ln f_0}\right)}{\ln f_0}. \tag{18}$$

Let $\eta_{k1} = 1 - \frac{\ln\left(\frac{\bar{R}_k(1-p)}{\bar{R}_k - C_{\mathrm{kd}}}\right)}{\ln f_0}$ and $\eta_{k2} = 1 - \frac{\ln\left(\frac{\epsilon}{(\bar{R}_k - C_{\mathrm{kd}}) \ln f_0}\right)}{\ln f_0}$. Using these two bounds, we can define the regime where caching is beneficial for user k in terms of average rate. Specifically, for a minimum guaranteed rate defined by $\bar{r}_k \geq p\bar{R}_k$ and for an average rate variation $\frac{d\bar{r}_k}{d\eta_k} \geq \epsilon$, caching is gainful for user k (i.e. can satisfy these latter conditions) if its steepness factor is between these intervals, such as $\eta_{k1} \leq \eta_k \leq \eta_{k2}$.

C. Rate Maximization

The total transmission rate in our setup is a function of various parameters. Among these parameters, we focus on the number of pairs L. We investigate the optimal value of L by defining and solving an optimization problem which seeks to maximize the total average transmission rate. In fact, as it can be seen in (14), solving this problem for the general case is of high complexity. Therefore, before proceeding in the definition of this optimization problem and for the sake of simplicity, we make the following assumptions: (i) all the transmitters have the same number of streams d, (ii) all the users have the same steepness factor denoted by η, and (iii) we use the extended Wyner model (1D system) where the path loss coefficient from transmitter i to user k is given by $\zeta^{|k-i|}$. We can represent this path loss model using the matrix

$$\mathbf{A} = \begin{pmatrix} 1 & \zeta & \zeta^2 & \cdots & \zeta^{L-1} \\ \zeta & 1 & \zeta & \cdots & \zeta^{L-2} \\ \zeta^2 & \zeta & 1 & \cdots & \zeta^{L-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \zeta^{L-1} & \zeta^{L-2} & \zeta^{L-3} & \cdots & 1 \end{pmatrix}. \tag{19}$$

Under these assumptions and recalling that $C_{\mathrm{kd}} = \frac{C_{\mathrm{d}}}{L}$, we can re-express (14) as

$$\bar{r}_{T_{\mathrm{s}}} = \begin{cases} 2a \sum_{i=1}^{\frac{L}{2}} e^{a_i} E_1(a_i) + b & \text{if } L \text{ is even} \\ 2a \sum_{i=1}^{\lfloor \frac{L}{2} \rfloor} e^{a_i} E_1(a_i) + a e^{b_1} E_1(b_1) + b & \text{if } L \text{ is odd} \end{cases} \tag{20}$$

where $a_i = d\sigma^2 P^{-1} + d2^{1-\frac{B}{Q}}(1-\zeta)^{-1}(2\zeta - \zeta^{L-i+1} - \zeta^i)$, $b_1 = d\sigma^2 P^{-1} + d2^{1-\frac{B}{Q}}(1-\zeta)^{-1}2(\zeta - \zeta^{\lfloor \frac{L}{2} \rfloor + 1})$, $a = d\log_2(e)(1 - f_0^{1-\eta})$, $b = C_{\mathrm{d}} f_0^{1-\eta}$ and $\lfloor \frac{L}{2} \rfloor$ is the largest integer not greater than $\frac{L}{2}$.

Remark 2. *To ensure the feasibility of the IA problem, the system parameters should satisfy the following condition (given in [26]) $N_{\mathrm{t}} + N_{\mathrm{r}} \geq d(L+1)$. Without loss of generality, we assume that the number of pairs L satisfies this condition.*

