Queueing Theoretic Models for Multiuser MISO Content-Centric Networks With SDMA, NOMA, OMA and Rate-Splitting Downlink

Ramkumar Raghu, Mahadesh Panju, and Vinod Sharma, Senior Member, IEEE

Abstract—Multiuser, Multiple Input, Single Output (MU-MISO) systems are indispensable in next-generation wireless networks such as 5G and 6G. The spatial diversity of MISO has been leveraged in these wireless networks to improve the capacity. Furthermore, several recent studies have utilised redundancies in the content request along with this spatial diversity for further capacity improvements. In such systems, the MISO Max-Min Fair (MMF) Beamforming schemes based on SDMA, NOMA, OMA and Rate-Splitting are shown to improve the delivery rates. However, in these studies the key aspects such as the queueing delays in the downlink and the user dynamics have generally been ignored. In this work, we study how queueing, beamforming and user dynamics affect the Quality-of-Service (user delay) of downlink in MU-MISO content centric networks (CCNs). We also show that a Simple Multicast Queue (SMQ) developed for SISO, when adapted to MU-MISO provides good performance due to its inherent stability. However, we observe that MMF Beamforming coupled with SMQ can be unfair to users with good channels. Thus, we propose an alternative scheme, called Dual-SMQ to address this problem. We also provide theoretical approximations for the mean delay experienced by the users in such systems.

Index Terms—MISO, scheduling, multigroup-multicasting, multiplexing, quality of service, queueing delay, SDMA, NOMA, rate splitting, max-min fairness.

I. INTRODUCTION

Proliferation of the Over-the-Top (OTT) platforms such as Netflix, Amazon Prime Video, YouTube etc., has increased the demand for high definition (HD) video contents over wireless networks [1]. It is seen that, depending on the geographic location, a finite subset of the contents hosted by these platforms can sometimes become highly popular [2]. The wireless base stations (BS) receive repeated requests for these contents from different users in a localised geographic location [2]. Such characteristics of request traffic in the current generation wireless networks have spawned an array of studies focused on Content Centric Networks (CCNs), to leverage such redundancies and improve the network capacity. Further, the broadcast nature of wireless transmitters has been exploited by many of these studies, to provide improvements to both the physical layer and the network layer, by serving the redundant requests in a multigroup-multicast manner.

In physical layer, multiple antenna, MISO (multiple input, single output) based designs [3], [4], [5], [6], [7], [8] have gained traction due to increased delivery rates they provide in multigroup, multicast transmissions. These schemes utilise high degrees of spatial diversity provided by multiple antennas. The diversity is two fold: First the diversity comes from the broadcast nature of the wireless channel. Multiple antennas in MISO systems accentuate this advantage by enabling simultaneous multicasting of different contents. The second diversity comes from the enhanced spatial multiplexing capabilities provided by the multiple antennas. In other words, more than one content can be delivered to the same user (e.g., small base station) simultaneously. Naturally, MISO (and MIMO) has become the cornerstone of 5G, 6G and eMBB (Enhanced Mobile Broadband) networks.

However many of these works study the physical layer performance in MISO CCNs in isolation. There are other aspects of networks which affect the performance/QoS in a wireless networks such as queueing, user dynamics etc. The importance of cross-layer design with queues in a multiuser MIMO network and the effect of user dynamics was identified in early works, [9]. However, cross-layer studies of MISO networks, particularly with queues, is seemingly limited, even today. In this paper, we take a holistic approach to the analysis of multigroup-multicasting in a MISO network. We study the interplay between queueing, multiantenna beamforming and user dynamics and the effect it has on the network performance. Unlike, the physical layer studies which study the delivery rate, we concentrate on the user experienced QoS, namely the user’s mean delay (sojourn time). In a practical system we will see that the queueing design plays an important role in reducing the delay, as queueing is what controls the scheduling of service transmissions, to a time-varying set of users and their requests.

In the following we provide the relevant literature survey. In [3], the authors show that in a MISO network, that the sum rate maximization, has no QoS or Fairness guarantees.
We study various beamforming schemes based on SDMA, OMA, NOMA and Rate Splitting (RS or RSMA) from a cross-layer, queueing perspective. We provide necessary adaptations to these beamforming schemes, required for queueing and transmission to time-varying and common users across different multicast groups. We show that the performance of RS differs significantly from [4] and [5] in such a setup. To the best of our knowledge, this is the first work, which studies the effects of queueing on Rate-Splitting Multiple Access.
We show that, in a heterogenous network, MMF beamforming schemes can be pretty unfair to the users with good channel statistics. We propose a novel two queue architecture, named Dual Simple Multicast Queue (DSMQ), to address this issue. We show that DSMQ provides a decoupled, fair QoS to users with good channels, in a network with heterogeneous channels. The users with bad channel gains also may not suffer much.

We provide theoretical analysis and insights to the proposed queues.

- First, we prove the stationarity of SMQ and DSMQ and show that they are always stable even in MISO setup.
- Second, we derive queueing theoretical approximations to the user experienced delay (mean sojourn time) in a Multiuser MISO CCNs with queues and MMF beamforming.

Finally, we present extensive performance study via simulations, for the queueing and MMF beamforming schemes considered in this paper. In the extended version of this paper [22], we also show that for Multiuser MISO CCNs, the schemes developed for SISO systems such as Loopback, Defer [18] and [23] and reinforcement learning based Power control in time [20], [21] are not useful. This is because, the time diversity utilised by these schemes in a SISO system is compensated by the spatial diversity of the MISO system.

Rest of the paper is organised as follows. Section II describes the system model, assumptions and SMQ. Section III presents the beamforming schemes considered in this paper. Section IV describes a new type of queue called Dual Simple Multicast Queue (DSMQ) which improves over SMQ in terms of fairness. Section V provides proof of stationarity for SMQ and DSMQ. In Section VI we present the theoretical approximations for mean delay (sojourn time) for SMQ and DSMQ in the MISO system. Section VII provides simulation results and comparisons different schemes. Finally, Section VIII concludes the paper.

**Notation:** \{ \}^H, \{ \}^T represent Hermitian and Transpose operations respectively, \([ N] \) represents the set of natural numbers upto \( N \), \| \|^2 represents \( L_2 \) norm, \text{diag}(g) \) represent a diagonal matrix formed by elements of vector \( g \).

**II. System Model**

We consider a wireless content-centric network with one BS endowed with \( L \) transmit antennas and \( K \) user equipments (UE). The system model is shown in Fig. 1. Each UE requests contents from a library of \( N \) files. Each file is of size \( F \) bits and the receiver bandwidth is, \( B \) MHz. In practical networks the UEs can be either a mobile user or a small BS (SBS). The channel between transmit antennas and a UE follows flat fading. In other words the channel stays constant for the duration of each file transmission and independently changes in the next transmission. The request traffic from each user follows Independent Reference Model (IRM). In IRM, the request process of each content \( n \in [N] \) from each UE \( k \), is an independent Poisson process with rate \( \lambda_{nk} \). The overall rate of request traffic to the BS is given as \( \lambda = \sum_{n,k} \lambda_{nk} \). These requests are queued and served by the BS using SMQ, [18].

**Simple Multicast Queue (SMQ):** In this queue, the files are served in a First-Come-First-Serve fashion with a slight modification: when a file is transmitted, it serves all the users who have requests for it in the queue. At a given instance the \( j^{th} \) entry of the queue is denoted by \( (n, L_n) \), where \( n \in [N] \) is the file index, \( L_n \neq \emptyset \) is the list of users requesting file \( n \). When a new request for a file \( n \) from user \( k \) arrives at SMQ, it is merged at \( j^{th} \) location as \( \mathbb{L}_n = \mathbb{L}_n \cup \{k\} \). If none of the entries of SMQ has the file \( n \) then a new entry \( (n, L_n = \{k\}) \) is added to the tail of the queue. The BS simultaneously transmits first \( \min(s, S) \) files starting from the head of the line of SMQ using one of the schemes in Section III, where \( s \) is the queue length and \( S \) is a configurable system parameter. New requests for a file under service are added to the tail of SMQ and the subsequent requests are merged as before. Thus the queue length does not exceed \( N \) at any given time. Further, the sojourn time, \( D \) of a request arriving at the BS at time \( t_r \) and serviced at time \( t_c \) is given by, \( D = t_c - t_r \). Here \( D \) includes the service time \( T^* \) (defined for each scheme in Section III). The mean sojourn time under stationarity (see Section V) is denoted as \( E[D] \).

