Wireless Multicasting for Content Distribution: Stability and Delay Gain Analysis

Bahman Abolhassani[1], John Tadrous[2], Atilla Eryilmaz[1]
[1,2] Department of Electrical and Computer Engineering
[1] The Ohio State University, Columbus, 43210
[1] Email: abolhassani.2@osu.edu, eryilmaz.2@osu.edu
[2] Gonzaga University, Spokane, WA 99202
[2] Email: tadrous@gonzaga.edu

***Abstract*—In this work, we provide a comprehensive analysis of stability properties and delay gains that wireless multicasting capabilities, as opposed to more traditional unicast transmissions, can provide for content distribution in mobile networks. In particular, we propose a model and characterize the average queue-length (and hence average delay) performance of unicasting and various multicasting strategies for serving a dynamic user population at the wireless edge. First, we show that optimized static randomized multicasting (we call it 'blind multicasting') leads to stable-everywhere operation irrespective of the network loading factor (given by the ratio of the demand rate to the service rate) and the content popularity distribution. In contrast, traditional unicasting suffers from unstable operation when the loading factor approaches one, although it outperforms blind multicasting at small loading factor levels. This motivates us to study 'work-conserving multicast' policies next that always outperform unicasting while still offering stable-everywhere operation. Then, in the worst-case of uniformly-distributed content popularity, we explicitly characterize the scaling of the average queue-length (and hence delay) under a first-come-first-serve multicast strategy as a function of the database size and the loading factor. Consequently, this work provides the fundamental limits, as well as the guidelines, for the design and performance analysis of efficient multicasting strategies for wireless content distribution.**

***Index Terms*—Wireless Content Distribution, Multicast, Delay Gains, Information-Centric Networking.**

I. Introduction

The recent advances in the development of capable smart wireless devices and mobile internet services have resulted in groundbreaking levels of data traffic over cellular networks. This excessive data demand is depleting the limited spectrum resources of wireless transmissions, especially the wireless connection between the base stations and the end-users. Consequently, wireless resources are becoming scarce due to the tremendous development of throughput-hungry applications including video streaming and online gaming. Thus, more sophisticated resource management strategies are needed in order to effectively meet the growing demand.

This work was supported by the NSF grants: CCSS-EARS-1444026, CNS-NeTS-1514260, CNS-NeTS-1717045, CMMI-SMOR-1562065, CNS-ICN-WEN-1719371, and CNS-SpecEES-1824337; the DTRA grants: HDTRA1-15-1-0003; HDTRA1-18-1-0050.

To tackle this problem, several techniques have already been proposed such as WiFi offloading, proactive caching, and wireless multicasting. WiFi offloading is a straight-forward approach to address the unprecedented explosion of data traffic. This approach is based on offloading some of the traffic to WiFi networks (e.g., [1]). Different approaches to implement WiFi offloading and to improve the performance of WiFi offloading have been investigated in [2]. In the aforementioned approaches, scheduling of wireless demand is applied reactively so that data requests are initiated beforehand, and the service provider utilizes the delay tolerance from end-users to schedule them efficiently. Thus, cost reduction comes at the expense of disturbed user activity patterns as the service is postponed to off-peak times, or the next available WiFi connection. Another possible solution to address the problem is to cache popular contents on the user's site (e.g., [3]). Cache system can help reduce the total response time of users' requests. Cached data can be shared by users at the same site. It also enables reduced peak-to-average traffic ratio for the original data management system [4]. By knowing the popularity of contents, caching efficiency can be improved by pre-downloading popular contents during off-peak times and serving predictable peak-hour demands, which is referred to as proactive caching (see [5]). However, because of the limited capacity of caching storage, this technique has also its limitations.

In this work, we consider another natural alternative strategy to alleviate the growing traffic load of wireless content distribution, namely, *multicasting* whereby content of common interest is transmitted to multiple users at once. As an example to illustrate the benefits of multicasting, consider a football stadium full of people watching the game and after a goal, many of them may request (at different times) the related footage to watch it at their smart device, giving the opportunity to broadcast content of common interest to multiple users with small delay. Even though multicasting is widely acknowledged to be a promising approach in such scenarios, its potential gains have not fully been investigated or realized. *Unicasting* is the predominant policy that is widely used in wireless content distribution [6]. Transition from IP based networks to information-centric networks (see [7]) encourages us to

rigorously investigate the multicasting gain in such networks.

In particular, we focus on the distribution of data content to dynamic users over wireless channels, whereby the wireless network can simultaneously serve all the requests awaiting the same data content at the time. We aim to reveal the stability conditions and the delay gains that multicasting can offer over its unicast counterpart. Our contributions, along with the organization of the paper, are as follows.

- In Section II, we present a tractable content distribution model for serving dynamically arriving demand over wireless broadcast channels.
- In Section III-A, for a database of n items with an arbitrary popularity distribution, we develop the optimal static-randomized multicasting strategy (called *blind multicasting*) that minimizes the aggregate average number of requests in the system. While unicast transmissions can only stabilize the system when the loading factor ρ (given by the ratio of the demand rate to the service rate) is less than 1, we show in Theorem 1 (proved in Section IV-A) that under our blind multicasting, the system is always stable for all $\rho \geq 0$.
- Moving beyond stability for the worst-case uniform popularity distribution, in Section III-B we expand the policies to the more efficient class of *work conserving* multicasting policies in order to improve the delay gains. In Theorem 2, we explicitly characterize the scaling delay gains of the *First-Come-First-Serve* work-conserving multicasting strategy as a function of the loading factor ρ and the database size n. The proof of Theorem 2, presented in Section IV-B, may be of independent-value as it utilizes a novel approach for dealing with the nontraditional abruptly-changing (as opposed to the traditional incremental) nature of queueing dynamics under multicasting transmissions.
- In Section V, we provide numerical simulations to validate the analytical results and compare the performance to other service strategies such as Max-Weight-based multicasting. Finally, we conclude in Section VI.

II. System Model

We consider a wireless network comprising a content provider that serves a population of users through a wireless base station (BS) deployed at the network edge. In a continuous time fashion, the users[1] dynamically send requests targeting content from a set of n data items with certain popularity distribution offered by the content provider. The wireless BS enqueues the incoming requests in n distinct queues, one queue per data item, in order to serve them.

Demand Generation: The population of users covered by the BS are assumed to generate data requests according to a Poisson process with rate λ. That is, for $A_{tot}(t), t \geq 0$ being the aggregated number of generated requests by time t, then $A_{tot}(t)$ is a Poisson random variable with mean λt.

The incoming requests at any point in time are split *independently* over the n data items based on their respective popularity. We capture the popularity of a data item k by the probability of that item k being intended by a request given a request is already generated. We denote such probability by α_k, $k = 1, \cdots, n$, where $\sum_{k=1}^{n} \alpha_k = 1$. Thus, the aggregate request generation process $\{A_{tot}(t)\}_t$ is the superposition of n independent Poisson processes $A_{tot}(t) := \sum_{k=1}^{n} A_k(t)$, where $A_k(t), t \geq 0$ is the request arrival process for item k which is Poisson with rate $\alpha_k \lambda$. We consider the vector $\boldsymbol{\alpha} := (\alpha_k)_{k=1}^{n}$ as the popularity profile of the system.