Now, we can define our optimization problem which seeks to maximize the total average transmission rate in (20) with respect to the number of pairs L. This is formally stated as

$$\underset{L}{\text{maximize}} \quad \bar{r}_{T_{\mathrm{s}}}(L) \tag{21}$$

$$\text{subject to} \quad L^2(L-1)B \leq \left(C_{\mathrm{c}} + (1 - f_0^{1-\eta})C_{\mathrm{d}}\right)\tau \tag{22}$$

where τ is the slot duration. The term at the left hand side of (22) represents the total number of bits (needed for CSI sharing) and is obtained from the fact that we have L transmitters, each of which shares $L-1$ channels (using LB bits for each channel) to $L-1$ other transmitters. The right hand side of (22) shows how caching mitigates the backhaul usage, allowing higher capacity of backhaul links which are used for CSI sharing. In detail, caching saves $(1 - f_0^{1-\eta})C_{\mathrm{d}}$ of the backhaul capacity usage, and thus this saved part can be used, in addition to C_{c}, in the CSI sharing process. For the optimization problem, we first describe the behavior of $\bar{r}_{T_{\mathrm{s}}}$ in the following result.

Proposition 2. *The total average rate $\bar{r}_{T_{\mathrm{s}}}$ is an increasing function with respect to the number of pairs L (with $L \geq 3$), for sufficiently small ζ values.*

Proof. The proof is provided in Appendix B. □

Using the above proposition, the optimal number of pairs (denoted by L_{opt}) can be easily obtained by setting $L = 3$ and increasing it until condition (22) is not satisfied. Note that Proposition 2 holds for sufficiently small values of ζ. To solve the optimization problem for arbitrary ζ values ($\zeta < 1$), we use the following procedure.

Step 1: Compute $\bar{r}_{T_s}$ for all L that satisfy conditions (22) and $d(L+1) \leq N_t + N_r$.

Step 2: Select the maximum among the computed $\bar{r}_{T_s}$ values and take the corresponding L as L_{opt}.

Notice that for a fixed number of pairs L, the same analysis can be done for the number of bits B. Using the condition in (22) and since $\bar{r}_{T_s}$ is an increasing function with B, an increase of bound $C_c + (1 - f_0^{1-\eta})C_d$ allows us to use more number of bits for the quantization process, and thus to get better total average rate $\bar{r}_{T_s}$.

IV. Numerical Results

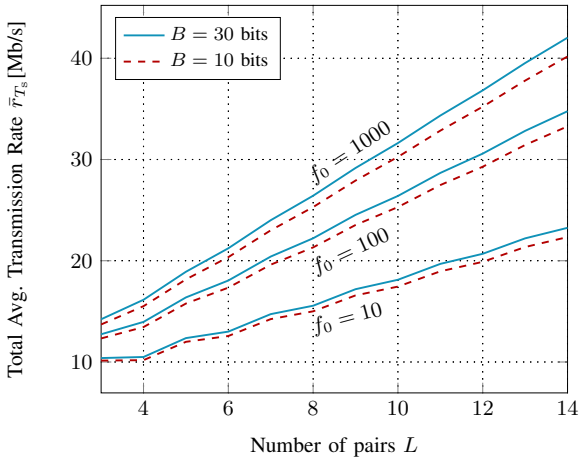

Figure 2: $\bar{r}_{T_s}$ vs. L, with $\eta = 1.2$.

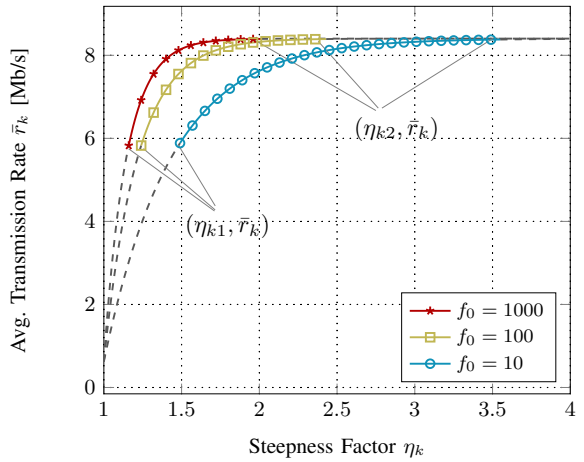

Figure 3: $\bar{r}_k$ vs. η_k, with $L = 8$, $B = 30$ bits, $p = 0.7$ and $\epsilon = 0.05$.