In the following sections we describe the beamforming strategies considered in this paper. We assume complete channel state information at the transmitter (CSIT). In each channel use, the channel matrix \( \mathbf{H} \in \mathbb{C}^{L \times K} \) is drawn independently as \( \mathbf{H} \sim \mathcal{C}\mathcal{N}(0, \Sigma) \), the complex Gaussian distribution with mean 0 and covariance \( \Sigma \). We consider two cases, namely, homogeneous channels and heterogeneous channels. In homogenous channel case, \( \Sigma = g\mathbf{I} \), where \( \mathbf{I} \) is the identity matrix of size \( KL \times KL \), and \( g \) is the mean of channel gain of each channel when all the users have similar channel statistics. In a heterogenous network different users channel statistics may be different. Here, \( \Sigma = \text{diag}(g_1^T, \cdots, g_K^T) \), where, \( g_k = g_k \mathbf{I} \), \( g_k \) is the mean fading (or) channel gain of user \( k \) and \( \mathbf{I} \) is the vector of all ones of size \( L \). Further the transmitter may choose to transmit one or more files from the queue depending on the beamforming strategy. \( S \) files in the head of the line of SMQ...
are denoted by $\mathcal{X}_S \triangleq \{X_1, X_2, \ldots, X_S\}$. We assume that the transmitter uses Gaussian codebook to assign a codeword $\tilde{X}_s$ for each file $X_s$ and the codebook is known to the receiver as well (This is a widely accepted assumption in all MAC layer studies [7], [8], [11]. For adaptations to practical coding and modulation schemes, see [8], [24]). Since SMQ merges different requests across users for the same content, it is possible that a user might have requested more than one content in $\mathcal{X}_S$.

III. Beamforming Schemes

We now describe the SDMA, OMA, NOMA and RS based beamforming schemes used in this paper. We note again that the objective of the paper is to optimize the QoS (mean delay)/mean sojourn time experienced by the users of the system described in Section II. We will see in Section VI that this QoS is affected directly by the service time moments. Larger the service moments, higher the delays and similarly, as the service moments get smaller, the delays also reduce. Thus, the most natural way to reduce the service moments is to minimize the transmission times of each service. Towards this, we design beamforming schemes either to minimize the maximum of transmission times among the streams/users directly (as in Sections III-B and III-C) or to maximize the minimum rate among the streams/users (as in Section III-A). The later also reduces the service moments as the rate is inversely proportional to the transmit time. Further, for $S = 1$, schemes in Sections III-A and III-B are same and correspond to OMA. For $S \geq 2$, the scheme in Section III-A, corresponds to SDMA when there are no common users and otherwise, a combination of NOMA (for common) and SDMA (for other users). For $S \geq 2$, Section III-B gives a NOMA based scheme. Finally, Section III-C presents our RS based scheme. Thus, these beamforming schemes are designed for queueing, time varying sets of active users and also cater to common users across multiple groups.

A. Max-Min Fair (MMF) Beamforming

Let the subset $X \subset \{X_1, \ldots, X_S\}$, be the set of files requested by user $k$. In this strategy the transmitter chooses a precoding vector $\bf w_s \in \mathbb{C}^L \times 1$ for each file $s \in [S]$ from $S$ of the head line files. Thus received signal $y_k$ at user $k$ is given as

$$y_k = \sum_{s: X_s \in X} h^H_w \tilde{X}_s + \sum_{t: X_t \in X \setminus X_s} h^H_w \tilde{X}_t + n_k,$$

where $k \in \mathcal{U}$, $n_k \sim \mathcal{CN}(0, \sigma^2_k)$, $\bf w_s \in \mathbb{C}^L \times 1$, is the spatial precoding vector selected by the transmitter for file/stream $s$ and $h^H_w \in \mathbb{C}^{L \times 1}$ is the $k^{th}$ column of matrix $H$. The SINR of transmitted file $s \in X$, requested by the $k^{th}$ user is:

$$\gamma^*_k = \frac{|h^H_w w_s|^2}{\sigma^2_k + \sum_{t \in X \setminus X_s} |h^H_w w_t|^2}.$$

Let $\mathcal{U}_k$ be the set of users requesting file $s$ and let $\mathcal{U}_A$ be the set of users requesting subset $A \subset [S]$. Further let $\mathcal{A}$ be the collection of all the subsets of $A \subset [S]$ with cardinality greater than one. That is, if $|B| > 1$.

Let $R_k^s$ denote the rate of transmission/service, corresponding to user $k$ requesting file $s$. The precoding weights $w_s, s \in [S]$ are obtained by solving the following optimization problem $P1$:

$$\begin{align*}
\max_{R_k^s, w_k, s \in [S]} & \quad \min_{k \in \mathcal{U}, s \in [S]} R_k^s \\
\text{s.t.} & \quad R_k^s \leq \log_2(1 + \gamma_k^s), \forall k \in \mathcal{U}, \ s \in [S], \\
& \quad \sum_{j \in B} R_j \leq \log_2(1 + \gamma_j), \forall B \in \mathcal{A}, u \in \mathcal{U}_A, A \subset [S], \text{and} \\
& \quad \sum_{s \in [S]} ||w_s||^2_2 \leq P. \\
\end{align*}$$

The first set of inequalities are the rate constraints based on the Gaussian capacity for a single stream, considering the unwanted streams as noise. The users who want more than one file (common users), decode the required files using Successive Interference Cancellation (SIC). The second set of inequalities are for SIC MAC constraints for the common users or users who have more than one file requests in the queue. Finally, the last inequality is the total power constraint with power $P$.

Since $S$ files are being transmitted simultaneously, the total transmit time, $T^s = F/(R^s B)$, where $R^s$ is the optimal rate (in bits/sec/Hz) obtained by solving $P1$.

B. Max-Min Fair Beamforming With Full SIC (MMF-SIC)

In this strategy $S$ head of the line requests, $X_S$ are transmitted to all the users $\mathcal{U} = \bigcup_{s \in [S]} \mathcal{U}_s$. Thus SINR for stream, $s$ at user, $k$ is $\gamma_k^s = |h^H_w w_s|^2 / \sigma^2_k$. All the messages are decoded at all the users using SIC. Let $R_k^s$ be the rate allotted to stream $s$ for user $k$, $\forall s \in [S]$ and $k \in \mathcal{U}$. Since all the users receive all the files, we have the flexibility to consider $S$ files in $X_S$ as a single file of size $SF$ and rearrange the sizes of files to be transmitted in different streams as $X^s_{\beta} \triangleq \{X^s_{\beta_1}, \ldots, X^s_{\beta_S}\}$, where the file $X^s_{\beta}$ is of size $\beta_s SF$, $\beta_s \in [0,1]$, $s \in [S]$ and $\sum_{s \in [S]} \beta_s = 1$. Thus transmit time of stream $s$ to user $k$ is given by $T_k^s = \beta_s SF$. Therefore, to minimize the transmit time of $S$ files, we have the following optimization problem $P2$:

$$\begin{align*}
\min_{R_k^s, \beta_s, w_s, s \in [S]} & \quad \max_{k \in \mathcal{U}, s \in [S]} T_k^s \\
\text{s.t.} & \quad \sum_{s \in \mathcal{S}} R_k^s \leq \log_2(1 + \sum_{s \in \mathcal{S}} \gamma_k^s), \forall \mathcal{S} \subset [S], k \in \mathcal{U}, \\
& \quad \sum_{s \in [S]} \beta_s = 1, \sum_{s \in [S]} ||w_s||^2_2 \leq P. \\
\end{align*}$$

The channel from the BS to a given user $k$ forms a MAC channel with $S$ messages. The first set of constraints ensure $R_k^s$ for $s \in [S]$ and $k \in \mathcal{U}$, lie in the achievable region for every user. This ensures that every user can decode all the $S$ streams using SIC. The second equality constraint on file size fraction ensures that all the $SF$ bits are split to $S$ streams. The fraction of file size associated to stream $s$ is decided by the optimization variable $\beta_s$. This way stream with the lowest min rate will have the lowest file size for transmission and the stream with the highest rate will have the highest file size.
size, thereby making the transmission time the same across the streams. The last inequality is the total power constraint.

