# Fundamental Limits of Wireless Caching under Uneven-Capacity Channels

Eleftherios Lampiris, Jingjing Zhang, Osvaldo Simeone, Petros Elia

*Abstract*— **This work identifies the fundamental limits of cache-aided coded multicasting in the presence of the well-known 'worst-user' bottleneck. This stems from the presence of receiving users with uneven channel capacities, which often forces the rate of transmission of each multicasting message to be reduced to that of the slowest user. This bottleneck, which can be detrimental in general wireless broadcast settings, motivates the analysis of coded caching over a standard Single-Input-Single-Output (SISO) Broadcast Channel (BC) with $K$ cache-aided receivers, each with a generally different channel capacity. For this setting, we design a communication algorithm that is based on superposition coding that capitalizes on the realization that the user with the worst channel may not be the real bottleneck of communication. We then proceed to provide a converse that shows the algorithm to be near optimal, identifying the fundamental limits of this setting within a multiplicative factor of $4$. Interestingly, the result reveals that, even if several users are experiencing channels with reduced capacity, the system can achieve the same optimal delivery time that would be achievable if all users enjoyed maximal capacity.**

## I. Introduction

The seminal work in [1] showed how adding caches to receiving nodes can substantially reduce the time required to deliver content. Specifically, reference [1] studied the case in which a transmitter with access to a library of $N$ unit-sized files serves – via a wired, single-stream, unit-capacity bottleneck link – $K$ cache-aided receivers/users. Each user is equipped with a cache of size equal to a fraction $\gamma \in [0,1]$ of the size of the library, so that $K\gamma$ is the cumulative cache size normalized by the library size. For this setting, the authors of [1] proposed a novel cache placement algorithm and a novel multicast transmission policy that delivers any set of $K$ files to the receivers with (normalized) delay at most

$$T_{MN} = \frac{K(1-\gamma)}{K\gamma+1} \tag{1}$$

thus revealing a speed-up factor of $K\gamma+1$ compared to the delay $K(1-\gamma)$ corresponding to a standard scheme that serves each user in turn.

The delay (1) is obtained by a *coded caching* approach that is based on the transmission of a sequence of multicast messages that convey information to several users at a time (even if these users requested different content), with users decoding their desired information by means of cache-aided interference cancellation. In this scheme, each multicast message consists of a XOR $X_\sigma$ that carries information to a subset $\sigma \subset [K] \triangleq [1,2,\dots,K]$ of $|\sigma| = K\gamma+1$ users at a time.

While the promised speedup factor of $K\gamma+1$ in (1) is proportional to the normalized *cumulative cache size of the network*, it was quickly realized that a variety of bottlenecks severely hamper this performance. These include the subpacketization bottleneck [2]–[8], the uneven cache sizes bottleneck [9]–[12], and the bottleneck studied here that arises from uneven channel capacities between the transmitter and the users. This last bottleneck is particularly relevant in wireless scenarios with multicasting. Such networks produce "slower" users that can force the multicast rates to be reduced down to a level that can be decoded by these users. This can diminish the coded caching gains and could pose a serious limitation to any effort to implement cache-aided coded multicasting in wireless settings.

***Example* 1.** *Let us consider the wireless Single-Input-Single-Output (SISO) Broadcast Channel (BC) with $K$ users, each equipped with a cache of normalized size $\gamma$, and let us further assume that all users have maximal normalized unit capacity, except for one user that has a normalized link capacity equal to $\frac{1}{K}+\gamma<1$. It is easy to see that a (naive) transmission of the sequence of the XORs from [1] would induce the delay*

$$\begin{aligned} T &= \frac{1-\gamma}{\frac{1+K\gamma}{K}} + \frac{(K-K\gamma-1)(1-\gamma)}{1+K\gamma} && (2)\\ &= 2T_{MN} - (1-\gamma) \approx 2T_{MN} && (3)\end{aligned}$$

*which is approximately double the delay $T_{MN}$ in (1) that we would have if all users enjoyed unit normalized link capacities. It is also worth noting that approximately the same delay $T$ in (2) would be obtained if we treated the slow user separately from the rest using time sharing. Essentially, whether with a naive or with a separated approach that excludes the slow user from coded caching, a single slow user can cause the worst-case delivery time to double, and the overall multicasting gain to be cut in half.*

### A. Related Work

The importance of the uneven-channel bottleneck in coded caching has been acknowledged in a large number of recent