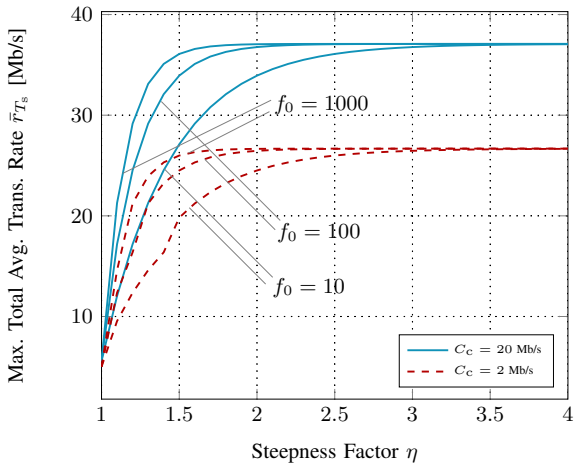

Figure 4: Maximum $\bar{r}_{T_s}$ vs. η, with $B = 30$ bits.

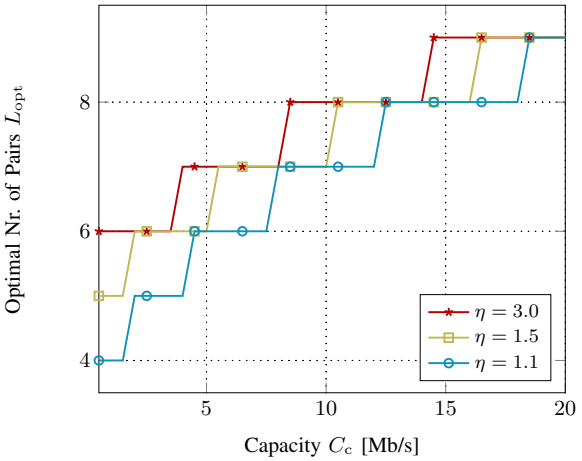

Figure 5: L_{opt} vs. C_c, with $B = 30$ bits and $f_0 = 10$.

In this section we present our numerical results to validate the analysis conducted in the previous section. For ease of exposition, we consider a setup with $N_t = N_r = 15$, SNR $= 10\log_{10}\left(\frac{P}{\sigma^2}\right) = 10$ dB, $d = 2$, $\zeta = 0.3$, $\tau = 1$ ms, $C_d = 5$ Mb/s and bandwidth $BW = 10$ MHz per transmitter.

In Fig. 2 we plot the variation of the total average transmission rate with respect to the number of active pairs L. It can be seen that $\bar{r}_{T_s}$ can be significantly increased by increasing the size of the catalog in transmitters, namely f_0. Furthermore, the impact of increasing the number of bits B is higher for larger f_0.

The evolution of average transmission rate with respect to the steepness factor is depicted in Fig. 3. By looking into the feasible values of $\bar{r}_k$ in which η_k is between η_{k1} and η_{k2} (recall Section III-B), we can notice that $\bar{r}_k$ increases more dramatically as the size of catalog increases. Additionally, keeping aside the fact that the transmission rate is not guaranteed below η_{k1}, the variations after η_{k2} are almost constant regardless of different catalog sizes. This confirms our expressions derived for the operational caching regime.

Fig. 5 illustrates the variation of $L_{\rm opt}$ with respect to the capacity $C_{\rm c}$, for different values of steepness factor η. It can be noticed that, for the same η, $L_{\rm opt}$ increases with $C_{\rm c}$ and can reach larger values for higher steepness factor η. Recall that $L_{\rm opt}$ also depends on the capacity $C_{\rm d}$ and the cache size f_0 (see the bound in (22)).

The impact of steepness factor on the maximum total average rate is shown in Fig. 4 for different values of the backhaul capacity dedicated to the CSI sharing (namely $C_{\rm c}$). Given the fact that maximum total average rate is achieved by finding the optimal number of pairs $L_{\rm opt}$, improvement of this rate for a specific range of η (as in operational caching regime) can be further fueled by increasing $C_{\rm c}$ and/or f_0. This behaviour in fact validates our analysis.