C. Max-Min Fair Rate Splitting (MMF-RS) Beamforming

In this section we give a new formulation of Rate Splitting (RS) proposed in [5]. In RS [5], the idea is to split a particular file into two parts, map the parts to two symbols and transmit both the parts simultaneously with different precoding weights. At the receiver the first symbol is decoded considering the other part as interference and then the decoded symbol is cancelled from the received signal and the second signal is recovered. The optimal weights are obtained by optimizing the sum rates of both the streams. While this is a good objective for physical layer, the queueing layer has to wait for transmission of both the files before the next can be served. Thus instead of maximizing the sum rate, we need to minimize the maximum transmit time for both the parts. For our MISO case, the problem is formulated as follows:

As in section III-A, we consider $S$ files at the head of the line of the queue. Each file $X_s$, $s \in S$ of size $F$ bits is split into two parts $X^0_s$ and $X^1_s$ with corresponding sizes $\alpha_s F$ and $(1 - \alpha_s)F$ correspondingly, where $\alpha_s \in [0, 1]$ is an optimization parameter for file $s$. The transmitter chooses precoders $w_s$, $s \in S$ for transmitting mapped symbols $X^1_s \rightarrow X^1_s$, $s \in S$. The parts $\{X^0_s, \ldots, X^0_S\}$ are mapped to a single symbol as $\{X^1_s, \ldots, X^0_S\} \rightarrow X^1_s$. The subscript $D$ represents degraded transmission as in [5]. The SINR for the degraded stream at user $k$ is given as:

$$\gamma^D_k = \frac{|h^H_kw_D|^2}{\sigma_k^2 + \sum_{s \in S}|h^H_kw_s|^2}. \quad (5)$$

Let $R_s$ be the rate allocated to stream $s$ and $R_D$ be the rate allocated to stream $D$. The transmitter chooses precoder $w_D$ for transmitting $X^0_D$. The transmission times of symbols $X^1_s$, $s \in S$ and $X^0_D$ are $(1 - \alpha_s)F/(R_s B^s)$, $s \in S$ and $\sum_{s \in S} \alpha_s F/(R_D B^D)$ respectively. Now we want to minimize the maximum transmission time. This leads to the following optimization problem $P3$:

$$\min_{R^D, w_D} \max_{0 \leq \alpha S \leq 1, s \in S} \left\{ \frac{\sum_{s \in S} \alpha_s F}{BR^D_k}, \frac{(1 - \alpha_s)F}{BR^s_k} \right\}$$

s.t. $R^D \leq \log_2 (1 + \frac{R^D_k}{\gamma^D_k}), \forall k \in \mathcal{U}, s \in [S]$, $\|w_D\|^2 + \sum_{s \in S} \|w_s\|^2 \leq P$,

and rest of the constraints, as in (3). \quad (6)

The constraint on $R_D$ ensures delivery of the degraded stream to all the users. Finally the last inequality is the total power constraint. In both, $P2$ and $P3$, the service time $T^*$ is the optimal value of the respective objective function.

D. Optimization Reformulation and Service Time

For a tractable queueing system, the transmitter serving multiple groups simultaneously should start the next transmission after all the transmissions are complete. Thus imposing symmetric rate across all groups (i.e., transmitting all groups with the optimal min rate), does not change the optimal transmit time obtained by solving $P1-P3$. Thus by imposing symmetric rate requirement our problem $P1$ can be reformulated as:

$$\max_{r, s, r \in [S]} \sum_{s \in [S]} R_s$$

s.t. $r \leq \log_2 (1 + \gamma_k^D), \forall k \in \mathcal{U}, s \in [S]$, $r \leq \frac{1}{|B|} \log_2(1 + \sum_{j \in B} \gamma_j^D), \forall B \in \mathcal{A}, u \in \mathcal{U}, A \subset [S]$, and $\sum_{s \in [S]} \|w_s\|^2 \leq P$. \quad (7)

The optimization $P2$ and $P3$ however require a slightly different reformulation when we impose a symmetric rate constraint. In $P2$, we impose that the time taken to transmit each stream to be the same. Towards this we consider $R_s = \min_{k \in \mathcal{U}} R^s_k$ for all $s \in [S]$. Further, for achieving same transmit time, we set the fraction $\beta_s = R^s/(\sum_{s \in [S]} R^s)$, $\forall s \in [S]$. The symmetric transmission time is thus, $T = SF/(B \sum_s R^s)$ for each stream. We define the symmetric rate of transmission as $r = 1/T$. Thus $P2$ can be reformulated as maximization of $\sum_s R_s$. This leads us to the following simplified formulation of $P2$:

$$\max_{R^s, \forall s \in [S]} \sum_{s \in [S]} R^s$$

s.t. $\sum_{s \in [S]} R^s \leq \log_2 \left( 1 + \sum_{s \in [S]} |h^H_kw_s| \right), \forall s \in [S], k \in \mathcal{U}$, and $\sum_{s \in [S]} \|w_s\|^2 \leq P$. \quad (8)

Similarly, in $P3$ we impose $r = \frac{BR^D}{\sum_{s \in [S]} \alpha_s F} = \frac{BR^D}{\sum_{s \in [S]} \alpha_s F}$ for all feasible $s, k$’s. This leads to the following reformulation of the optimization problem $P3$:

$$\max_{r, w_D, \forall s \in [S]} \sum_{s \in [S]} R^s$$

s.t. $r \sum_{s \in [S]} \alpha_s F \leq B \log_2 (1 + \frac{\gamma^D_k}{\gamma^D_k}), r(1 - \alpha_s)F \leq B \log_2 (1 + \gamma^D_k)$, $\forall k \in \mathcal{U}, s \in [S], 0 \leq \alpha_s \leq 1$, $r \sum_{j \in B} (1 - \alpha_s)F \leq B \log_2(1 + \sum_{j \in B} \gamma_j^D)$, $\forall B \in \mathcal{A}, u \in \mathcal{U}, A \subset [S]$, and $\|w_D\|^2 + \sum_{s \in [S]} \|w_s\|^2 \leq P$. \quad (9)

**Service Time:** For the above reformulations, the service time $T^*$, is given by $T^* = \frac{1}{r^*}$ for MMF (7), MMF-SIC (8) and $T^* = \frac{1}{r^*}$ for MMF-RS (9) Beamforming, where $r^*$ are the optimal values of the objectives in (7), (8) and (9).

**Solver and Complexity:** The MMF problems (like P1-P3) are known to be NP-Hard [7]. However good sub-optimal
algorithms such as SCA, SOCP are available [7]. We use the reformulations (7), (8) and (9) along with Python’s, SciPy SLSQP, to directly solve for the beamformers. We have seen that this gives better performance than SOCP in our simulations. Further, successive convex approximation methods like SLSQP are shown to have a better performance and computational complexity trade-off [25]. For a simpler case of non-overlapping groups, the per-iteration computational complexity from [25] is $O(SK^2 + S^2KL^2)$. For overlapping groups as in our setup, the complexity may be slightly higher.

IV. DUAL SIMPLE MULTICAST QUEUE (DSMQ)

Performance of SMQ in MISO with MMF beamforming schemes is severely affected by the presence of bad channel users. This is because (3), (4) and (6) maximise the min rate, which is controlled by users with bad channel statistics. Under stationarity, by PASTA [26], the users with good channel also experience the same mean sojourn times as bad users. Hence the performance of the overall system degrades. To improve the performance of good channel users while maintaining fairness to bad channel users in multi-antenna case, we modify the SMQ scheme.

We call the new kind of queue as Dual Simple Multicast Queue (DSMQ). In DSMQ, the requests from the good and the bad channel users are put in two different simple multicast queues, SMQ-G and SMQ-B, respectively. We assume that the BS keeps track of the statistics (e.g., mean/moving average of channel gains) of each user and thus can differentiate between good and bad channel users. When a user channel changes (e.g., due to movement) the BS appropriately changes the queue (SMQ-G/B) for the user (such variations are not in the scope of this work).

Further, only one queue is serviced at a time, using all the antennas. Depending on the setting the BS solves $P1$, $P2$ or $P3$ and serves users in first $S$ head of the line files of the queue. We fix a number $C$ and allow the SMQ-B to be serviced once in every $C$ channel uses or when SMQ-G is empty. If both SMQ-G and SMQ-B are empty the first arrival to the system is served. In all the other conditions, only SMQ-G is serviced. This way we decouple the QoS (delay) of good channel users from that of bad channel users.

We will see in Section VII that an appropriate choice of $C$ and beamforming strategy improves the QoS of good channel users without drastically affecting the bad channel users.

DSMQ can be generalised to the case when there are $G(\geq 2)$ groups of users with sufficiently different channel statistics (e.g., 15 dB difference in $g_{i,j}$). Here, the number of queues can be increased to, $G$. However, increasing the number of queues reduces the multicast opportunities and perhaps the performance. Thus choosing the number of queues is a tradeoff between fairness and the system performance.

Remarks: We record here, that the DSMQ, is a culmination of evaluation of multiple schemes that could serve as a potential candidate for providing differentiated QoS (delay). One possible candidate (named “2Q Simultaneous” [22]) is the one that maintains two different queues as DSMQ and serves all the users in head-of-the-line, in both the queues simultaneously, via two streams using one of the MMF beamforming schemes ($P1 - P3$). We have seen via simulations [22], that this gives poor performance mainly because the service times become coupled for both the queues, thus bringing down the performance of the good user queue and the overall system. Further, we also point out that DSMQ is useful only in heterogeneous case. For homogenous case, where we are interested in overall mean sojourn time, we note that [22] DSMQ gives poor performance.