Service Dynamics: The base station serves requests one at a time. The service time of an individual request is considered to follow an exponential distribution with mean $1/\mu$ and the service times are assumed independent and identically distributed over time and requests. While, in practice, service times may exhibit heavily-tailed distributions due to data item length and retransmissions over the wireless medium, we adopt the exponential distribution to allow tractable characterization of the multicasting gains and contrast it with the well-known unicast results that are already derived for exponentially distributed service times.

The n queues maintained at the BS hold the requests awaiting service with queue k has all the pending requests for item k. We consider these queues to be of infinite length, hence we are not concerned with outage events due to lost requests. Instead, we care about the average delay these requests incur as our metric of interest. Since the set of items requested by users in a typical content distribution network is very large, considering that $n \to \infty$ is an reasonable assumption.

We denote the number of requests in queue k at time t by $Q_k(t)$, $k = 1, \cdots, n$. We define the service completion of a request from queue k as an ON-OFF process $B_k(t)$ where $B_k(t) = 1$ if a request from queue k has completed service at time t, otherwise $B_k(t) = 0$. We can thus define the service completion of any request from any queue as the ON-OFF process $B_{tot}(t) := \sum_{k=1}^{n} B_k(t)$.

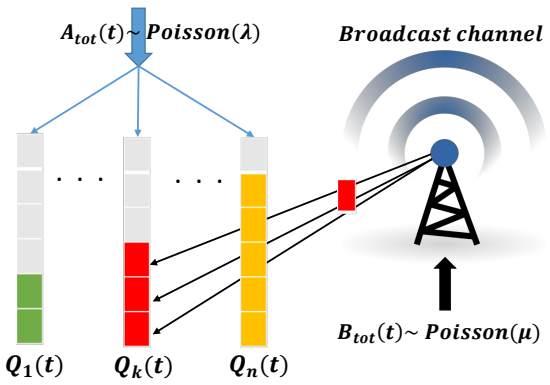


Fig. 1: Queuing system model

Fig. 1 shows the model for our queueing system. Requests are generated at a rate of λ and based on the item being requested, each request is placed in a queue dedicated for that item. Then requests are served at the BS with a rate of $\mu > 0$. In this paper, we are interested in the comparative and comprehensive study of *unicast* (as the baseline that is widely

[1]Note that the number of users that generate demand is unbounded, as in the infinite-population setting of classical Aloha networks.

adopted by today's wireless technologies) and *multicast* modes of service that are described next.

Unicast and Multicast Operation: Through unicast operation, the BS has to individually serve the requests in each queue, one request at a time. Thus, when an a request is served from any queue, the length of such queue is decremented by one. Let $Q_k^U(t)$ be the number of requests in queue k at time t under the unicast operation, then for dt being an infinitesimal increment in time, then[2]

$$Q_k^U(t+dt) = [Q_k^U(t) - B_k(t)]^+ + A_k(t+dt) - A_k(t),$$

where $[x]^+ = \max\{0, x\}$.

In the multicast operation, the BS relies on the broadcast nature of the wireless medium to send the requested data simultaneously to all the requesting users, consuming the same amount of resources required by a single unicast transmission. Thus, if $Q_k^M(t)$ is the number of requests in Queue k under the multicast operation, then

$$Q_k^M(t+dt) = (Q_k^M(t))(1 - B_k(t)) + A_k(t+dt) - A_k(t),$$

that is, as shown in Fig. 1, the service of a single request from queue k collectively serves all of the requests in queue k yielding an empty queue after each service. This is the key difference between multicast and unicast dynamics.

Performance Metric: We use the time-average expected number of requests in the system as our performance metric to quantify the gains of multicasting. At any time t, the number of requests in the system under Unicast and Multicast operations are $Q_{tot}^U(t)$ and $Q_{tot}^M(t)$, respectively, where $Q_{tot}^o(t) := \sum_{k=1}^n Q_k^o(t), \quad o \in \{U, M\}$.

For any queue-length process $Q_k(t)$, we use the notation $\overline{Q}_k$ to indicate its time-average expected value, that is,

$$\overline{Q}_k := \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbb{E}[Q(t)]dt.$$

Accordingly, the time-average of the expected total number of requests in the system under unicast and multicast operation is denoted by $\overline{Q}_{tot}^U$, $\overline{Q}_{tot}^M$, respectively.

We finally define the loading factor $\rho := \frac{\lambda}{\mu}$ as a key parameter shaping the traffic intensity of the system. We then investigate the system's performance with the number of data items n in different regimes of ρ. We begin with the unicast operation as it constitutes our baseline model. From the well known results of an M/M/1 queue [8], we have

$$\overline{Q}_{tot}^U = \frac{\rho}{1-\rho}, \quad \rho \in [0, 1), \tag{1}$$

which clearly shows that the system can be stabilized by unicasting only for $\rho < 1$. We can also observe that $\overline{Q}_{tot}^U$ depends neither on the number of data items n, nor on the individual popularity of data items, since the service of requests is carried out on an individual request basis. In the following sections, we investigate the behavior of $\overline{Q}_{tot}^M$ and compare it to that of its unicast counterpart.

[2]We note that the main results of this work will remain essentially the same if we use: $Q_k^U(t+dt) = [Q_k^U(t) - B_k(t) + A_k(t+dt) - A_k(t)]^+$.

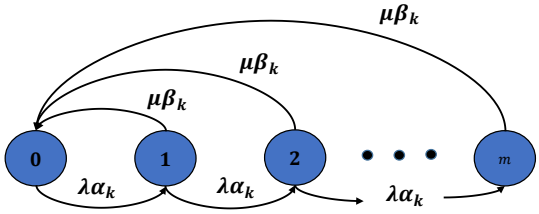


Fig. 2: Markov chain diagram of queue k under blind multicast operation.

III. Stability and Delay Gain Results of Blind and Work-Conserving Multicast Policies

This section presents the main results of this paper and highlights the significant multicasting gains, with their detailed proofs postponed to Section IV. We first show the endless stability operation furnished by simple multicasting strategies (cf. Theorem 1). Then, we explore further multicasting gains under a first-come-first-serve work-conserving operation (cf. Theorem 2). We conclude this section with a discussion of key insights from these results.

A. Endless Stability of Blind Multicast

We begin by considering a simple static multicasting strategy which we label *blind multicast*. This strategy is suitable for scenarios whereby the individual requests are not known by the BS, and multicasting decisions are made blindly based on the statistical popularity information. As such, it is convenient in conditions where it is not feasible to receive feedback from the individual users.