V. Conclusion

In this paper, we have analyzed the performance of the interference alignment technique applied to a L-user MIMO system, under the limited backhaul capacity and caching capabilities at the transmitters. Under some specific assumptions and considerations, we derived expressions of the total average transmission rate $\bar{r}_{T_{\rm s}}$ and the operational caching regime has been determined based on the content popularity profile. A key observation of this work is that, under this regime, cache-enabled base stations can significantly increase the $\bar{r}_{T_{\rm s}}$ as compared to traditional BSs. We also showed the existence of an optimum number of pairs for the total average rate, and that this optimum number depends on several parameters such as capacity $C_{\rm c}$, steepness factor η and storage size f_0.

The implication of caching in wireless networks is of high interest and requires further investigations. For instance, solving the optimization problems for the general case would be an interesting result. In addition, the impact of caching on other interference management techniques can be investigated. Lastly, heterogeneous network scenarios, including macro cells and small cells deployments, can be added as an additional layer to reveal the benefits of caching and IA methods for future networks.

References

[1] J. G. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?" *arXiv preprint arXiv:1405.2957*, 2014.

[2] E. Baştuğ, M. Bennis, and M. Debbah, "Living on the Edge: The role of proactive caching in 5G wireless networks," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82–89, August 2014.

[3] E. Baştuğ, M. Bennis, M. Kountouris, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," *EURASIP Journal on Wireless Communications and Networking*, Accepted (2015).

[4] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," *arXiv preprint arXiv:1409.7626*, 2014.

[5] K. Hamidouche, W. Saad, and M. Debbah, "Many-to-many matching games for proactive social-caching in wireless small cell networks," in *12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt'14)*, May 2014, pp. 569–574.

[6] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Match to cache: Optimizing user association and backhaul allocation in cache-aware small cell networks," in *IEEE International Conference on Communications (ICC'2015)*, Submitted (2014).

[7] N. Karamchandani, U. Niesen, M. A. Maddah-Ali, and S. Diggavi, "Hierarchical coded caching," in *IEEE International Symposium on Information Theory (ISIT'14)*, June 2014, pp. 2142–2146.

[8] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Communications Magazine*, vol. 51, no. 4, pp. 142–149, 2013.

[9] K. Poularakis, G. Iosifidis, and L. Tassiulas, "Approximation algorithms for mobile data caching in small cell networks," *IEEE Transactions on Communications*, vol. 62, no. 10, pp. 3665–3677, October 2014.

[10] J. Pääkkönen, C. Hollanti, and O. Tirkkonen, "Device-to-device data storage with regenerating codes," *arXiv preprint arXiv:1411.1608*, 2014.

[11] A. Liu and V. Lau, "Cache-enabled opportunistic cooperative MIMO for video streaming in wireless systems," *IEEE Transactions on Signal Processing*, vol. 62, no. 2, pp. 390–402, January 2014.

[12] ——, "Cache-induced opportunistic MIMO cooperation: A new paradigm for future wireless content access networks," in *IEEE International Symposium on Information Theory (ISIT'14)*, June 2014, pp. 46–50.

[13] V. R. Cadambe and S. A. Jafar, "Interference alignment and degrees of freedom of the k-user interference channel," *IEEE Transactions on Information Theory*, vol. 54, no. 8, pp. 3425–3441, August 2008.

[14] T. Gou and S. A. Jafar, "Degrees of freedom of the k user $M \times N$ MIMO interference channel," *IEEE Transactions on Information Theory*, vol. 56, no. 12, pp. 6040–6057, December 2010.

[15] H. Bolcskei and I. Thukral, "Interference alignment with limited feedback," in *IEEE International Symposium on Information Theory (ISIT'09)*. IEEE, June 2009, pp. 1759–1763.

[16] R. T. Krishnamachari and M. K. Varanasi, "Interference alignment under limited feedback for MIMO interference channels," *IEEE Transactions on Signal Processing*, vol. 61, no. 15, pp. 3908–3917, August 2013.

[17] X. Chen and C. Yuen, "Performance analysis and optimization for interference alignment over MIMO interference channels with limited feedback," *arXiv preprint arXiv:1402.0295*, 2014.