We also note that in literature [13], [14], [15], [16], [17], queueing systems for MIMO/MISO use individual queues for individual users and the transmission/service is performed simultaneously for a subset of users in the head-of-the-line of these queues. These are generalizations of the “2Q Simultaneous” [22] queueing system, albeit without merging of redundant requests. We point out that splitting queues this way (particularly for homogenous system), is detrimental to the performance, since the splitting of queues reduces the multicast opportunities offered by redundant requests for the same files in CCNs. We have shown this via simulations in [22].

V. STATIONARITY OF SMQ AND DSMQ

Before we proceed it is important to establish the existence of stationarity of the proposed queueing systems. Towards this, we define the state of the queue at a given time $t$, as $X_t = \{(i_1, L_{i_1}), \cdots, (i_q, L_{i_q})\}$, where the tuple $(i_j, L_{i_j})$ represents the $j^{th}$ queue entry of file $i_j$ and $L_{i_j}$ is the list of users requesting file $i_j$ and $q$ is the queue length. Let, $E[T]$ be the mean service time when all the users request all the $S$ head of the line files (for any given beamformer setting, $P1$, $P2$ or $P3$). We assume that $E[T] < \infty$. Let $D_j$, denote the sojourn time of $j^{th}$ request arrival to the queue (SMQ/DSMQ). For DSMQ let $X_{i}^G, X_{i}^B$ be the state of the queue, where $X_{i}^G$ and $X_{i}^B$ are defined for the good user queue and the bad user queue in a similar manner. We have the following proposition:

Proposition 1: Under IRM, if $E[T] < \infty$, then for SMQ and DSMQ, $\{D_j\}$ is an aperiodic regenerative process with finite mean regeneration interval and hence has a unique stationary distribution. Also, starting from any initial distribution, $\{D_j\}$ converges in total variation to the stationary distribution.

Proof: Let, $Y_n$, be the state of the queue just after $n^{th}$ departure. Since, there are only $N$, finite number of files in the library, by IRM assumption, $\{Y_n\}$ is a finite state, irreducible discrete time Markov chain (DTMC), [26]. Now, we show that $\{Y_n\}$ is also aperiodic. To see this, consider the state $Y_n = \{\phi\}$, that is the queue is empty. We note that $P(Y_{n+1} = \{\phi\}|Y_n = \{\phi\}) > 0$, since starting from $Y_n = \{\phi\}$, the event that there is exactly one arrival, has positive probability. Thus the DTMC has a unique stationary distribution.

Next, consider the delay $D_j$ of the $j^{th}$ arrival to the system just after $Y_n = \{\phi\}$. The epochs, $Y_n = \{\phi\}$ are also regeneration epochs for $\{D_j\}$. Let $E[\tau]$ be the mean regeneration length of $\{Y_n\}$. The mean number of total request arrivals to the BS during these regeneration epochs, defined as $\pi$ is bounded by $\lambda E[\tau]E[T] + 1$ which is finite. Further $\pi$ is also the mean regeneration length of the $\{D_j\}$ process. Therefore, $\{D_j\}$ also has finite mean regeneration length and
is aperiodic by the argument given for \( \{ Y_n \} \). Thus \( \{ D_j \} \) has a unique stationary distribution. Also, starting from any initial distribution, \( \{ D_j \} \) converges in total variation to the stationary distribution.

Similarly, we can show that \( \{ X_t \} \) also has a unique stationary distribution and that starting from any initial distribution, converges in total variation to the stationary distribution.

We can also show stationarity for DSMQ, in a similar manner by considering the state of the two queue system \( \{ Y^G_n, Y^B_n \} \) just after the \( n^{th} \) departure. We remark that the stationary distribution itself can be quite complicated, given that the dimensionality of \( Y_n \) and \( \{ Y^G_n, Y^B_n \} \) will be very large.

Further, since \( E[T] < \infty \) and the queue length is bounded by \( N \), the mean sojourn time is upper bounded by \((N + 1)E[T]\). This is a unique feature of our queue.

The assumption that \( E[T] < \infty \), can be satisfied with slight modification to the service of SMQ/DSMQ. Let, \( r^* = 1/T^*, \) where \( T^* \) is the optimal service time obtained after solving (7), (8) or (9). Further we, fix \( r_s > 0 \), such that, if \( r^* < r_s \), we do not transmit for \( 1/r_s \) secs. This time is typically greater than the coherence time after which the channel, \( H \) changes. We then solve for the new \( H \) and repeat the procedure, say \( n \) times, till \( r^* \geq r_s \). Here, \( n \) has a geometric distribution with parameter \( p \), where \( p \) is the probability, \( P(r^* \geq r_s) \). We choose \( r_s \) such that \( 1-p \) is close to zero. Thus \( E[T] \leq (1+p)/(pr_s) < \infty \).

VI. THEORETICAL APPROXIMATIONS OF MEAN SOJOURN TIME: SMQ AND DSMQ

In this section we derive approximate mean delays/sojourn times for SMQ and DSMQ. In [18] we derived approximate mean sojourn time for SMQ for single antenna case where the service time distribution was i.i.d, its first and second order statistics were known and also when only head of the line was serviced \( (S = 1) \). However, in Multi-Antenna case the service time distribution is dependent on

1) the distribution of channel gains, of different users,
2) the solutions of the optimization problems \( P1 - P3 \) and
3) the stationary user distribution in the queue.

Among these the distribution of channel gains is known (from assumptions). However, a closed form solution to the optimization problems \( P1 - P3 \) is not available, as the problem is known to be NP Hard. Further, the stationary user distribution itself is dependent on service time distribution. Thus, although the service time distribution in principle can be obtained given the channel gain distribution and the user distribution, its closed form is not available. But it can be easily obtained in simulation. We denote the first and second moment of service time as \( T_S(\pi) \) and \( T_S^2(\pi) \) for a given channel gain distribution \( \Sigma \) and user distribution \( \pi \). In this section we derive approximations for the user distribution also. Following that we propose a method for getting the stationary mean sojourn times.

As a first step we approximate the mean sojourn time of SMQ, when the services have a known general and independent distribution \( (GI \) services) as in [18] and extend it to \( S \geq 2 \). This approximation will be further generalized to the multi-antenna case.

A. Approximation of Mean Sojourn Time of SMQ for GI Services

In SMQ for \( S = 1 \), we first typify the requests into two types, Type 1 (T1) and Type 2 (T2). Among the overall requests arriving at the Base Station queue, T1 requests correspond to the first request for a file that is either unavailable in the queue or is in service. Thus when a T1 request arrives at the queue it is put at the tail of the queue and the subsequent requests (T2) for the file are merged with this entry, as mentioned in Section II. Let \( \lambda'_i \) correspond to the rate of arrival of file \( i \) in T1. Now, let \( D_{i1} \) represent the stationary delay experienced by a T1 arrival of file \( i \). By PASTA ([26]), the mean number of total arrivals to the queue till this request is served is given as \( 1 + \lambda_i E[D_{i1}] \), where \( \lambda_i = \sum_{j=1}^{K} \lambda_{ij} \). Thus, \( \lambda'_i = \lambda_i/(1 + \lambda_i E[D_{i1}]) \).

Further, the rate of total T1 traffic for all the files is given as \( \lambda = \sum_{i=1}^{N} \lambda'_i \). We further assume \( E[D_{i1}] = d \), for all \( i \) where \( d < \infty \). In [18], we showed that this approximation is quite accurate and have verified that it holds even for varying file sizes. Finally, we approximate the mean delay of T1 traffic, using P-K formula for M/G/1 queues [26], as

\[
d = \frac{\rho_d}{1 - \rho_d} \left( \frac{E[T^2]}{2E[T]} \right),
\]

(10)

where \( \rho_d = E[T] \lambda' = E[T] \sum_{i=1}^{N} \frac{\lambda_i}{1 + \lambda_i d} \) and \( E[T] \) and \( E[T^2] \) are the first and second moments of i.i.d service times. The mean delay of T1 traffic can be calculated by recursively solving (10), starting from any \( d \). Further, if \( d^* \) is the T1 delay obtained from the recursion, the delay experienced by T2 traffic has mean \( d^*/2 \). This is because the traffic is Poisson and hence the T2 arrivals have uniform distribution in the duration \( d^* \). Thus the overall mean sojourn time is given as

\[
\bar{D} = \frac{\lambda'}{\lambda} d^* + \frac{\lambda - \lambda'}{\lambda} d^*/2 + E[T].
\]

(11)

For \( S > 1 \), since more than 1 file is being served in a service, the utility \( \rho_d \) reduces \( S \) times. Thus, we make another approximation in (10) and replace \( \rho_d \) by \( \rho_d/S \) and we get

\[
d = \frac{\rho_d}{S - \rho_d} \left( \frac{E[T^2]}{2E[T]} \right) = f(d, S, E[T], E[T^2]).
\]

(12)

Again, the mean sojourn time can be derived using (11). It is worth noting that the SMQ may not always have queue length greater than \( S \), particularly at low arrival rates. However, (12) assumes that the queue length is always at least \( S \). This is a source of error which could be large when \( S \) is large. Thus, (12) gives a lower bound on Type 1 delay in such cases, since \( \rho_d \) is reduced \( S \) times.