Definition 1 (Blind Multicast): Define the indicator $\sigma_k^{M,B}(t) \in \{0, 1\}$ to capture the service decision of queue k at time t such that $\sigma_k^{M,B}(t) = 1$ if the queue k is assigned the service resources at time t, otherwise $\sigma_k^{M,B}(t) = 0$, $k = 1, \cdots, n$. Then, blind multicast strategy is a randomized strategy through which the BS randomly assigns the service resources to n queues such that

$$\beta_k := \lim_{T\to\infty} \frac{1}{T} \int_0^T \sigma_k^{M,B}(t)dt, \quad k = 1, \cdots, n,$$

for $(\beta_k)_{k=1}^n$ is a vector of non-negative weights to be determined.

We can note from the definition that the allocation of the service resources to queues is independent of the queue length, hence the naming *blind*. A blind multicasting strategy thus assigns the service to queue k for a fraction β_k of the time irrespective of its instantaneous state.

The whole system under blind multicast can be split into n independent and parallel queues with queue k having an arrival rate of $\alpha_k\lambda$ and service rate $\beta_k\mu$ with state evolution as shown in Fig. 2. Each state represents the number of requests in the queue k. We have the following result for such multicasting.

Theorem 1 (Endless Stability of Delay-Optimizing Blind Multicast): Let $\overline{Q}_{tot}^{M,B}$ be the time-average expected number of all requests in the queues under blind multicasting. Then,

the average delay-minimizing choice of the design parameters $(\beta_k)_k$ is given by

$$\beta_k^\star = \frac{\sqrt{\alpha_k}}{\sum_{l=1}^{n}\sqrt{\alpha_i}}, \quad k = 1, \cdots, n. \tag{2}$$

Accordingly, the time-average expected number of requests under this delay-optimal blind multicast strategy is given by

$$\overline{Q}_{tot}^{M,B} = \rho \left(\sum_{i=1}^{n} \sqrt{\alpha_i} \right)^2, \tag{3}$$

which can be written as $\overline{Q}_{tot}^{M,B} = \rho ||\boldsymbol{\alpha}||_{\frac{1}{2}}$.
Note that, even in the worst-case of uniform popularities, we have $\overline{Q}_{tot}^{M,B} = \rho\, n$ under the optimal blind multicast.

B. Delay Gains of Work-Conserving Multicast

More multicast gains can be reaped under smarter policies that schedule services based on the instantaneous state of the queues. In particular, we consider *work-conserving* policies that utilize the BS resources for some pending request(s) unless all the queues are empty. However, due to the analytical complexity under a general popularity distribution $\boldsymbol{\alpha}$, we study the worst case scenario of uniformly distributed popularities that serve as a fundamental lower bound on the performance of a multicasting system besides allowing tractable closed form expressions for the behavior of the average expected number of requests in the system.

Definition 2 (work-conserving Multicast): Define the indicator $\sigma_k^{M,W}(t) \in \{0, 1\}$ to capture the service of queue k at time t such that $\sigma_k^{M,W}(t) = 1$ if the queue k is assigned the service resources at time t, otherwise $\sigma_k^{M,W}(t) = 0$, $k = 1, \cdots, n$. Also, let $Q_k^{M,W}(t)$ be the number of requests in queue k at time t under work-conserving multicasting. Then, a work-conserving multicast strategy is a strategy through which $\sigma_k^{M,W}(t) = 0$ if $Q_k^{M,W}(t) = 0$, and $\sum_k Q_k^{M,W}(t) > 0$ implies that $\sigma_{k^*}^{M,W}(t) = 1$ for some k^* such that $Q_{k^*}^{M,W}(t) > 0$.

We can note from the work-conserving operation that the allocation of the service resources to queues depends on the state of the queue. In this subsection, we consider the well-known *First-Come-First-Serve (FCFS)* work-conserving policy to characterize an upper bound on the average expected number of requests in the system. FCFS operates by serving the *queue* that contains the oldest unserved request first. We choose the FCFS for its time-based ordering of service which enables us to analytically derive our fundamental bound on the system's performance. As such, it possesses fairness characteristics within the class of work-conserving policies. We have the following result.

Theorem 2 (Scaling Delay Gains of Work-Conserving FCFS Multicast): Let $\overline{Q}_{tot}^{M,F}$ be the time-average expected number of all requests for the FCFS work-conserving multicast strategy under the worst-case of uniform popularities, i.e., $\alpha_k = 1/n$ for all k. Then, we have

$$\overline{Q}_{tot}^{M,F} \quad \dot{\leq} \begin{cases} \frac{1}{2}\left(\frac{\rho^2-1}{\rho}\right) n, & \rho > 1, \\ \sqrt{\frac{2}{\pi}} n, & \rho = 1, \\ \min(\frac{\rho}{1-\rho}, \frac{1}{2}\rho^3(1-\rho)^2 n), & \rho < 1, \end{cases} \tag{4}$$

where $a(n) \dot{\leq} b(n)$ means that $\lim\limits_{n\to\infty} \dfrac{a(n)}{b(n)} \leq 1$.

Note that the bound on $\overline{Q}_{tot}^{M,F}$ is directly related to the average delay experienced by the users via Little's law.

C. Discussion of Relevant Insights from the Results

Theorems 1 and 2 reveal the potential for content multicasting to extend the stable operation of the network significantly beyond that of unicasting. In the following remarks, we highlight some insights about those theorems.

Remark 1: Under unicast operation, when $\rho \uparrow 1$, we see from (1) that the average number of requests grows unboundedly, i.e., $\overline{Q}_{tot}^{U} \to \infty$ signifying the instability of unicast as the traffic intensity becomes higher. Theorem 1, on the other hand, shows that blind multicasting guarantees a finite total average of the number of requests for any popularity distribution $\boldsymbol{\alpha}$ and $\rho \geq 0$ as can be seen in (3). Hence, blind multicasting promises *endless stability* operation for any distribution of content popularity and for any number of content items.

Remark 2: Uniform and degenerate distributions of popularity are, respectively, the $\overline{Q}_{tot}^{M,B}$ maximizing and minimizing distributions. This can be seen by optimizing (3) for the maximum and minimum values over $\boldsymbol{\alpha}$, where the maximum value for $\overline{Q}_{tot}^{M,B}$ is $\rho\, n$ and the minimum value is ρ.

Intuitively, uniformly distributed popularities maximize the average number of distinct data items being requested in the system irrespective of the multicasting scheduling policy. Hence, more requests on average require individual service than under any popularity distribution. The degenerate distribution, on the other hand, implies that all of the incoming traffic is targeting the same data item. Hence, multicasting operation will reap the highest gains.

Note that, using (3), we can also find the delay performance of the optimal blind multicast strategy under more common popularity distributions, such as the Zipf distribution. In particular: a Zipf distribution with parameter $\gamma = 2$, the total queue-length $\overline{Q}_{tot}^{M,B}$ under blind multicast scales as $O(\rho\,(\log(n))^2)$; while Zipf distribution with parameter $\gamma \in (0, 2)$, it scales as $O(\rho\, n^{\min\{2-\gamma,1\}})$. We omit the details of these results due to limited space.