[18] M. Rezaee, M. Guillaud, and F. Lindqvist, "CSIT sharing over finite capacity backhaul for spatial interference alignment," in *IEEE International Symposium on Information Theory Proceedings (ISIT'13)*. IEEE, July 2013, pp. 569–573.

[19] N. B. Chang and M. Liu, "Optimal channel probing and transmission scheduling for opportunistic spectrum access," *IEEE/ACM Transactions on Networking*, vol. 17, no. 6, pp. 1805–1818, December 2009.

[20] P. Chaporkar and A. Proutiere, "Optimal joint probing and transmission strategy for maximizing throughput in wireless systems," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1546–1555, October 2008.

[21] P. Chaporkar, A. Proutiere, H. Asnani, and A. Karandikar, "Scheduling with limited information in wireless systems," in *Proceedings of the tenth ACM international symposium on Mobile ad hoc networking and computing*, ser. MobiHoc '09. New York, NY, USA: ACM, 2009, pp. 75–84.

[22] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-antenna downlink channels with limited feedback and user selection," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1478–1491, September 2007.

[23] K. Huang and V. Lau, "Stability and delay of zero-forcing SDMA with limited feedback," *IEEE Transactions on Information Theory*, vol. 58, no. 10, pp. 6499–6514, Oct 2012.

[24] M. Deghel, M. Assaad, and M. Debbah, "System performance of interference alignment under TDD mode with limited backhaul capacity," in *IEEE International Conference on Communications (ICC'15)*, London, UK, 2015, [Online] http://goo.gl/NhGBKk.

[25] O. El Ayach, A. Lozano, and R. Heath, "On the overhead of interference alignment: Training, feedback, and cooperation," *IEEE Transactions on Wireless Communications*, November 2012.

[26] C. M. Yetis, T. Gou, S. A. Jafar, and A. H. Kayran, "On feasibility of interference alignment in MIMO interference networks," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4771–4782, September 2010.

Appendix A
Proof of Lemma 1

We start by calculating the inner expectation in (10) given by $\mathbb{E}\left[\left|\left(\hat{\mathbf{u}}_k^m\right)^* \mathbf{H}_{ki} \hat{\mathbf{v}}_i^j\right|^2\right]$. From (8), we have the following:

$\mathbb{E}\left|(\hat{\mathbf{u}}_k^m)^* \mathbf{H}_{ki} \hat{\mathbf{v}}_i^j\right|^2 = \mathbb{E}\left[\|\mathbf{h}_{ki}\|^2 e_{ki} \left|\mathbf{w}_{ki} \mathbf{s}_{k,i}^{m,j}\right|^2\right]$. According to [17, Appendix A], $\|\mathbf{h}_{ki}\|^2 e_{ki} \left|\mathbf{w}_{ki} \mathbf{s}_{k,i}^{m,j}\right|^2$ is equal to $2^{-\frac{B}{N_t N_r - 1}} \chi^2(2)$ in distribution. Since $\chi^2(2)$ has a mean equal to 2, then we have $\mathbb{E}\left[\left|(\hat{\mathbf{u}}_k^m)^* \mathbf{H}_{ki} \hat{\mathbf{v}}_i^j\right|^2\right] = 2^{1-\frac{B}{N_t N_r - 1}} = 2^{1-\frac{B}{Q}}$. Thus, the expression in (10) can be re-expressed as the following:

$$\begin{aligned}\bar{R}_k &= \sum_{m=1}^{d_k} \mathbb{E}\left[\log_2\left(1 + \frac{\alpha_{kk}\left|(\hat{\mathbf{u}}_k^m)^* \mathbf{H}_{kk} \hat{\mathbf{v}}_k^m\right|^2}{\sigma^2 + \sum\limits_{\substack{i=1 \\ i \neq k}}^{L} \alpha_{ki} d_i 2^{1-\frac{B}{Q}}}\right)\right] \\ &= \sum_{m=1}^{d_k} \mathbb{E}\left[\log_2\left(1 + \frac{P\zeta_{kk}\left|(\hat{\mathbf{u}}_k^m)^* \mathbf{H}_{kk} \hat{\mathbf{v}}_k^m\right|^2}{d_k(\sigma^2 + P 2^{1-\frac{B}{Q}} \sum\limits_{\substack{i=1 \\ i \neq k}}^{L} \zeta_{ki})}\right)\right].\end{aligned} \tag{23}$$