Following comments are in order:

- The intuition behind the assumption \( E[D_{i1}] = d \) is that the T1 process can be approximated as Poisson (see [18], [22] for justification). Then PASTA holds and hence the assumption.
- Further, the M/G/1 approximation also holds true due to the above reason.
- Similar to [18], via a qualitatively typical example we show that the fixed point equation in (12) has a unique solution for any service time. We have verified that the
Fig. 2. Uniqueness of solution to the fixed point (12) for GI Service times (Constant in this case): K = 40, λ = 40, N = 100, γ = 1, S1 = .15 secs, S2 = .2 secs. The blue curve represents the line y = d. The red and yellow curves represent the RHS of (12) for S1(= E[T]) = .15 and S2(= E(T)) = .2 secs respectively.

curves for different λ, K, S, N, γs are typically similar to those presented in this section. We consider a system with λ = 40, K = 40, S = 1, N = 100 and the file popularity has Zipf distribution with γ = 1. Further, let us consider two pairs of first and second moments of service times (E[T], E[T^2]) = (S1, S1^2) and (S2, S2^2) where S1 = 0.15 and S2 = 0.2. From Figure 2 we see that for both the pairs of service time moments have unique solutions for d = f(d, 1, E[T], E[T^2]). It is also worth noting that the unique solutions lie in the region (right of black vertical lines in figure 2) for each pair of service moments) where ρ2/S < 1. We also see from Figure 2, that d* increases, with increase in moments. This is typically the case for all system parameters. We have seen that d* is non-decreasing in both E[T], E[T^2] for different sets of system parameters.

- To see the stability of the queue analytically, in (12), if λ → ∞, then f → E[T^2]N/d(4E[T]N - 8E[T^2]NS)/4S < ∞. Thus solving the quadratic equation from (12), arising in the limit, we get

\[ d = \frac{E[T]N \pm \sqrt{4E[T]N^2 - 8E[T^2]NS}}{4S}. \]  

(13)

Now, ignoring the negative (irrelevant) root, for large N, we get d ~ E[T]N/S. Thus Type-1 delay (and hence the mean sojourn time in (11)) is upper bounded by E[T]N/S for all λs. Hence, also our queue is stable for all arrival rates, a feature unique to our queue.

Now, we approximate the stationary distribution of users for different file requests, in the queue.

**B. Approximate User Distribution for SMQ**

We derive the distribution of users requesting first S head of the line files (service times depend only on this). We first derive for S = 1 and then generalize it to S ≥ 1. Note that the probability that file \( i \) in \([N]\) is in the queue is \( P(i \in Q) = \lambda_i / X' \). Further, the probability that a T1 request for file \( i \) is made by user \( j \) is given by \( P_{ij} = \lambda_{ij} / \lambda_i \).

Let \( L_{ij} \) represent the set of users requesting file \( i \) when the T1 request was made by user \( j \). Now, just before the service of file \( i \), the \( L_{ij} \) would contain all the users who have made at least one request within time \( D_{1i} \). Then,

\[ q_{ij}^k = P(k \in L_{ij}) = 1 - e^{-\lambda_{ik} E[D_{1i}]} = 1 - e^{-\lambda_{ik} d}. \]  

(14)

Now, let \( U_1 \in \{0, 1\}^K \) represent a K-length vector where \( k^{th} \) element is 1, if the \( k^{th} \) user has requested the file in head of the line and 0 otherwise. We can now approximate the conditional user distribution for head of the line entry given the requested file is \( i_1 \) as

\[ P(U_1 = [u_1^1, \ldots, u_1^K] | i_1) = \prod_{j=1}^{K} P_{i_1j} \prod_{k \neq j, k \neq [K]} (u_k^1 q_{ij}^k + (1 - u_k^1)(1 - q_{ij}^k)). \]  

(15)

Using \( P(i \in Q) \) and (15) we get the head of the line user distribution as

\[ \pi_d = P(U_1 = [u_1^1, \ldots, u_1^K]) = \sum_{i_1 \in [N]} P(i_1 \in Q) P(U_1 = [u_1^1, \ldots, u_1^K] | i_1) \]  

(16)

\[ S \geq 1 \text{ case: Proceeding as for } S = 1, \text{ the probability that first } S \text{ entries have files } i_s, s = [S], i_t \neq i_q, \text{ for } t \neq q \]

\[ P(i_s \in Q, s = [S]) = \lambda_i^s \prod_{s=[S-1]}^{s} \lambda_{s+1}^{s} = (1 - \lambda_i^{s+1}). \]  

(17)

where \( Q \) is the set of files in SMQ.

Note that (17) is an approximation of the actual distribution since the Type 1 delay will actually be less than \( d \) for queue entries (file request positions) other than the head of the line. However, we calculate all \( \lambda_i^s, \forall i_s \in Q, s = [S] \) using the same \( d \). The error due to this approximation may increase for large \( S \), but work well for \( S = 1, 2 \) and 3, the relevant values.

Now, let \( U_s \in \{0, 1\}^K \) represent a K-length vector where \( k^{th} \) element is 1, if the \( k^{th} \) user has requested the file in \( s^{th} \) entry of the queue and 0 otherwise. We can now define the conditional user distribution in the first \( S \) head of the line entries given the files in positions \( 1, \ldots, S \), as,

\[ P(U_s = [u_s^1, \ldots, u_s^K], s = [S], i_s, s = [S]) = \prod_{s=[S], j=[K]} P_{i_tj} \prod_{k \neq j, k \neq [K]} (u_k^s q_{ij}^k + (1 - u_k^s)(1 - q_{ij}^k)). \]  

(18)

Combining (17) and (18) we get the approximate user distribution \( \pi_d^S \) in the \( S \) head of the line entries for a given T1 delay \( d \).

Thus given the parameter \( S \), the user distribution \( \pi_d^S \) and channel gain parameter \( \Sigma \) we can now calculate \( T_d(\pi_d^S) \) and \( T_{\Sigma}(\pi_d^S) \). Using the moments of service time along with the M/G/1 approximation in the previous section, enables calculation of the Mean Sojourn Time of SMQ for the MISO System. For simplicity, we write \( T_{\Sigma}(\pi_d^S) \) and \( T_{\Sigma}^2(\pi_d^S) \) as \( T(d) \) and \( T^2(d) \) respectively.
C. Theoretical Mean Sojourn Time for SMQ MISO

Having derived $\pi_y^d$, we now derive the mean sojourn times of SMQ for multi-antenna case. We rewrite, (12) as

$$d' = f(d', S, T(d), T^2(d)) = \frac{T(d)X'(d')}{S - T(d)X'(d')} \frac{T^2(d)}{2T(d)}. \quad (19)$$

Notice that the fixed point is over the variable $d'$ and also (19) brings in the dependence of service times on $d$. Though the fixed point equation is with $d'$ replacing $d$, keeping $d$ and $d'$ as different variables is useful in algorithm formulation. Also, the M/G/1 approximation (19) assumes that the service times are i.i.d which is not true. However, this approximation works well. Also, unlike (12), in the MISO system, since service time depends on the parameter $S$, equation (19), gives an upper bound at lower arrival rates (as seen via simulations). The algorithm to calculate mean sojourn time is described below.

**Algorithm 1:**

- For any given $d$ the moments $T(d)$ and $T^2(d)$ are bounded (the boundedness arguments are similar to the $E[T] < \infty$ argument in Section V).

- Now, for a given $d$ we get $T(d) = t_1$ and $T^2(d) = t_2$ as follows: $M$ samples of service times, $T_m$, $m \in [M]$, are obtained by solving $P_X = P1, P2$ or $P3$, on $M$ samples of user distribution, channel gains using $\pi_y^d$, $\Sigma$. We get,

$$t_1 = \frac{\sum_{m \in [M]} T_m}{M} \text{ and } t_2 = \frac{\sum_{m \in [M]} T^2_m}{M}.$$

- Now iterating the fixed point equation (19) with $t_1$ and $t_2$, we get a unique $d^*$ such that $d^* = f(d^*, S, T(d), T^2(d))$.

- We replace $d \leftarrow d^*$ and repeat the above steps till $|d_{n} - d_{n-1}| < \epsilon$ for a given $\epsilon$, where $d_n$ is the value of $d$ in the $n^{th}$ iteration. The mean sojourn time is obtained by plugging the final solution $d^*$ in (11).

The pseudo-code for Algorithm 1, to calculate the mean sojourn time for SMQ is given in the extended version of this paper [22]. We have seen that, starting from any $d$, Algorithm 1 converges to a unique value of Type 1 delay and hence to a unique mean sojourn time. To show this we continue with our typical example in Figure 2 and provide some heuristics for the convergence of Algorithm 1 to a unique fixed point of (19). We fix $L = 16$, $g = 1$ and $P_X = P1$. In general, for any system parameters, the plots and the following heuristics are similar.