Remark 3: Note that the result of (1) is obtained assuming work-conserving *unicast* operation. For stable operation, i.e., $\rho < 1$, we see that $\overline{Q}_{tot}^{U}$ is independent of the number of content items n irrespective of their popularity distribution. This is not the case under blind multicast operation for the same range of $\rho < 1$ where $\overline{Q}_{tot}^{M,B}$ is determined by both n and $\boldsymbol{\alpha}$. In fact, for large values of n, and several distributions,

e.g., Zipf with $\gamma \leq 2$, we have $\overline{Q}_{tot}^{U} \leq \overline{Q}_{tot}^{M,B}$. Thus, unicast outperforms blind multicast for $\rho < 1$ in this case.

Remark 4: Assume uniform distribution of popularities and $\rho > 1$. As $n \to \infty$, the average expected number of requests per queue under multicast operation satisfies

$$\lim_{n\to\infty} \frac{\overline{Q}_{tot}^{M,o}}{n} \begin{cases} \leq \frac{1}{2}(\frac{\rho^2-1}{\rho}), & o = F, \\ = \rho, & o = B. \end{cases}$$

That is, FCFS work-conserving multicasting attains at most an expected value of $\frac{1}{2}(\frac{\rho^2-1}{\rho})$ requests per queue while blind multicasting attains ρ. Thus, FCFS experiences at most half the delay of blind multicasting for $\rho > 1$.

Remark 5: Our analysis reveals important practical insights that, while work-conserving multicast always outperforms unicast and blind-multicast: (i) unicast strategy can be sufficiently satisfactory under lightly-loaded conditions, i.e., when $\rho << 1$; and (ii) blind-multicast strategy tends to suffer a delay performance loss within a factor of 2 under over-loaded conditions, i.e., when $\rho >> 1$. The gains of work-conserving multicasting is highest in the regime (that is explicitly characterized by our analysis in terms of ρ and n) where the loading factor is neither too small, nor too large.

IV. Proofs of the Stability and Delay Gain Results

In this section, we provide the full proofs of the main results discussed in the previous section. The proof of Theorem 1 (in Section IV-A) is based on decomposing the system into parallel queues to optimize the delay. However, the proof of Theorem 2 (in Section IV-B) requires a much more sophisticated strategy due to the coupling between the queues and their nontraditional dynamics.

A. Endless Stability of Blind Multicast (Theorem 1)

We start by obtaining the expected queue-length under the blind multicast operation with a general $(\beta_k)_k$ choice.

Lemma 1: Let $Q_k^{M,B}(t)$ be the number of requests in queue k under blind multicast operation, then

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbb{E}[Q_k^{M,B}(t)]dt = \rho \frac{\alpha_k}{\beta_k}. \tag{5}$$

Proof. For a queue with input rate $\alpha_k\lambda$ and service rate $\beta_k\mu$ using the multicast operation, when a new request arrives, number of requests increases by one but when there is a service available for queue k, because of the multicast nature, after serving the total number of requests in queue k becomes 0. Markov chain for queue k is shown in Fig. 2. The average number of requests in the system is given by:

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbb{E}[Q_k^{M,B}(t)]dt = \sum_{m=0}^{\infty} m p_m. \tag{6}$$

Which p_m is the probability of having m requests in queue k. Using the markov chain and by induction we have:

$$p_m = \frac{\mu\beta_k}{\lambda\alpha_k + \mu\beta_k} \left(\frac{\lambda\alpha_k}{\lambda\alpha_k + \mu\beta_k}\right)^m.$$

Substituting p_m in equation 6 and using the definition of loading factor $\rho = \frac{\lambda}{\mu}$, we have:

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbb{E}[Q_k^{M,B}(t)]dt = \frac{\alpha_k\lambda}{\beta_k\mu} = \rho\frac{\alpha_k}{\beta_k}.$$

■

We thus have $\overline{Q}_{tot}^{M,B} = \rho \sum_{k=1}^{n} \frac{\alpha_k}{\beta_k}$. Noting the convexity of this expression with respect to $(\beta_k)_k$, we use the KKT optimality conditions to find that the choice of $\beta_k^\star$ in (2) minimizes $\overline{Q}_{tot}^{M,B}$ subject to the constraints that $\beta_k \geq 0$, $\forall k$, and $\sum_{k=1}^{n} \beta_k = 1$. Finally, the direct substitution of $\beta_k^\star = \frac{\sqrt{\alpha_k}}{\sum_{l=1}^{n}\sqrt{\alpha_i}}$ in (5) completes the proof of Theorem 1.

B. Delay Gains of Work-Conserving Multicast (Theorem 2)

Traditional queuing dynamics, under which requests are served one by one, have been investigated in various works (see [9] for a survey). However, in our multicasting scenario, due to the service of all pending demands at once, previous well-known techniques such as Lyapunov-drift [10] or fluid-limit [11] analysis techniques do not apply. In order to analyze and prove the results of multicasting systems, we take a different approach based on the number of *active queues* defined next.

Definition 3 (Active Queue): We define an active queue as a nonempty queue, i.e., a queue that has at least one request in it. Formally, queue k is **active** at time t, if $Q_k^{M,W}(t) > 0$.

Utilizing the statistics of active queues, we derive a novel upper bound for the average number of requests in the system. Each proof is broken down into pieces in order to facilitate the understanding. Some of the results in these proofs may be of independent-interest, especially in the case of Theorem 2.

Let $N(t)$ be the Markov process giving the number of active queues at time t under any given work-conserving multicast strategy. The evolution of this process is shown in Fig. 3. We are interested in limit of $N(t) \xrightarrow[t\to\infty]{d} \bar{N}(\rho, n)$, which is characterized next and studied subsequently[3].

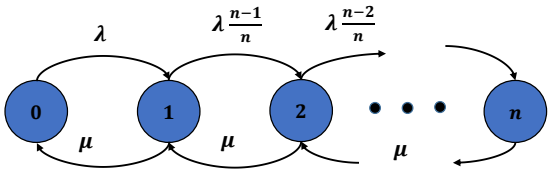


Fig. 3: Markov chain for active queues $N(t)$ under any work-conserving multicast

Lemma 2: Let $\pi_k = P(\bar{N}(\rho, n) = k)$ be the probability of having k active queues under work-conserving multicast operation, then

$$\pi_k = \pi_0 \prod_{m=0}^{k-1} (1 - \frac{m}{n})\rho, \tag{7}$$

[3]Such a steady state behavior holds since $N(t), t \geq 0$ follows a finite state ergodic Markov chain.

where $\pi_0 = \left(\sum_{k=0}^{n} \prod_{m=0}^{k-1}(1-\frac{m}{n})\rho\right)^{-1}$.