We now need to calculate the outer expectation. For this, we use the result in [25]: $\mathbb{E}\left[\log_2\left(1 + \frac{P\zeta_{kk}|(\hat{\mathbf{u}}_k^m)^* \mathbf{H}_{kk} \hat{\mathbf{v}}_k^m|^2}{d_k \sigma_k^2}\right)\right] = \log_2(e) e^{\frac{1}{\beta_k}} E_1\left(\frac{1}{\beta_k}\right)$, where $\sigma_k^2 = \sigma^2 + P 2^{1-\frac{B}{Q}} \sum\limits_{i=1, i \neq k}^{L} \zeta_{ki}$, $\beta_k = \frac{P\zeta_{kk}}{d_k \sigma_k^2}$ and $E_1(.)$ is the exponential integral function. Therefore, the average rate for user k can be given by

$$\begin{aligned}\bar{R}_k &= \sum_{m=1}^{d_k} \log_2(e) e^{\frac{1}{\beta_k}} E_1\left(\frac{1}{\beta_k}\right) \\ &= d_k \log_2(e) e^{\frac{1}{\beta_k}} E_1\left(\frac{1}{\beta_k}\right).\end{aligned} \tag{24}$$

This concludes the proof. ■

Appendix B
Proof of Proposition 2

We recall that $\bar{r}_{T_s}$ is given by the following

$$\bar{r}_{T_s} = \begin{cases} 2a \sum\limits_{i=1}^{\frac{L}{2}} e^{a_i} E_1(a_i) + b & \text{if } L \text{ is even} \\ 2a \sum\limits_{i=1}^{\lfloor \frac{L}{2} \rfloor} e^{a_i} E_1(a_i) + a e^{b_1} E_1(b_1) + b & \text{if } L \text{ is odd} \end{cases} \tag{25}$$

where $a_i = d\sigma^2 P^{-1} + d 2^{1-\frac{B}{Q}} (1-\zeta)^{-1} (2\zeta - \zeta^{L-i+1} - \zeta^i)$, $b_1 = d\sigma^2 P^{-1} + d 2^{1-\frac{B}{Q}} (1-\zeta)^{-1} 2(\zeta - \zeta^{\lfloor \frac{L}{2} \rfloor + 1})$, $a = d \log_2(e)(1 - f_0^{1-\eta})$ and $b = C_{\mathrm{d}} f_0^{1-\eta}$. For sufficiently small values of ζ, we can suppose that $2\zeta + 2\zeta^2 + 2\zeta^3 + \cdots \approx 2\zeta$, or equivalently $\zeta + \zeta^2 + \zeta^3 + \cdots \approx \zeta$. To justify this, take for instance $\zeta = 0.1$ which yields $0.1 + 0.1^2 + 0.1^3 + \cdots = 0.11 \approx 0.1$.

Consequently, we get $(1-\zeta)^{-1}(\zeta - \zeta^{L-i+1}) = \zeta + \cdots + \zeta^{L-i} \approx \zeta$, $(1-\zeta)^{-1}(\zeta - \zeta^{\lfloor \frac{L}{2} \rfloor + 1}) = \zeta + \cdots + \zeta^{\lfloor \frac{L}{2} \rfloor} \approx \zeta$ and also $(1-\zeta)^{-1}(\zeta - \zeta^i) \approx \zeta$ (for $i > 1$). Therefore, the expression in (25) simplifies to

$$\bar{r}_{T_s} \approx 2a e^{c_1} E_1(c_1) + (L-2) a e^{c_2} E_1(c_2) + b, \tag{26}$$

where $c_1 = d\sigma^2 P^{-1} + d 2^{1-\frac{B}{Q}} \zeta$ and $c_2 = d\sigma^2 P^{-1} + d 2^{1-\frac{B}{Q}} 2\zeta$. Based on expression (26), we conclude that the total average rate $\bar{r}_{T_s}$ is linear with the number of pairs L. Hence, the desired result holds. ■