Eleftherios is with the Electrical Engineering and Computer Science Department, Technische Universität Berlin, 10587 Berlin, Germany (lampiris@tu-berlin.de). Petros is with the Communication Systems Department at EURECOM, Sophia Antipolis, 06410, France (elia@eurecom.fr). Their work is supported by the ANR project ECOLOGICAL-BITS-AND-FLOPS and the European Research Council under the EU Horizon 2020 research and innovation program / ERC grant agreement no. 725929. (ERC project DUALITY). Jingjing and Osvaldo are with the Department of Informatics, King's College London, London, UK (jingjing.1.zhang@kcl.ac.uk, osvaldo.simeone@kcl.ac.uk). Their work is supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 725731). This work was conducted while Eleftherios was employed by EURECOM.

works that seek to understand and ameliorate this limitation [13]–[28]. For example, reference [13] focuses on the uneven link-capacity SISO BC where each user experiences a distinct channel strength, and proposes algorithms that outperform the naive implementation of the algorithm of [1] whereby each coded message is transmitted at a rate equal to the rate of the worst user whose message appears in the corresponding XOR operation. Under a similar setting, the work in [16] considered feedback-aided user selection that can maximize the sum-rate as well as increase a fairness criterion that ensures that each user receives their requested file in a timely manner. In the related context of the erasure BC where users have uneven probabilities of erasures, references [17] and [18] showed how an erasure at some users can be exploited as side information at the remaining users in order to increase system performance. Related work can also be found in [19]–[21].

The uneven-capacity bottleneck was also studied in the presence of multiple transmit antennas [15], [29]. Reference [15] exploited transmit diversity to ameliorate the impact of the worst-user capacity, and showed that employing $\mathcal{O}(\ln K)$ transmit antennas can allow for a transmission sum-rate that scales with $K$. Similarly, the work in [29] considered multiple transmit and multiple receive antennas, and designed topology-dependent cache-placement to ameliorate the worst-user effect.

In a related line of work, the papers [22] and [23] studied the cache-aided topological interference channel where $K$ cache-aided transmitters are connected to $K$ cache-aided receivers, and each transmitter is connected to one receiver via a direct "strong" link and to each of the other receivers via "weak" links. Under the assumption of no channel state information at the transmitters (CSIT), the authors showed how the lack of CSIT can be ameliorated by exploiting the topology of the channel and the multicast nature of the transmissions.

Recently, significant effort has been made toward understanding the behavior of coded caching in the finite Signal-to-Noise Ratio (SNR) regime with realistic (and thus often uneven) channel qualities. In this direction, the work in [24] showed that a single-stream coded caching message beamformed by an appropriate transmit vector can outperform some existing multi-stream coded caching methods in the low-SNR regime, while references [25], [26] (see also [27]) revealed the importance of jointly considering caching with multicast beamformer design. Moreover, the work in [28] studied the connection between rate and subpacketization in the multi-antenna environment, accounting for the unevenness naturally brought about by fading.

Our work is in the spirit of all the above papers, and it can be seen specifically as an extension of [14]. This reference considered a specific binary topological case, for which it proposed a two-level superposition-based transmission scheme to alleviate the worst-user bottleneck. Further, a similar approach has been proposed in the work of [30], where though the closely related scheme places focus on minimizing the power.

### B. *Overview of Results*

In this paper, we study a cache-aided SISO BC where each receiver $k$ experiences a link of some normalized capacity $\alpha_k \in [0,1]$. We establish the optimal worst-case delivery time $T(K,\gamma,\{\alpha_k\})$ within a factor of at most 4 for any number of $K$ users, fractional cache capacity $\gamma$, and capacity set $\{\alpha_k\}$. Key to this result is a new algorithm that uses superposition coding, where (assuming without loss of generality that the users are labeled from weaker to stronger, i.e., such that $\alpha_k \leq \alpha_{k+1}$) we split the power into $K - K\gamma - 1$ layers, and in layer $k$, we transmit *only* XORs whose weakest user is user $k$. While this design indeed encodes some XORs at lower rates (matching the capacity of the worst user for that message), it also allows the simultaneous transmission of other XORs in the remaining power layers. The main result reveals that the optimal performance (1) achievable when $\alpha_k = 1$, for all $k \in [K] \triangleq [1,2,\dots,K]$, is in fact achievable even if each user $k$ has reduced link capacity such that the condition

$$\alpha_k \gtrsim 1 - e^{-k\gamma}, \quad \forall k \in [K] \tag{4}$$

is satisfied. This quantifies the intuitive fact that systems with smaller caches can be better immune to the negative effects of channel unevenness.