D. Heuristics for Convergence of Algorithm 1

We begin by noting that as $d$ increases the number of users in the $S$ entries starting from head of the line, increases. This can be easily seen from (14) and $\pi_y^d$. Thus, naturally the service time increases with $d$ as the number of users in the system increase for all $P_X = P1, P2$ and $P3$. Further for very large $d$, we see that the user distribution is such that all the users are present in all the $S$ entries starting from head of the line. However, we note that the service moments remain bounded (from arguments similar to that of $E[T] < \infty$ in Section V).

Further, we have seen that the equation in (12) has a unique fixed point $d^*$ for a given set of values of service moments. Similarly, (19) also has a unique fixed point (here the fixed point equation is in variable $d'$) for each value of $d$.

Thus in each iteration, for a given $d$ Algorithm 1 gives a unique $T(d)$, $T^2(d)$ and $d^*$, where $d^*$ is the fixed point of the equation $d^* = f(d^*, S, T(d), T^2(d))$. In Figure 3, we see that $f(d^*, S, T(d), T^2(d))$ is monotonically increasing with $d$. This is again because of monotonic increase in service time moments with $d$, because of which $d^*$ is non-decreasing as seen in Section VI-A. Further, since Type-I delay $d^*$ and $T(d), T^2(d)$ are bounded, $f(d^*, S, T(d), T^2(d))$ is also bounded for all values of $d$. This can be seen in Figure 3 where $f$ is bounded and non-decreasing.

Now we note that, $T(d)$ and $T^2(d)$ are strictly positive even for $d = 0$ since $P_X = P1, P2, P3$ give strictly positive service times even with one user in each entry (the distributions (14) and $\pi_y^d$ are such that there is atleast one user in each entry even for $d = 0$). Thus the fixed point equation (19) gives a non-zero positive $d^*$ even for $d = 0$. Further assuming continuity (proof of which is beyond the scope of analysis since continuity of $f$ depends on continuity of service moments $T(d)$ and $T^2(d)$ which is further dependent on the NP-Hard problems $P1, P2$ or $P3$) the straight line $d$ intersects $f$ at at least one point. We have seen from evaluations of Algorithm 1 for different system parameters that this point is unique.

From Figure 3 we see that for a typical case, there exists a unique fixed point which the algorithm achieves, starting from any $d$. For every $d$, the algorithm computes the service moments and the corresponding $d^*$ using (19). The algorithm then proceeds iteratively by replacing the $d$ with the new $d^*$ as explained in Section VI-C. The dashed arrows in Figure 3 show different paths, the algorithm takes starting from different $d$, to reach this unique fixed point.

E. Evaluation of Algorithm 1

In this section, we evaluate the accuracy of the approximation of mean sojourn time provided by Algorithm 1. Towards this we look at a system with $L = 16$, $P = 10$, $N_0 = 1$, $N = 100$, $F = 100MHz$, $B = 100MHz$ and the file popularity follows Zipf distribution with parameter $\gamma = 1$. The channels are modelled as complex Gaussian flat fading channels with $g = 1$. To evaluate the goodness of fit of our approximation we vary the system parameters as $\lambda = 10, 20$ and 40 representing...
Let $\overline{T}_1$ and $\overline{T}_2$ be the mean service times and let $\overline{T}_1^2$ and $\overline{T}_2^2$ be the second moments of service times of SMQ-G and SMQ-B, respectively. Now, SMQ-G can be approximated as an independent M/G/1 queue with the first and second moments of service times defined as

$$
\overline{T}_G = \overline{T}_1 + \overline{T}_2/(C-1), \quad \overline{T}_G^2 = \overline{T}_1^2 + \overline{T}_2^2/(C-1)
$$

(20)

and effective arrival rate as in Section VI-B considering only rates $\Lambda_G = \{\lambda_{ij}, i \in [N] \text{ and } j \in K_G\}$, where $K_G$ is the set of good channel users. Similarly, SMQ-B can be approximated as an M/G/1 queue with the first and second order mean service times given as

$$
\overline{T}_B = \overline{T}_1(C-1) + \overline{T}_2, \quad \overline{T}_B^2 = \overline{T}_1^2(C-1) + \overline{T}_2^2.
$$

(21)

Also, the effective arrival rate as in section VI-B considering only rates $\Lambda_B = \{\lambda_{ij}, i \in [N] \text{ and } j \in K_B\}$, where $K_B$ is the set of bad channel users. Now similar to SMQ, we can approximate mean T1 delay of SMQ-G using the fixed point equation

$$
d_1 = f_G(d_1, C, S, \overline{T}_G, \overline{T}_G^2) = \frac{\rho d_1}{S - \rho d_1} \left( \frac{\overline{T}_G}{2\overline{T}_G^2} \right),
$$

(22)

where $\rho d_1 = \overline{T}_G \lambda'_G(d_1)$, and $\lambda'_G = \sum_{i=1}^{N} \frac{\sum_{j \in K_G} \lambda_{ij}}{1 + \rho d_1 \sum_{j \in K_G} \lambda_{ij}}$.

Also, for SMQ-B the mean T1 delay is obtained using the fixed point equation

$$
d_2 = f_B(d_2, C, S, \overline{T}_B, \overline{T}_B^2) = \frac{\rho d_2}{S - \rho d_2} \left( \frac{\overline{T}_B}{2\overline{T}_B^2} \right),
$$

(23)

where $\rho d_2 = \overline{T}_B \lambda'_B$, and $\lambda'_B = \sum_{i=1}^{N} \frac{\sum_{j \in K_B} \lambda_{ij}}{1 + \rho d_2 \sum_{j \in K_B} \lambda_{ij}}$.

Thus T1 delays, $d_1$ and $d_2$ for both SMQ-G and SMQ-B can be calculated using recursive equations (22) and (23) independently. The uniqueness of $d_1$ and $d_2$ holds true even in this case as both the queues are treated independently. Further, the user distributions $\pi^B_S$ and $\pi^G_S$ for SMQ-G and SMQ-B can be derived in the same manner as in Section VI-B, with rates $\Lambda_G$ and $\Lambda_B$ respectively.

G. Theoretical Mean Sojourn Time for DSMQ MISO

We can extend Algorithm 1 to derive mean sojourn time for the DSMQ MISO system. The DSMQ can again be considered as a Polling system with $\Lambda_G$ arrivals to SMQ-G and $\Lambda_B$ arrivals to SMQ-B. Similar to $T(d)$ and $T^2(d)$ in the SMQ case, the first and second moments of the service times of SMQ-G and SMQ-B are represented as $T_n(d_n), T^2_n(d_n)$, $n = 1, 2$, respectively. The moments $T_n(d_n), T^2_n(d_n)$, $n = 1, 2$ are dependent on $\Sigma, \pi^B_S$ and $\pi^G_S$. Similar to the SMQ case we replace $\overline{T}_G$ and $\overline{T}_G^2$, $n = 1, 2$, in equations (20) and (21) with $T_n(d_n), T^2_n(d_n)$, $n = 1, 2$ respectively. Thus, unlike SMQ case, the $T_G$ and $T_B$ are both dependent on both $d_1$ and $d_2$. Using the new $T_G$ and $T_B$ in (22) and (23) we get two new fixed point equations similar to (19).

Algorithm 2:

The algorithm to calculate the mean sojourn time of DSMQ, proceeds in a manner similar to Algorithm 1, except for the coupling in service times coming from the dependence of
Further to calculate a mean sojourn time with less than (P2 and P3) have similar accuracies. We fix $g = 0 dB$, $\forall k \in [K_G]$ and for bad users is $g_k = -15 dB$, $\forall k \in [K] \setminus [K_G]$. We fix $\lambda = 40$. The rest of the parameters are same as in Section VI-E. We also vary the system parameter $S = 1, 2$. The algorithm parameters are fixed as follows: $P_X = P1$, $\epsilon = 0.1$ and $M_1 = 500$, $M_2 = 500$. The comparison here is shown for MMF beamforming (P1). Other beamforming cases (P2 and P3) have similar accuracies.

From Figure 5 we see that Algorithm 2 (Theory) gives a mean sojourn time with less than 10% error for the case $\lambda = 40$. Here we have considered only the high load condition.

The Algorithm 2 performs similar to Algorithm 1 in low load conditions as well, however, we do not present here for clarity of presentation.

Further our Algorithms are not restricted to beamforming cases presented in this work but can be adapted to any beamforming strategies a designer may wish to apply by appropriately changing the optimization. It would also be much faster if one could get a closed form expressions for the moments of service time for a given user distribution (which typically is not the case in the current multi-user MIMO literature).