Proof. Global balance equations of Fig. 3 yields

$$\pi_k = (\frac{\lambda}{\mu})^k.1.(1-\frac{1}{n})(1-\frac{2}{n})...(1-\frac{k-1}{n})\pi_0$$
$$\pi_0(\frac{\lambda}{\mu})^k \prod_{m=0}^{k-1}(1-\frac{m}{n}) = \pi_0 \prod_{m=0}^{k-1}((1-\frac{m}{n})(\frac{\lambda}{\mu})).$$

Replacing $\frac{\lambda}{\mu}$ with loading factor ρ gives (7). Then, setting the sum of probabilities to 1 gives the result for π_0. ■

We introduce a new parameter $s_i(\rho, n)$ as:

$$s_i(\rho, n) := \sum_{k=1}^{n} k^i \prod_{m=0}^{k-1}(1-\frac{m}{n})\rho, \tag{8}$$

which we use in deriving the moments of $\bar{N}(\rho, n)$. In the light of $s_i(\rho, n)$, we can rewrite the π_0 as $\pi_0 = \frac{1}{1+s_0(\rho,n)}$. We also make the following connection between $s_i(\rho, n)$ and $s_{i-1}(\rho, n)$.

Lemma 3: For $s_i(\rho, n)$ defined in (8),

$$s_i(\rho, n) = n(1-1/\rho)s_{i-1}(\rho, n) + \frac{n}{\rho}\sum_{m=2}^{i}(-1)^m \binom{i-1}{m-1} s_{i-m}(\rho, n) + n\delta(i-1) \tag{9}$$

such that $i \in \{1, 2, ...\}$.

Proof. We prove this by induction.

$$\begin{aligned} s_i(\rho, n) &- n(1-1/\rho)s_{i-1}(\rho, n) \\ &+ \frac{n}{\rho}\sum_{q=1}^{i-1}(-1)^q \binom{i-1}{q} s_{i-1-q}(\rho, n) \\ &= \sum_{k=1}^{n}\Bigg[k^i - n(1-1/\rho)k^{i-1} \\ &+ \frac{n}{\rho}\sum_{q=1}^{i-1}\binom{i-1}{q}(-1)^q k^{i-1-q}\Bigg]\prod_{m=0}^{k-1}(1-\frac{m}{n})\rho. \end{aligned} \tag{10}$$

We have n terms on the right hand side of (10), by induction we can show that sum of p last terms is equal to:

$$\frac{n}{\rho}(n-p)^{i-1}\prod_{m=0}^{n-p}(1-\frac{m}{n})\rho.$$

Since we have n total terms, putting $p = n$ gives $\frac{n}{\rho}(n-n)^{i-1}\rho = n\delta(i-1)$. ■

Lemma 4: The first and second moments of the number of active queues, $\bar{N}(\rho, n)$, are given by:

$$E[\bar{N}(\rho, n)] = n(1-\frac{1}{\rho})\frac{s_0(\rho,n)}{1+s_0(\rho,n)} + \frac{n}{1+s_0(\rho,n)}, \tag{11}$$

$$E[\bar{N}(\rho, n)^2] = (n^2(1-\frac{1}{\rho})^2 + \frac{n}{\rho})\frac{s_0(\rho,n)}{1+s_0(\rho,n)}. \tag{12}$$

Proof.

$$E[\bar{N}(\rho, n))] = \pi_0 \sum_{k=1}^{n} k \prod_{m=0}^{k-1}(1-\frac{m}{n})\rho = \frac{s_1(\rho,n)}{1+s_0(\rho,n)}.$$

From Lemma 3, writing $s_1(\rho, n)$ as a function of $s_0(\rho, n)$ gives Equation 11. Similarly,

$$E[\bar{N}(\rho, n)^2] = \pi_0 \sum_{k=1}^{n} k^2 \prod_{m=0}^{k-1}(1-\frac{m}{n})\rho = \frac{s_2(\rho,n)}{1+s_0(\rho,n)}.$$

Writing $s_2(\rho, n)$ in terms of $s_0(\rho, n)$ gives the result in Equation (12). ■

Lemma 5: For $\rho = 1$, $s_0(\rho, n)$ asymptotically achieves

$$s_0(1, n) \doteq \sqrt{\frac{\pi}{2}n}, \tag{13}$$

where $a(n) \doteq b(n)$ means $\lim_{n\to\infty}\frac{a(n)}{b(n)} = 1$.

Proof. For $\rho = 1$, $s_0(1, n) = \sum_{k=1}^{n}\prod_{m=0}^{k-1}(1-\frac{m}{n})$. Rewriting and changing the variable $j = n - k$ gives:

$$s_0(1, n) = \frac{n!}{n^n}\sum_{j=0}^{n-1}\frac{n^j}{j!}$$

Now by using the fact that $\sum_{j=0}^{n-1}\frac{x^j}{j!} = e^x\frac{\Gamma(n,x)}{\Gamma(n)}$, such that $\Gamma(n, x) = \int_x^\infty t^{n-1}e^{-t}dt$ and $\Gamma(n) = \Gamma(n, 0)$ [12], we can rewrite $s_0(1, n)$ as:

$$s_0(1, n) = \frac{n!}{n^n}e^n\frac{\Gamma(n, n)}{\Gamma(n)}. \tag{14}$$

Since from [13], $\lim_{n\to\infty}\frac{\Gamma(n,n)}{\Gamma(n)} = \frac{1}{2}$, and utilizing the asymptotic behavior of Stirling's approximation, we obtain

$$s_0(1, n) \doteq \sqrt{2\pi n}(\frac{n}{e})^n\frac{1}{2}e^n = \sqrt{\frac{\pi}{2}n}.$$

■

Lemma 6: Let $f(\rho, n) = \frac{s_0(\rho,n)}{1+s_0(\rho,n)}$, then:

$$\lim_{n\to\infty} f(\rho, n) = \begin{cases}\rho, & \rho < 1,\\ 1, & \rho > 1,\end{cases} \tag{15}$$

Proof. First we show that for $\rho < 1$, $\lim_{n\to\infty} f(\rho, n) = \rho$.

$$\begin{aligned} s_0(\rho, n) &= \sum_{k=1}^{n}\prod_{m=0}^{k-1}(1-\frac{m}{n})\rho = \sum_{k=1}^{n}\rho^k\prod_{m=0}^{k-1}(1-\frac{m}{n}) \\ &\le \sum_{k=1}^{n}\rho^k = \frac{\rho}{1-\rho}(1-\rho^n) \end{aligned} \tag{16}$$

On the other hand we have:

$$\begin{aligned} s_0(\rho, n) &\geqslant \sum_{k=1}^{\sqrt{n}}\rho^k\prod_{m=0}^{k-1}(1-\frac{m}{n}) \\ &\geqslant \sum_{k=1}^{\sqrt{n}}\rho^k(1-\frac{k-1}{n})^k \geqslant \sum_{k=1}^{\sqrt{n}}\rho^k(1-\frac{\sqrt{n}-1}{n})^k \\ &\geqslant \frac{\rho(1-\frac{\sqrt{n}-1}{n})}{1-\rho(1-\frac{\sqrt{n}-1}{n})} \times (1-\rho^{\sqrt{n}}(1-\frac{\sqrt{n}-1}{n})^{\sqrt{n}}). \end{aligned} \tag{17}$$

From (16) and (17), and letting $n \to \infty$, we have:

$$\frac{\rho}{1-\rho} \leq \lim_{n\to\infty} s_0(\rho, n) \leq \frac{\rho}{1-\rho}$$

This proves the fact that $\lim_{n\to\infty} s_0(\rho, n) = \frac{\rho}{1-\rho}$. By using the definition of $f(\rho, n)$, we have that $\lim_{n\to\infty} f(\rho, n) = \rho$.