## II. System Model

We consider the $K$-user wireless SISO BC, with the transmitter having access to a library of $N$ files $\{W^n\}_{n=1}^N$, each of normalized unit size, and the $K$ receivers having a cache whose size is equal to a fraction $\gamma \in [0,1]$ of the library size. Communication takes place in two distinct phases, namely the pre-fetching and the delivery phases. In the first phase, the caches of the users are filled with content from the library without any knowledge of future requests or of channel capacities. Then, during the delivery phase, each user $k$ requests[1] a single file $W^{d_k}$, after which the transmitter – with knowledge of the requests and the link capacities – delivers the requested content.

After transmission, at each user $k \in [K]$, the received signal takes the form

$$y_k = \sqrt{P^{\alpha_k}} h_k x + z_k, \tag{5}$$

where $P$ represents the transmitting power; $x \in \mathbb{C}$ is the power-normalized transmitted signal satisfying $\mathbb{E}\{|x|^2\} \leq 1$; $h_k \in \mathbb{C}$ is the channel coefficient of user $k$; $z_k \sim \mathbb{C}\mathcal{N}(0,1)$ represents the Gaussian noise at user $k$; and $\alpha_k \in (0,1]$ is such that at each user $k \in [K]$ the average SNR equals

$$\mathbb{E}\{|y_k|^2\} = P^{\alpha_k}. \tag{6}$$

Under the simplified Generalized Degrees of Freedom (GDoF) framework of [31]–[33], condition (6) amounts to a (normalized, by a factor $\log P$) user rate of $r_k = \alpha_k \in [0,1]$. Without loss of generality, $\alpha = 1$ corresponds to the highest possible channel strength. We assume an arbitrary set of such

[1] We are interested in the worse-case delivery time and thus we will assume that each user will ask for a different file.

normalized capacities $\boldsymbol{\alpha} \triangleq \{\alpha_k\}_{k=1}^K$ and we assume them without loss of generality to be ordered in ascending order ($\alpha_k \leq \alpha_{k+1}$).

The objective is to design the caching and communication scheme that minimizes the worst-case delivery time $T(K, \gamma, \boldsymbol{\alpha})$ for any capacity vector $\boldsymbol{\alpha}$.

## III. Main Results

Before presenting the main results, we remind the reader that the naive implementation of coded caching which sequentially transmits the sequence of XORs $X_\sigma$ to all subsets $\sigma \in [K]$ of $|\sigma| = K\gamma+1$ users, requires a worst-case delivery time

$$T_{uc}(K, \gamma, \boldsymbol{\alpha}) = \frac{1}{\binom{K}{K\gamma}} \sum_{\sigma \subseteq [K],\ |\sigma| = K\gamma+1} \max_{i \in \sigma} \left\{ \frac{1}{\alpha_i} \right\}. \tag{7}$$

This follows since this conventional uncoded scheme allocates, for each XOR $X_\sigma$, a transmission time $T_\sigma = \max_{w \in \sigma} \left\{ \frac{1}{\alpha_w} \right\}$ to allow the weakest user in $\sigma$ to decode the message[2].

We now proceed with the main result.

**Theorem 1.** *In the $K$-user SISO BC with receiver channel strengths $\{\alpha_k\}_{k=1}^K (\alpha_k \leq \alpha_{k+1})$ and with receivers equipped with a cache of normalized size $\gamma$, the worst-case delivery time*

$$T_{sc}(K, \gamma, \boldsymbol{\alpha}) = \max_{w \in [K]} \left\{ \frac{1}{\alpha_w} \cdot \frac{\binom{K}{K\gamma+1} - \binom{K-w}{K\gamma+1}}{\binom{K}{K\gamma}} \right\} \tag{8}$$

*is achievable and is within a multiplicative factor of at most 4 from the optimal delay $T^*(K, \gamma, \boldsymbol{\alpha})$.*

*Proof.* The achievability part of the scheme is described as Algorithm 1 in Section IV, while the converse and the derivation of the gap to optimal are presented in Section V. □

One of the main conclusions from the above result is summarized in the following corollary.

**Corollary 1.** *In the same $K$-user SISO BC with $\gamma$-sized caches and (ordered) capacities $\{\alpha_k\}_{k=1}^K$, the baseline performance*

$$T(K, \gamma, \boldsymbol{\alpha} = \mathbf{1}) = T_{MN} = \frac{K(1-\gamma)}{1+K\gamma} \tag{9}$$

*associated to the ideal case $\alpha_k = 1\ \forall k \in [K]$, can be achieved even if the capacities satisfy the inequalities*

$$\alpha_k \geq \alpha_{th,k} \triangleq 1 - \frac{\binom{K-k}{K\gamma+1}}{\binom{K}{K\gamma+1}} \approx 1 - e^{-k\gamma}, \quad \textit{for all } k \in [K]. \tag{10}$$