We would also like to point out that the Algorithm 2 can be used to analyse any M/G/1 type E-limited Polling system with zero switchover times to get the theoretical waiting times for users, which maybe of independent interest.

VII. SIMULATION RESULTS AND DISCUSSION

In this section we present, simulation results and comparison of different beamforming schemes with SMQ and DSMQ. We consider two cases of channel statistics. First with homogenous channel statistics across users and second with heterogeneous channel statistics where there are users with good and bad channel statistics. All our simulations use complex Gaussian flat fading channels as explained in Section II. To avoid arbitrarily large service times, we fix $r_e = 0.01$, (see Section V). This ensures that $E[T] < \infty$, needed in Proposition 1, for all our schemes. All our simulations are run for 10000 services of the queue at the BS, to let the queues reach stationarity. The mean sojourn times are estimated using sample average of sojourn times seen during the simulation.

Case 1 (Homogeneous Channel Statistics): We consider a MISO network with $L = 16$ antennas, $N = 100$ files, each file of size $F = 100 MB$ and system bandwidth $B = 100 MHz$. Channels between each antenna and users are i.i.d complex Gaussian with mean fading $g = 1$. The popularity of files follows Zipf distribution with popularity $\gamma = 1$. This is a common assumption and is shown to reflect the content request traffic in servers as youtube, netflix [2]. We fix $P = 10$ and assume that the average noise power $\sigma_k^2 = N_0 = 1$, $\forall k \in [K]$ in all our systems. To cater for all scenarios (low, medium and high traffic) we consider systems with $K = 10, 40$ users, and the arrival rates $\lambda$ = 10, 20, 40, 60, 80 files/sec. The first case of $K = 10$ represents low load scenario. Since the number of antennas are more than the number of users, the beamformer has higher degrees of freedom to null the inter-stream interference (if any) in all kinds of traffic. However, to bring in the effect of queueing we also look at the different arrival rates $\lambda = 10, 20, 40, 60, 80$. The second case of $K = 40$ represents moderately loaded condition. Here the total active users (for each file in the queue) may actually be less than the total number of antennas, when the traffic is low (e.g., $\lambda = 10, 20$) and greater when the traffic is high $\lambda = 40, 60, 80$.

Figures 6 and 7 show the comparison of mean sojourn times for different schemes for different arrival rates for cases $K = 10$ and $40$ respectively. The first observation we make is that, increasing the number of streams $(S \geq 2)$ is beneficial for SMQ MMF and SMQ MMF-RS, only in
low and moderate arrival rates. Compared to $S = 1$, both $S = 2,3$ provide $50$ – $75\%$ improvement for $\lambda = 10, 20$ and $30$ – $50\%$ improvement for $\lambda = 40$. At very high traffic $\lambda = 80$, increasing $S$ provides negligible gain. The reason is that at very high traffic each file entry in the queue has enough requests to provide multicast opportunities and adding more streams provides no advantage. However, at lower and medium arrival rates the spatial multiplexing gains are provided by $S = 2,3$ in addition to the multicast gain provided by the SMQ. Note that the performance reduction of $S = 2,3$ streams is also due to the fact that there may exist common users in $S$ groups which may limit the rate, thus reducing the multiplexing gain.

Our second observation is that the MMF-SIC is beneficial only when the total number of users are less than the total number of antennas, $(K = 10, L = 16)$.

Further we make another important observation that MMF-RS beamforming in SMQ performs similar to (or) worse than MMF beamforming case for all cases of $K = 10, 20, 40$ and $S = 2,3$. For $S = 1$ the performances of MMF and MMF-RS are similar. This is because of the optimization of the max transmit time (6) among the degraded and designated streams, which is inevitable in queued systems such as SMQ. This result is in stark contrast to what is observed in a system without queues [4], [5] where, RS based beamforming always does better than any other beamforming scheme. However, we will see in the following that MMF-RS provides significant advantage in the heterogeneous case, even in a queued system.

Case 2 (Heterogeneous Channel Statistics): We consider only MMF and MMF-RS in this section (MMF-SIC performs poorly in heterogeneous case as well. Hence, we do not present it here for the sake of clarity of presentation). We consider the system with $K = 40$ users among which $K_G = 20$ are good users and the rest $K_B = 20$ are bad users. The mean fading for good users is $g_k = 0\, dB$, $\forall k \in [K_G]$ and for bad users is $g_k = -15\, dB$, $\forall k \in [K] \setminus [K_G]$. In other words bad users undergo $15dB$ deeper fading than the good users. In practical systems, this does happen. All other parameters stay the same.

We compare the SMQ and DSMQ in terms of mean sojourn times experienced by good users in Figure 8 and bad users in Figure 9.

As explained before, in SMQ both good and bad users experience the same mean sojourn time. When we compare MMF SMQ $S = 1$ for $\lambda = 40$ in Figures 8, 9 with Figure 7, we see that the presence of bad users increases the mean sojourn time of all users (good and bad), from 7 to 145 secs. This is a significant degradation. Further, in contrast to the homogeneous case, MMF-RS SMQ $S = 1$ improves the mean sojourn time to 70 secs, in Figure 8. Nevertheless, the delay is still significantly high. Now consider MMF DSMQ with
$C = 8$, and $C = 15$. We see from Figure 8 that $C = 15$ mostly recovers the mean sojourn time for good users to $\sim 20$ secs as compared to $\sim 7$ secs for MMF SMQ in Figure 7. However, the delay of bad users is severely degraded to 300 secs. This is not desirable. Setting $C = 8$ improves this situation by providing mean sojourn time of 30 secs to good users and 210 secs to bad users. Thus, $C$ can be tuned to get desirable fairness.

Next, in Figure 9 we observe that for $C = 8$, MMF-RS DSMQ further reduces the mean sojourn time of bad channel users from 210 secs for MMF DSMQ to 135 secs, for $\lambda = 40$. This also results in a slight improvement of 5 secs for the good users, Figure 8. Further, fine tuning of MMF-RS DSMQ with $C = 5$ controls the fair allocation of QoS to good and bad channel users, resulting in delays of 30 and 100 secs, respectively, for $\lambda = 40$. We can see similar trends for $\lambda = 10, 20$ in Figures 8 and 9. Finally, we remark that increasing $S (\geq 2)$ is only beneficial in MMF SMQ for $\lambda = 10, 20$ for reasons similar to the homogeneous case.

Such QoS allocation by DSMQ is quite useful in practical CCN networks to prevent situations where bad users might restrict good users from watching HD content. We would like to point out that we have evaluated other content centric queues such as Loopback ([18]) which were designed for the same purpose (providing optimal fairness), for the MISO system with heterogeneous channels. In the extended version of this paper [22], we have also seen via simulations that such queues have inferior performance to DSMQ. We have seen via simulations that the reinforcement learning based Power Control in time [21], which was very useful in SISO, does not provide any performance improvement to the queues that are considered in this paper. This is mainly because the spatial diversity in a MISO system, makes up for the time diversity utilised by these algorithms. Therefore we can conclude that the choice of queueing in a MISO system is an important problem and that the QoS can be significantly improved with careful design.

From these simulations, it is clear that MMF-RS DSMQ and SMQ-MMF perform the best in heterogeneous and homogenous case, respectively. Therefore we can conclude that the choice of a queueing and a beamforming scheme is a coupled problem and that the QoS and fairness are cross layer objectives which can be significantly improved with careful design.

VIII. CONCLUSION AND FUTURE WORK
We have considered practical adaptations of beamforming strategies in a MISO CCN with a queue and evaluated their performance. We show that the Simple Multicast Queue (SMQ) can be adapted to such a MISO setup. For the homogenous channel case, we show that the SMQ combined with the simplest MMF scheme performs the best, thus avoiding the complexities of Rate Splitting (RS). This is in contrast to the results in [4] and [5] where RS performs the best.

Further, we have identified SMQ’s shortcomings in the heterogeneous user channel case. We have proposed a new simple queueing scheme called Dual Simple Multicast Queue (DSMQ) which gives flexibility in allocating different QoS for users with good and bad channels. Here, we have shown that MMF-RS DSMQ performs the best, among all schemes. We have also shown that power control and loopback schemes in [18] and [21], are ineffective in MISO setup. We have also proved the stationarity of the queues and have shown that they are always stable across all arrival rates. We have provided queueing theoretic approximations to the mean sojourn time, which could be very useful in cross-layer analysis of Multiuser MISO CCNs with queues. Finally, we conclude that the selection of the queueing strategy and beamforming is a coupled problem. The pairs (SMQ, MMF) and (DSMQ, MMF-RS) are optimal strategies for homogeneous and heterogeneous cases respectively, among the ones considered in this paper. Future work may include analysis with user movement, imperfect CSIT etc. Similarly, adaptations of SISO SMQ and DSMQ to coded caching and extension to a full MIMO setting with queueing adaptations of the beamforming techniques similar to those in [29] are possible.