In order to prove the $\lim_{n\to\infty} f(\rho, n) = 1$ for $\rho > 1$, we show that for any $\rho > 1$, $s_0(\rho, n)$ grows exponentially in n.

$$\begin{aligned} s_0(\rho, n) &= \sum_{k=1}^{n} \prod_{m=0}^{k-1} (1 - \frac{m}{n})\rho = \sum_{k=1}^{n} \rho^k \prod_{m=0}^{k-1} (1 - \frac{m}{n}) \\ &= \frac{n!}{(\frac{n}{\rho})^n} \sum_{k=1}^{n} \frac{(\frac{n}{\rho})^{n-k}}{(n-k)!} = \frac{n!}{(\frac{n}{\rho})^n} \sum_{j=0}^{n-1} \frac{(\frac{n}{\rho})^j}{j!} \end{aligned} \tag{18}$$

Now by the fact that $\sum_{j=0}^{n-1} \frac{(\frac{n}{\rho})^j}{j!} = e^{\frac{n}{\rho}} \frac{\Gamma(n, \frac{n}{\rho})}{\Gamma(n)}$, we can rewrite $s_0(\rho, n)$ as

$$s_0(\rho, n) = \frac{n!}{(\frac{n}{\rho})^n} e^{\frac{n}{\rho}} \frac{\Gamma(n, \frac{n}{\rho})}{\Gamma(n)}.$$

For $\rho > 1$, we have $\Gamma(n, \frac{n}{\rho}) > \Gamma(n, n)$, which implies:

$$\lim_{n\to\infty} \frac{\Gamma(n, \frac{n}{\rho})}{\Gamma(n)} > \lim_{n\to\infty} \frac{\Gamma(n, n)}{\Gamma(n)} = \frac{1}{2},$$

and applying Stirling's Inequality $n! \geq \sqrt{2\pi n}(\frac{n}{e})^n$ yields

$$s_0(\rho, n) \geq \sqrt{\frac{\pi n}{2}} (\frac{\rho}{e})^n e^{\frac{n}{\rho}} = \sqrt{\frac{\pi n}{2}} e^{n(\frac{1}{\rho} + \log\rho - 1)}.$$

Setting $g(\rho) := \frac{1}{\rho} + \log\rho - 1$, since $g(1) = 1$ and $g'(\rho) > 0$ for $\rho > 1$, $s_0(\rho, n)$ grows exponentially in n for $\rho > 1$. ■

After having introduced a number of crucial lemmas, we proceed to our investigation of the number of requests in the system under work-conserving multicasting. To that end, we focus on the FCFS strategy.

Let $Q_k^{M,F}(t)$ be the number of requests in queue k at time t under FCFS work-conserving multicasting and $Q_{tot}^{M,F}(t) := \sum_{k=1}^{n} Q_k^{M,F}(t)$ be the aggregate number of requests in all queues at time t. The following lemma characterizes an upper bound for the average total number of requests in the system under FCFS multicast operation.

Lemma 7: Let

$$\overline{Q}_{tot}^{M,F} := \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbb{E}[Q^{M,F}(t)]dt,$$

be the time-average expected number of aggregate requests in the system operating under FCFS multicasting, then

$$\overline{Q}_{tot}^{M,F} \leq n \frac{\frac{\rho}{2n}\mathbb{E}[\bar{N}(\rho, n)^2] + (1 + \frac{\rho}{n})\mathbb{E}[\bar{N}(\rho, n)] + \frac{\rho}{n} + 1}{\mathbb{E}[\bar{N}(\rho, n)] + \frac{n}{\rho} + 1}. \tag{19}$$

Proof. Since the request arrivals and services are statistically indistinguishable across the n queues, we have under steady state operation that $\mathbb{E}[Q_k^{M,F}(t)] = \mathbb{E}[Q_l^{M,F}(t)]$, $\forall k, l$. Thus, it suffices to study $\mathbb{E}[Q_k^{M,F}(t)]$ and then obtain $\overline{Q}_{tot}^{M,F} = n\overline{Q}_k^{M,F}$, where

$$\overline{Q}_k^{M,F} := \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbb{E}[Q_k^{M,F}(t)]dt.$$

Let $t_0 = 0$ and t_i be the time instant at which queue k has completed service for the i^{th} time. So at time instants $\{t_i\}_i$, we will have:

$$Q_k^{M,F}(t_i) = 0, \text{ and } Q_k^{M,F}(t_i - \epsilon) > 0, \quad i = 0, 1, \cdots,$$

for some $0 < \epsilon < t_i - t_{i-1}$. Let X_i be the time it takes queue k to become active for the $(i+1)^{th}$ time since it has been last served (emptied) at time t_i. So $Q_k^{M,F}(t_i + X_i) = 1$ and $Q_k^{M,F}(t_i + X_i - s) = 0$, $\forall s \in (0, t_i]$. Since the popularity distribution is uniform, X_i follows exponential distribution with mean $\frac{n}{\lambda}$. Let $N_k(t)$ be the number of active queues in the system, given that queue k has just become active at time t. We should note that $N_k(t)$ includes queue k itself. That is, $N_k(t) \geq 1$. Let τ_i denote the duration queue k must wait while being active in order to be fully serviced for the $(i+1)^{th}$ time. So $t_{i+1} = t_i + X_i + \tau_i$. We define $T_i := X_i + \tau_i$. Given $N_k(t) = v, \tau_i = \sum_{j=1}^{v} Y_j$, where Y_j is the service time of an active queue which has an exponential distribution with mean $\frac{1}{\mu}$. We are now interested in the time-average value of expected number of requests in queue k, $\overline{Q}_k^{M,F}$, where

$$\overline{Q}_k^{M,F} = \lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbb{E}[Q_k^{M,F}(t)]dt,$$