*Proof.* The proof is direct from Eq. (8), after using the Sterling approximation $\binom{n}{k} \approx \left(\frac{n}{k}\right)^k$ and the limit

$$\lim_{K \to \infty} \left(1 - \frac{b}{K}\right)^K = e^{-b}. \tag{11}$$

□

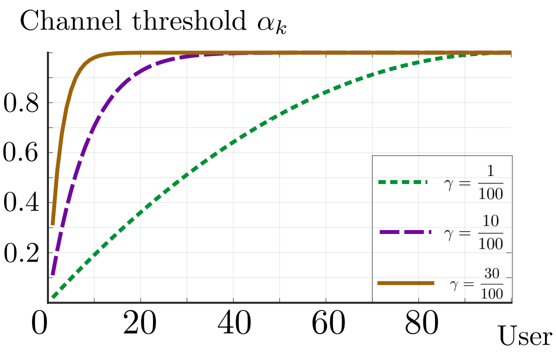

Fig. 1. The plot presents the threshold $\alpha_{th,k}$ for the case of $K = 100$ users. We can see that as $\gamma$ decreases, an ever increasing fraction of users can have a further reduced channel capacity without any performance degradation with respect to the maximal-capacity delay.

Given any user $k$, $\alpha_{th,k} = 1 - \frac{\binom{K-k}{K\gamma+1}}{\binom{K}{K\gamma+1}}$ provides a threshold channel capacity that allows the algorithm to achieve the baseline unit-capacity performance $T_{MN}$.

## IV. Placement and Delivery Algorithms

We here present the superposition-based communication scheme with the corresponding cache placement, transmission, and decoding that achieves the delay in Theorem 1.

### A. Cache Placement

During the placement phase, we apply directly the placement algorithm of [1] without exploiting any knowledge of the channel capacities. To this end, each file $W^n, n \in [N]$, is subpacketized into $S = \binom{K}{K\gamma}$ subfiles

$$W^n \to \{W_\tau^n, \tau \subset [K],\ |\tau| = K\gamma\} \tag{12}$$

and the cache $\mathcal{Z}_k$ of user $k$ is filled as

$$\mathcal{Z}_k = \{W_\tau^n : \tau \subset [K], \forall n \in [N]\} \tag{13}$$

which, as can easily be shown, adheres to the cache-size constraint.

### B. Delivery Algorithm

After each user $k \in [K]$ requests a file $W^{d_k}$ as in [1], the transmitter delivers the $\binom{K}{K\gamma+1}$ XORs

$$X_\sigma = \bigoplus_{k \in \sigma} W_{\sigma \setminus \{k\}}^{d_k} \tag{14}$$

for all subsets $\sigma$ of users of size $|\sigma| = K\gamma+1$. To this end, in every communication slot, we split the available transmission power into $K - K\gamma - 1$ "power layers". In power layer $k$ we encode XORs from the set

$$\mathcal{X}_k \triangleq \{X_\sigma : \min\{\sigma\} = k\}. \tag{15}$$

This contains all the XORs intended for set of users $\sigma$ for which the slowest user is user $k$ i.e., all the XORs intended for user $k$ except those desired by any user whose channel is

[2] This is a well known expression that has been calculated in a variety of works such as in [13], [24].

weaker than user $k$. It can be easily shown[3] that the sets $\mathcal{X}_k$ are disjoint; that for any $k \leq K - K\gamma - 1$, we have

$$|\mathcal{X}_k| = \binom{K-k+1}{K\gamma+1} - \binom{K-k}{K\gamma+1} = \binom{K-k}{K\gamma} \tag{16}$$

XOR messages in power layer $k$ and that the total number of XOR messages in the first $k$ power layers is

$$\left|\bigcup_{m=1}^{k} \mathcal{X}_m\right| = \binom{K}{K\gamma+1} - \binom{K-k}{K\gamma+1}. \tag{17}$$

For example, Layer 1 (which will correspond to the highest-powered layer) contains all the XORs in set $\mathcal{X}_1$ i.e., all the XORs that are intended for the weakest user (user 1). Similarly Layer 2 will contain the XORs from $\mathcal{X}_2$, i.e., those XORs that are intended for user 2, but not for user 1, and so on. The power allocation for each XOR is designed so that the weakest user of the XOR can decode it, implying that any other user that needs to decode that same XOR is able to do so. The chosen power allocation seeks to minimize the overall delay.