REFERENCES
[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2016–2021 White Paper, Cisco, San Jose, CA, USA, 2016.
[2] M. Laterman, M. Arlitt, and C. Williamson, “A campus-level view of Netflix and twitch: Characterization and performance implications,” in Proc. Int. Symp. Perform. Eval. Comput. Telecommun. Syst. (SPECTS), Jul. 2017, pp. 1–8.
[3] E. Karipidis, N. D. Sidiropoulos, and Z.-Q. Luo, “Quality of service and max-min fair transmit beamforming to multiple cochannel multicast groups,” IEEE Trans. Signal Process., vol. 56, no. 3, pp. 1268–1279, Mar. 2008.
[4] B. Clerckx, Y. Mao, R. Schober, and H. V. Poor, “Rate-splitting unifying SDMA, OMA, NOMA, and multicasting in MISO broadcast channel: A simple two-user rate analysis,” IEEE Wireless Commun. Lett., vol. 9, no. 3, pp. 349–353, Mar. 2020.
[5] H. Joudeh and B. Clerckx, “Rate-splitting for max-min fair multigroup multicast beamforming in overloaded systems,” IEEE Trans. Wireless Commun., vol. 16, no. 11, pp. 7276–7289, Nov. 2017.
[6] N. Naderializadeh, M. A. Maddah-Ali, and A. S. Avestimehr, “Fundamental limits of cache-aided interference management,” IEEE Trans. Inf. Theory, vol. 63, no. 5, pp. 3092–3107, May 2017.
[7] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. H. Khalaj, “Multi-antenna interference management for coded caching,” IEEE Trans. Wireless Commun., vol. 19, no. 3, pp. 2091–2106, Mar. 2020.
[8] M. Aloehd et al., “Symbol-level and multicast precoding for multiuser multiantenna downlink: A state-of-the-art, classification, and challenges,” IEEE Commun. Surveys Tuts., vol. 20, no. 3, pp. 1733–1757, 3rd Quart., 2018.
[9] C. Antón-Haro, P. Svedman, M. Bengtsson, A. Alexiou, and A. Gameiro, “Cross-layer scheduling for multi-user MIMO systems,” IEEE Commun. Mag., vol. 44, no. 9, pp. 39–45, Sep. 2006.
[10] A. Z. Yalcin, M. Yuksel, and B. Clerckx, “Rate splitting for multi-group multicasting with a common message,” IEEE Trans. Veh. Technol., vol. 69, no. 10, pp. 12281–12285, Oct. 2020.
[11] E. Castañeda, A. Silva, A. Gameiro, and M. Kountouris, “An overview on resource allocation techniques for multi-user MIMO systems,” IEEE Commun. Surveys Tuts., vol. 19, no. 1, pp. 239–284, 1st Quart., 2017.
[12] B. Bellalta, V. Daza, and M. Oliver, “An approximate queueing model for multi-rate multi-user MIMO systems,” IEEE Commun. Lett., vol. 15, no. 4, pp. 392–394, Apr. 2011.
[13] M. Deghel, M. Assaad, M. Debbah, and A. Ephremides, “Queueing stability and CSI probing of a TDD wireless network with interference alignment,” IEEE Trans. Inf. Theory, vol. 64, no. 1, pp. 547–576, Jan. 2018.
[14] Z. Xie and W. Chen, “A joint channel and queue aware scheduling method for multi-user massive MIMO systems,” in Proc. IEEE Int. Conf. Commun. (ICC), May 2019, pp. 1–6.
[15] K. Huang and V. K. N. Lau, “Stability and delay of zero-forcing SDMA with limited feedback,” IEEE Trans. Inf. Theory, vol. 58, no. 10, pp. 6499–6514, Oct. 2012.
[16] J. Arnau and M. Kountouris, “Delay performance of MISO wireless communications,” in Proc. 16th Int. Symp. Model. Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt), May 2018, pp. 1–8.

[17] J. Chen and V. K. N. Lau, “Large deviation delay analysis of queue-aware multi-user MIMO systems with two-timescale mobile-driven feedback,” IEEE Trans. Signal Process., vol. 61, no. 16, pp. 4067–4076, Aug. 2013.

[18] M. Panju, R. Raghu, V. Sharma, V. Aggarwal, and R. Ramachandran, “Queueing theoretic models for uncoded and coded multicast wireless networks with caches,” IEEE Trans. Wireless Commun., vol. 21, no. 2, pp. 1257–1271, Feb. 2022.

[19] B. Zhou, Y. Cui, and M. Tao, “Optimal dynamic multicast scheduling for cache-enabled content-centric wireless networks,” IEEE Trans. Commun., vol. 65, no. 7, pp. 2956–2970, Jul. 2017.

[20] R. Raghu, M. Panju, V. Aggarwal, and V. Sharma, “Scheduling and power control for wireless multicast systems via deep reinforcement learning,” Entropy, vol. 23, no. 12, p. 1555, Nov. 2021. [Online]. Available: https://www.mdpi.com/1099-4300/23/12/1555

[21] R. Raghu, P. Upadhyaya, M. Panju, V. Aggarwal, and V. Sharma, “Deep reinforcement learning based power control for wireless multicast systems,” in Proc. Allerton, 2019.

[22] R. Raghu, M. Panju, and V. Sharma, “Queueing theoretic models for multiten MISO content-centric networks with SDMA, NOMA, OMA and rate-splitting downlink,” 2021, arXiv:2111.06352.

[23] M. Panju, R. Raghu, V. Aggarwal, V. Sharma, and R. Ramachandran, “Queueing theoretic models for multicasting under fading,” in Proc. IEEE Wireless Commun. Netw. Conf. (WCNC), Apr. 2019, pp. 1–6.

[24] O. Dizdar, Y. Mas, W. Han, and B. Clerckx, “Rate-splitting multiple access for downlink multi-antenna communications: Physical layer design and link-level simulations,” in Proc. IEEE 31st Annu. Int. Symp. Pers., Indoor Mobile Radio Commun., Aug. 2020, pp. 1–6.

[25] A. Z. Yalcin, M. K. Cetin, and M. Yüksel, “Max-min fair precoder design and power allocation for MU-MIMO NOMA,” IEEE Trans. Veh. Technol., vol. 70, no. 6, pp. 6217–6221, Jun. 2021.

[26] S. Asmussen, Applied Probability and Queues, 2nd ed. New York, NY, USA: Springer, 2003.

[27] W. Yue, Y. Takahashi, and H. Takagi, Advances in Queueing Theory and Network Applications. Berlin, Germany: Springer, 2009.

[28] S. W. Fuhrmann and Y. T. Wang, “Analysis of cyclic service systems with limited service: Bounds and approximations,” Perform. Eval., vol. 9, no. 1, pp. 35–54, Nov. 1988.

[29] A. R. Flores and R. C. de Lamare, “Robust and adaptive power allocation techniques for rate splitting based MU-MIMO systems,” IEEE Trans. Commun., vol. 70, no. 7, pp. 4656–4670, Jul. 2022.

Ramkumar Raghu received the B.Tech. degree in electronics engineering from the Indian Institute of Technology (ISM) Dhanbad in 2009 and the M.S. (Eng.) and Ph.D. degrees in electrical communication engineering from the Indian Institute of Science, Bangalore, in 2023. He has been a Scientist with the Center for Airborne Systems, DRDO, since 2009. He is currently leading the development of MIMO digital radars. He has several publications in reputed IEEE conferences and journals. His research interests include MIMO systems, AESA digital radars and adaptive digital beamforming, wireless communication, content-centric networks, and machine learning. He has won several awards, including DRDO Agni Award and Technology Group Award.

Mahades Panju received the M.Tech. degree in electronics engineering from the Indian Institute of Technology Bombay in 2011. He is currently pursuing the Ph.D. degree in electrical communication engineering with the Indian Institute of Science, Bangalore. He is also with Qualcomm. His research interests include wireless communication, stochastic processes, and machine learning.

Vinod Sharma (Senior Member, IEEE) received the B.Tech. degree in EE from the Indian Institute of Technology Delhi, India, in 1978, and the Ph.D. degree in ECE from Carnegie Mellon University, in 1984. He was with Northeastern University, Boston, and University of California at Los Angeles, before joining the Indian Institute of Science (IISc), Bangalore, in 1988. In IISc, Bangalore, he has been the Chair of the ECE Department, a TATA CHEM Chair Professor, and the Dean of the Engineering Faculty and has superannuated in July 2022. Currently, he is a Professor with the Electrical Engineering Department, Shiv Nadar Institution of Eminence, Greater Noida. His research interests are in communication networks, wireless communications, and information theory. He has more than 200 papers in international journals and conferences in these areas. He is an elected fellow of the Indian National Academy of Engineering (INAE). He has won Prof. Rustom Choksi Award for excellence in research in 2017 and ACCS-CDAC Award in 2012.