Let $K_T = \max\{i \geq 0 | t_i \leq T\}$, which is the number of times that queue k has received service by time T. We have:

$$\begin{aligned} \int_0^T \mathbb{E}[Q_k^{M,F}(t)]dt &= \sum_{i=0}^{K_t - 1} \int_{t_i}^{t_i+1} \mathbb{E}[Q_k^{M,F}(t)]dt \\ &\quad + \int_{t_{K_T}}^{T} \mathbb{E}[Q_k^{M,F}(t)] \leq \sum_{i=0}^{K_T} M_k[i], \end{aligned}$$

where $M_k[i] = \int_{t_i}^{t_i+1} E[Q_k^F(t)]dt$ and $\{M_k[i]\}_i$ is an identically distributed sequence of random variables. Then:

$$\frac{1}{T} \int_0^T \mathbb{E}[Q_k^{M,F}(t)]dt \leq \frac{1}{T} \sum_{i=0}^{K_T} M_k[i] \leq \frac{K_T}{T} \frac{1}{K_T} \sum_{i=0}^{K_T} M_k[i],$$

We note that $K_T \to \infty$ and $\frac{K_T}{T} \to \frac{1}{E[T_i]}$ as $T \to \infty$, both with probability 1. We thus have

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbb{E}[Q_k^{M,F}(t)]dt \leq \frac{1}{E[T_i]} E[M_k[i]],$$

For $E[M_k[i]]$, we have:

$$\begin{aligned} &E[M_k[i]] = \mathbb{E}[\int_{t_i}^{t_i+1} \mathbb{E}[Q_k^F(t)]dt] = \\ &\mathbb{E}_{X_i,\tau_i}\left[\tau + \int_{t_i+x}^{t_i+x+\tau} \int_{t_i+x}^{t_i+x+\tau} \mathbb{E}[A_k(t)]dldt | x_i = x, \tau_i = \tau\right] \\ &= \mathbb{E}_{X_i,\tau_i}\left[\tau + \frac{\lambda}{2n}\tau^2\right] = \frac{\mathbb{E}[N_k(t)]}{\mu} + \frac{\lambda}{2n\mu^2}\mathbb{E}[N_k(t)^2], \end{aligned}$$

where $A_k(t)$ is the arrival process to queue k at time t. We thus obtain

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbb{E}[Q_k^{M,F}(t)]dt \le \frac{\frac{\mathbb{E}[N_k(t)]}{\mu}+\frac{\lambda}{2n\mu^2}\mathbb{E}[N_k(t)^2]}{\frac{n}{\lambda}+\frac{\mathbb{E}[N_k(t)]}{\mu}}.$$

Since $N_k(t)$ is the number of active queue in the system when queue k turns active at time t, we have $\mathbb{E}[N_k(t)] \le \mathbb{E}[\bar{N}(\rho,n)]+1$. Substituting in the previous inequality, we get

$$\overline{Q}_k^{M,F} \le \frac{\frac{\mathbb{E}[\bar{N}(\rho,n)]}{\mu}+\frac{1}{\mu}+\frac{\lambda(\mathbb{E}[\bar{N}(\rho,n)^2]+2\mathbb{E}[\bar{N}(\rho,n)]+1)}{2n\mu^2}}{\frac{n}{\lambda}+\frac{\mathbb{E}[\bar{N}(\rho,n)]}{\mu}+\frac{1}{\mu}}.$$

By multiplying with n and substituting $\frac{\lambda}{\mu}$ with loading factor ρ, we will have the result. ■

Lemma 8: For $f(\rho,n)$ defined in Lemma 6,

$$\overline{Q}_{tot}^{M,F} \le \frac{n(\frac{\rho^2-1}{2\rho}f(\rho,n)+1-f(\rho,n))+(\rho+1-\frac{f(\rho,n)}{2})+\frac{\rho}{n}}{(1+\frac{1-f(\rho,n)}{\rho})+\frac{1}{n}}. \tag{20}$$

Proof. From Lemmas 4 and 6, we can write

$$E[\bar{N}(\rho,n)] = n\left(1-f(\rho,n)/\rho\right),$$
$$E[\bar{N}(\rho,n)^2] = \left(n^2(1-1/\rho)^2+n/\rho\right)f(\rho,n).$$

By direct substitution in (19) we have the result. ■

Corollary 1: For large enough n, the upper bound of (20) can be approximated as

$$\overline{Q}_{tot}^{M,F} \stackrel{\cdot}{\le} n\,\frac{\frac{1}{2}\frac{\rho^2-1}{\rho}f(\rho,n)+1-f(\rho,n)}{1+\frac{1-f(\rho,n)}{\rho}}. \tag{21}$$

At this point, by substituting from Lemmas 5 and 6, in Corollary 1, we conclude the proof of Theorem 2.

V. Numerical Results

The analytical results obtained in this paper are validated through numerical simulations in this section. Each of the following simulation results is an average behavior over 10^6 iterations. We first validate the main analytical results under uniform popularity, and then provide more numerical results for non-uniform popularity distributions (such as commonly used Zipf distribution). Moreover, we compare the performance of FCFS work-conserving policy to a heuristic *Max-Weight* work-conserving multicast policy that is expected to yield favorable delay-minimization merits.

A. Validation of Main Results under Uniform Popularities

In Fig. 4, we provide a numerical evaluation of $\overline{Q}_{tot}^{U}$ and $\overline{Q}_{tot}^{M,B}$ under different content popularity distributions for $n = 1000$. For degenerate distribution of content popularity, $\overline{Q}_{tot}^{M,B}$ is equal to ρ which is the minimum that blind multicast can achieve for given n and ρ. On the other hand, uniform distribution of content popularity gives the maximum value of $\overline{Q}_{tot}^{M,B}$ which equals $n\rho$. It is obvious from the figure that as ρ approaches 1, unicast system becomes unstable, while blind multicasting operation guarantees a finite total average number of requests upper bounded by $n\,\rho$ for any popularity distribution α and $\rho \ge 0$ as can be seen in Fig. 4. We can also observe from Fig. 4 that unicast outperforms blind multicast under the considered instances of Zipf distributed popularity which is consistent with the insights of Remark 3.