**Algorithm 1:** Delivery based on Superposition Coding

```
1 Let $\alpha_k \leq \alpha_{k+1},\ \forall k \in [K]$
2 Find $w \in [K]$ such that
      $w = \arg\max_{k\in[K]} \left\{ \frac{\binom{K}{K\gamma+1} - \binom{K-k}{K\gamma+1}}{\alpha_k} \right\}.$   (18)
3 Set $\beta_0 = 0$ and for $k \in [K - K\gamma - 1]$ set
      $\beta_k = \frac{|\cup_{i=1}^{k} \mathcal{X}_k|}{|\cup_{i=1}^{w} \mathcal{X}_k|}\alpha_w = \frac{\binom{K}{K\gamma+1} - \binom{K-k}{K\gamma+1}}{\binom{K}{K\gamma+1} - \binom{K-w}{K\gamma+1}}\alpha_w.$   (19)
  for all $k \in [K - K\gamma - 1]$ do
4     Encode $x_k$ selected from $\mathcal{X}_k$ without replacement
5     with power
          $P_k = P^{-\beta_{k-1}} - P^{-\beta_k}$   (20)
6     and rate
          $r_k = \beta_k - \beta_{k-1} = \frac{\binom{K-k}{K\gamma}}{\binom{K}{K\gamma+1} - \binom{K-w}{K\gamma+1}}\alpha_w.$   (21)
7 Transmit $x_k, \forall k \in [K]$ simultaneously.
```

The process is described in the form of pseudo-code in Algorithm 1. The algorithm begins by identifying (Step 2) the bottleneck user

$$w = \arg\max_{k\in[K]} \left\{ \frac{\binom{K}{K\gamma+1} - \binom{K-k}{K\gamma+1}}{\alpha_k} \right\}. \tag{22}$$

This is defined as the user $k$ that takes the longest time to decode all power layers from 1 to $k$. Then Step 3 calculates the power layer coefficients $\beta_i, i \in \{0, 1, .., K - K\gamma - 1\}$ for each power layer as explained below. In Step 4, for every $k \in [K - K\gamma - 1]$, a new XOR is selected from set $\mathcal{X}_k$, and is encoded in message $x_k$, with power $P_k = P^{-\beta_{k-1}} - P^{-\beta_k}$ (Step 5) and rate $\frac{\binom{K-k}{K\gamma}}{\binom{K}{K\gamma+1} - \binom{K-w}{K\gamma+1}}\alpha_w$ (Step 6). Finally in Step 7 all the $x_k, \forall k \in [K]$ are transmitted simultaneously using superposition coding.

### C. Decoding

In the received signal

$$y_k = h_k\sqrt{P^{\alpha_k}} \sum_{m_1=1}^{k} x_{m_1} + h_k\sqrt{P^{\alpha_k}} \sum_{m_2=k+1}^{K-K\gamma-1} x_{m_2} \tag{23}$$

at user $k \in [K]$, the second term $\sum_{m_2=k+1}^{K-K\gamma-1} x_{m_2}$ contains the lower power layers, which carry no valuable information for user $k$ and are treated as noise. This part of the message is transmitted with power $P^{-\beta_k}$, where $\beta_k = \frac{\binom{K}{K\gamma+1} - \binom{K-k}{K\gamma+1}}{\binom{K}{K\gamma+1} - \binom{K-w}{K\gamma+1}}\alpha_w$. Due to the power and rate allocation for each of these messages (cf. Eq. (20) and Eq. (21)), using successive interference cancellation[4] (SIC), receiver $k$ can decode the first term that encodes the messages that potentially contain information that is valuable for user $k$.

### D. Delay Calculation

The total delay of the scheme is

$$T_{sc}(K, \gamma, \boldsymbol{\alpha}) = \max_{k\in[K-K\gamma-1]} \left\{ \frac{|\mathcal{X}_k|}{\binom{K}{K\gamma}} \cdot \frac{1}{r_k} \right\} \tag{24}$$

$$= \frac{1}{\alpha_w} \cdot \frac{\binom{K}{K\gamma+1} - \binom{K-w}{K\gamma+1}}{\binom{K}{K\gamma}}. \tag{25}$$

This corresponds to the maximum delay required to deliver all XORs $X_\sigma \in \mathcal{X}_k$ across all values of $k \in [K - K\gamma - 1]$.

## V. Converse and Gap to Optimality

In this section, we provide a lower bound on the optimal delay for any given set of parameters $K, \gamma, \boldsymbol{\alpha}$, and then we prove that the achievable delay $T_{sc} \triangleq T_{sc}(K, \gamma, \boldsymbol{\alpha})$ from Theorem 1 is within a factor of at most 4 from the optimal delay $T^*(k, \gamma, \boldsymbol{\alpha})$.

To lower bound the minimum delay $T^*(k, \gamma, \boldsymbol{\alpha})$, we consider an augmented system where the capacities of the first $w$ users, with $w$ selected as (18), are increased to $\alpha_k = \alpha_w \triangleq \alpha$, for all $k \in [w]$, while the capacities of the remaining users are increased to 1. For such a system, the delay is lower bounded as

$$T_{\text{aug}} \geq \overbrace{\frac{1}{\alpha}}^{t_1} \overbrace{\frac{1}{2}\frac{w(1-\gamma)}{1+w\gamma}}^{t_2}, \tag{26}$$

[3] The last equality follows directly from Pascal's triangle.

[4] In successive interference cancellation, a user first decodes the highest powered message by treating the remaining messages as noise, then proceeds to remove this – known at this point – message and decodes the second message by treating the remaining as noise, and so on until all messages have been decoded.

where term $t_1$ corresponds to the channel capacity of the first $w$ users, while term $t_2$ corresponds to a lower bound on the minimum possible worst-case delivery time[5] associated to a system with $w$ cache-aided users (cf. [36]).

To bound the ratio $T_{sc}/T_{\text{aug}}$, we first consider the case of $w\gamma < 1$ for which we have the inequalities

$$\frac{T_{sc}}{T_{\text{aug}}} \leq \frac{\frac{\binom{K}{K\gamma+1}-\binom{K-w}{K\gamma+1}}{\binom{K}{K\gamma}}}{\frac{1}{2}\frac{w(1-\gamma)}{(1+w\gamma)}} \leq \frac{w(1-\gamma)}{\frac{1}{2}\frac{w(1-\gamma)}{(1+w\gamma)}} \leq 4, \tag{27}$$

where we used the inequality $\frac{\binom{K}{K\gamma+1}-\binom{K-w}{K\gamma+1}}{\binom{K}{K\gamma}} \leq w(1-\gamma)$ which we prove in Appendix A.

When $w\gamma \geq 1$, the bound – after a few basic algebraic manipulations – takes the form

$$\frac{T_{sc}}{T_e} = \frac{\frac{\binom{K}{K\gamma+1}-\binom{K-w}{K\gamma+1}}{\binom{K}{K\gamma}}}{\frac{1}{2}\frac{w(1-\gamma)}{(1+w\gamma)}} \leq \frac{\frac{\binom{K}{K\gamma+1}}{\binom{K}{K\gamma}}}{\frac{1}{2}\frac{w(1-\gamma)}{(1+w\gamma)}} \tag{28}$$

$$= \frac{\frac{K(1-\gamma)}{1+K\gamma}}{\frac{1}{2}\frac{w(1-\gamma)}{1+w\gamma}} = 2\frac{K(1+w\gamma)}{w(1+K\gamma)} = 2+2\frac{K-w}{w+Kw\gamma} \tag{29}$$

$$< 2\left(1+\frac{K+w}{w+wK\gamma}\right) < 2\left(1+\frac{K}{wK\gamma}\right) \leq 4, \tag{30}$$

which concludes the proof.

## VI. Conclusions and Ramifications

In this work, we studied a cache-aided SISO BC in which users have different channel capacities. This model is motivated by the well-known worst-user bottleneck of coded caching, which, when left untreated, can severely deteriorate coded caching gains. The new algorithm establishes, together with the converse, the fundamental limits of performance within a factor of 4, revealing that it is in fact possible to achieve the full-capacity performance even in the presence of many users with degraded link strengths.

Pivotal to our approach is the identification of a 'bottleneck (threshold) user', which may not necessarily be the user with the worst channel. From an operational point of view, this reveals that to increase performance, we must not necessarily focus on enhancing only the weakest users, but rather should focus on altering this bottleneck threshold.

## Appendix A
## Bound on the difference of Binomials

Our aim is to prove the following corollary.

**Corollary 2.** *For every integer $m > 0$, the following inequality holds*

$$\frac{\binom{K}{K\gamma+1}-\binom{K-m}{K\gamma+1}}{\binom{K}{K\gamma}} \leq m(1-\gamma). \tag{31}$$

*Proof.* We first note that

$$\frac{\binom{K-n-1}{K\gamma}}{\binom{K}{K\gamma}} \leq (1-\gamma), \quad n \geq 0 \tag{32}$$

holds, because for any $m > p$, we have that $\binom{K-m}{K\gamma} < \binom{K-p}{K\gamma}$, which yields that Eq. (32) holds for any $n \geq 0$.

With this in place, in order to prove the inequality in Eq. (31), we employ proof by induction. Toward this, we first see that Eq. (31) holds for $m = 1$ because

$$\frac{\binom{K}{K\gamma+1}-\binom{K-1}{K\gamma+1}}{\binom{K}{K\gamma}} = \frac{\binom{K}{K\gamma+1}-\frac{K-K\gamma-1}{K}\binom{K}{K-K\gamma-1}}{\binom{K}{K\gamma}} \tag{33}$$

$$= \frac{\binom{K}{K\gamma+1}}{\binom{K}{K\gamma}}\frac{K\gamma+1}{K} = (1-\gamma). \tag{34}$$

Now we assume that Eq. (31) holds for some $n \geq 1$, and to prove that it also holds for $n+1$, we see that

$$\frac{\binom{K}{K\gamma+1}-\binom{K-n-1}{K\gamma+1}}{\binom{K}{K\gamma}} = \frac{\binom{K}{K\gamma+1}-\binom{K-n}{K\gamma+1}+\binom{K-n-1}{K\gamma}}{\binom{K}{K\gamma}} \tag{35}$$

$$= \frac{\binom{K}{K\gamma+1}-\binom{K-n}{K\gamma+1}}{\binom{K}{K\gamma}} + \frac{\binom{K-n-1}{K\gamma}}{\binom{K}{K\gamma}} \tag{36}$$

$$\leq n(1-\gamma)+(1-\gamma) \tag{37}$$

where in Eq. (35) we used the equality from Pascal's triangle, and then in Eq. (37) we used the inequality of Eq. (31). This concludes the proof. □

## References

[1] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. on Inf. Theory*, vol. 60, pp. 2856–2867, May 2014.

[2] K. Shanmugam, M. Ji, A. M. Tulino, J. Llorca, and A. G. Dimakis, "Finite-length analysis of caching-aided coded multicasting," *IEEE Trans. on Inf. Theory*, vol. 62, no. 10, pp. 5524–5537, 2016.

[3] E. Lampiris and P. Elia, "Adding transmitters dramatically boosts coded-caching gains for finite file sizes," *IEEE Journal on Selected Areas in Communication (JSAC), Special Issue on Caching*, June 2018.

[4] Q. Yan, M. Cheng, X. Tang, and Q. Chen, "On the placement delivery array design for centralized coded caching scheme," *IEEE Transactions on Information Theory*, vol. 63, pp. 5821–5833, Sep. 2017.

[5] L. Tang and A. Ramamoorthy, "Coded caching schemes with reduced subpacketization from linear block codes," *IEEE Transactions on Information Theory*, vol. 64, pp. 3099–3120, April 2018.

[6] C. Shangguan, Y. Zhang, and G. Ge, "Centralized coded caching schemes: A hypergraph theoretical approach," *IEEE Transactions on Information Theory*, vol. 64, pp. 5755–5766, Aug 2018.

[7] H. H. S. C and P. Krishnan, "Low subpacketization coded caching via projective geometry for broadcast and D2D networks," *CoRR*, vol. abs/1902.08041, 2019.

[8] X. Zhang and M. Ji, "A new design framework on device-to-device coded caching with optimal rate and significantly less subpacketizations," *CoRR*, vol. abs/1901.07057, 2019.

[9] A. M. Ibrahim, A. A. Zewail, and A. Yener, "Coded caching for heterogeneous systems: An optimization perspective," *IEEE Transactions on Communications*, pp. 1–1, 2019.

[10] B. Asadi, L. Ong, and S. J. Johnson, "Centralized caching with unequal cache sizes," in *IEEE Inf. Theory Workshop (ITW)*, Nov 2018.

[11] M. Mohammadi Amiri, Q. Yang, and D. Gündüz, "Decentralized caching and coded delivery with distinct cache capacities," *IEEE Transactions on Communications*, vol. 65, pp. 4657–4669, Nov 2017.

[12] E. Lampiris and P. Elia, "Full coded caching gains for cache-less users," *IEEE Information Theory Workshop (ITW)*, 2018.

[5] In fact, as we know from [36], this factor is slightly smaller than $\frac{1}{2}$.

[13] L. Zheng, Z. Wang, Q. Yan, Q. Chen, and X. Tang, “On the coded caching based wireless video transmission scheme,” in *IEEE/CIC Inter. Conf. on Comm. in China (ICCC)*, pp. 1–6, July 2016.

[14] J. Zhang and P. Elia, “Wireless coded caching: A topological perspective,” in *IEEE Int. Symp. on Inf. Theory (ISIT)*, June 2017.

[15] K. Ngo, S. Yang, and M. Kobayashi, “Scalable content delivery with coded caching in multi-antenna fading channels,” *IEEE Transactions on Wireless Communications*, vol. 17, pp. 548–562, Jan 2018.

[16] A. Destounis, M. Kobayashi, G. Paschos, and A. Ghorbel, “Alpha fair coded caching,” in *15th International Symp. on Modeling and Opt. in Mobile, Ad Hoc, and Wireless Networkss (WiOpt)*, pp. 1–8, May 2017.

[17] A. Ghorbel, M. Kobayashi, and S. Yang, “Content delivery in erasure broadcast channels with cache and feedback,” *IEEE Transactions on Information Theory*, vol. 62, pp. 6407–6422, Nov 2016.

[18] M. Mohammadi Amiri and D. Gündüz, “Cache-aided content delivery over erasure broadcast channels,” *IEEE Transactions on Communications*, vol. 66, pp. 370–381, Jan 2018.

[19] S. Kamel, M. Sarkiss, and M. Wigger, “Decentralized joint cache-channel coding over erasure broadcast channels,” in *IEEE Middle East and North Africa Comm. Conf. (MENACOMM)*, pp. 1–6, April 2018.

[20] S. Kim, S. Mohajer, and C. Suh, “Coding across heterogeneous parallel erasure broadcast channels is useful,” in *IEEE International Symposium on Information Theory (ISIT)*, pp. 1883–1887, June 2017.

[21] S. Saeedi Bidokhti, M. Wigger, and R. Timo, “Noisy broadcast networks with receiver caching,” *IEEE Transactions on Information Theory*, vol. 64, pp. 6996–7016, Nov 2018.

[22] E. Lampiris, J. Zhang, and P. Elia, “Cache-aided cooperation with no CSIT,” in *IEEE Int. Symp. on Inf. Theory (ISIT)*, June 2017.

[23] E. Piovano, H. Joudeh, and B. Clerckx, “Generalized degrees of freedom of the symmetric cache-aided MISO broadcast channel with partial CSIT,” *IEEE Transactions on Information Theory*, pp. 1–1, 2019.

[24] S. P. Shariatpanahi, G. Caire, and B. Hossein Khalaj, “Physical-layer schemes for wireless coded caching,” *IEEE Transactions on Information Theory*, vol. 65, pp. 2792–2807, May 2019.

[25] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. Khalaj, “Multi-antenna interference management for coded caching,” *arXiv preprint arXiv:1711.03364*, 2017.

[26] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. Khalaj, “Multicast beamformer design for coded caching,” in *IEEE International Symposium on Information Theory (ISIT)*, pp. 1914–1918, June 2018.

[27] J. Zhao, M. M. Amiri, and D. Gündüz, “A low-complexity cache-aided multi-antenna content delivery scheme,” in *IEEE Int. Workshop on Signal Processing Advances in Wireless Comm. (SPAWC)*, July 2019.

[28] M. Salehi, A. Tölli, S. P. Shariatpanahi, and J. Kaleva, “Subpacketization-rate trade-off in multi-antenna coded caching,” *arXiv preprint arXiv:1905.04349*, 2019.

[29] I. Bergel and S. Mohajer, “Cache-aided communications with multiple antennas at finite SNR,” *IEEE Journal on Selected Areas in Communications*, vol. 36, pp. 1682–1691, Aug 2018.

[30] M. M. Amiri and D. Gündüz, “Caching and coded delivery over gaussian broadcast channels for energy efficiency,” *IEEE Journal on Selected Areas in Communications*, vol. 36, pp. 1706–1720, Aug 2018.

[31] S. A. Jafar and S. Vishwanath, “Generalized Degrees of Freedom of the symmetric Gaussian $K$ user Interference Channel,” *IEEE Transactions on Information Theory*, vol. 56, pp. 3297–3303, July 2010.

[32] A. Gholami Davoodi and S. A. Jafar, “Aligned image sets under channel uncertainty: Settling conjectures on the collapse of degrees of freedom under finite precision CSIT,” *IEEE Transactions on Information Theory*, vol. 62, pp. 5603–5618, Oct 2016.

[33] A. Gholami Davoodi and S. A. Jafar, “Generalized degrees of freedom of the symmetric $K$ user interference channel under finite precision CSIT,” *IEEE Trans. Inf. Theory*, vol. 63, pp. 6561–6572, Oct 2017.

[34] E. Lampiris and P. Elia, “Bridging two extremes: Multi-antenna coded caching with reduced subpacketization and CSIT,” in *IEEE Int. Workshop on Signal Processing Advances in Wireless Comm. (SPAWC)*, 2019.

[35] E. Lampiris and P. Elia, “Achieving full multiplexing and unbounded caching gains with bounded feedback resources,” *IEEE International Symposium on Information Theory (ISIT)*, 2018.

[36] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, “Characterizing the rate-memory tradeoff in cache networks within a factor of 2,” *IEEE Transactions on Information Theory*, vol. 65, pp. 647–663, Jan 2019.