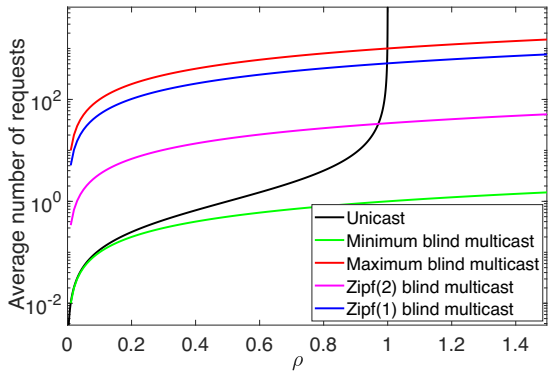


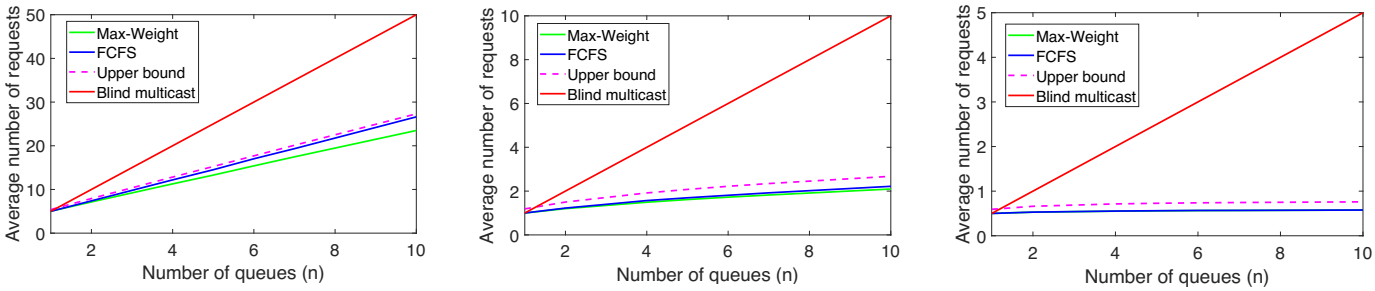
Fig. 4: Comparison between unicast and blind multicast for $n = 1000$ items.

Fig. 5 shows the average number of requests as a function of number of queues in a system with uniform distributions of content popularity under different values of loading factor ρ and scheduling policies. We can see that for different levels of the loading factor ρ, the FCFS multicast policy performs very close to the heuristic Max-Weight that serves a queue with the largest number of requests at the time of service, both of which outperform blind multicasting by a large margin. Also, we can see that our upper-bound is very accurate for different levels of loading factor ρ.

B. Performance Comparison against Max-Weight multicast for Uniform and Non-uniform Popularities

In this section, we aim to investigate the performance of our blind and work-conserving multicast strategies under non-uniform popularities and compare their performance to a heuristic Max-Weight multicast policy. Fig. 6 shows the total average number of requests in the system for different policies under the uniform popularity distribution. As it can be seen in this figure, the upper bound we derived for FCFS is very tight. Moreover, performance of FCFS is very close to that of Max-Weight. According to Fig. 6, for small loading factor ρ, unicast performance is very close to work-conserving multicast performance and it is much better than the blind multicast performance. For ρ close to 1, when the unicast becomes unstable, work-conserving multicast become substantially efficient compared to both unicast and blind multicast. For $\rho >> 1$, work-conserving multicast still outperforms blind multicast by a factor of 2 as it has been noted in Remark 4.

Fig. 7 shows the total average number of requests in the system for different policies under Zipf popularity distribution with parameter $\gamma = 1.2$. As it can be seen from the figure, performance of FCFS is very close to Max-Weight and our upper bound which we derived for FCFS in Theorem 2 under uniform popularity distribution, is also reasonable for non-uniform popularity distributions like Zipf.



(a) $\rho = 5$. (b) $\rho = 1$. (c) $\rho = 1/2$.

Fig. 5: Average number of requests in the system over n for different loading factors ρ.

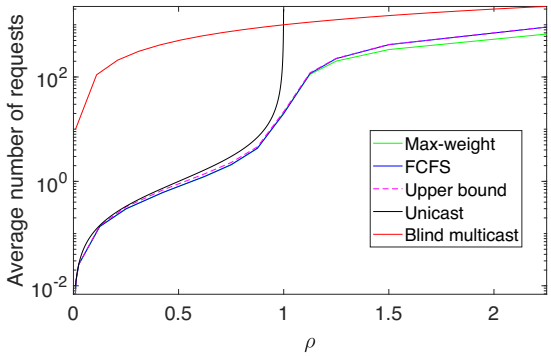


Fig. 6: Number of requests in the system for different policies under uniform popularity distribution and n=1000.

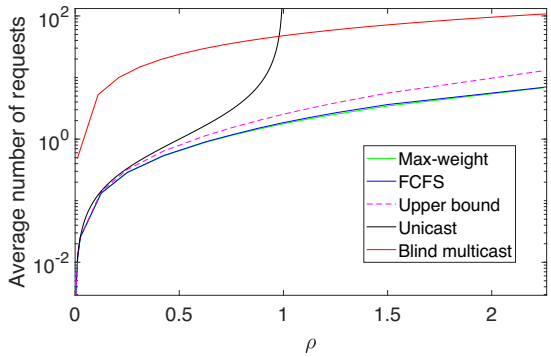


Fig. 7: Number of requests in the system for different policies under Zipf(1.2) popularity distribution and n=1000.

VI. Conclusion

In this work, we provided a comprehensive analysis of multicast gains for wireless content distribution networks serving a dynamic population of users that aim to access a content database with a given popularity distribution. In particular, we characterized the delay performance of two classes of multicasting strategies, namely, 'blind' multicasting whereby the pending requests are unknown to the transmitter, and 'work-conserving' multicasting whereby the pending requests are known. Our results establish that both types of multicasting yields *endless stability*, in that an unbounded traffic load can be supported by them by exploiting the multicast advantage of wireless communication. This is in contrast to the bounded stability of unicast mode of transmission whereby requests are fulfilled individually. Moreover, we show that work-conserving multicast based on a first-come-first-serve principle can yield further delay gains over its blind counterpart that are explicitly characterized in our analysis as a function of the traffic load and the database size. In addition to the explicit characterization of delay performance of these proposed multicast strategies, our work also revealed key insights on the conditions under which blind and work-conserving multicast solutions can yield most benefit.

References

[1] X. Kang, Y.-K. Chia, S. Sun, and H. F. Chong, "Mobile data offloading through a third-party wifi access point: An operator's perspective," *IEEE Transactions on Wireless Communications*, vol. 13, no. 10, pp. 5340–5351, 2014.

[2] J. Lee, Y. Yi, S. Chong, and Y. Jin, "Economics of wifi offloading: Trading delay for cellular capacity," *IEEE Transactions on Wireless Communications*, vol. 13, no. 3, pp. 1540–1554, 2014.

[3] J. Tadrous and A. Eryilmaz, "On optimal proactive caching for mobile networks with demand uncertainties," *IEEE/ACM Transactions on Networking*, vol. 24, no. 5, pp. 2715–2727, 2016.

[4] Y. Yasuda, S. Ata, and I. Oka, "Proactive cache management method for content hash based distributed archive system," in *IEEE International Conference on Information Networking (ICOIN)*, 2013, pp. 456–461.

[5] E. Baştuğ, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5g wireless networks," *arXiv preprint arXiv:1405.5974*, 2014.

[6] L. Al-Kanj, Z. Dawy, and E. Yaacoub, "Energy-aware cooperative content distribution over wireless networks: Design alternatives and implementation aspects." *IEEE Communications Surveys and Tutorials*, vol. 15, no. 4, pp. 1736–1760, 2013.

[7] G. M. De Brito, P. B. Velloso, and I. M. Moraes, *Information-centric Networks: A New Paradigm for the Internet*. John Wiley & Sons, 2013.

[8] L. Kleinrock, *Queueing systems, Volume 1: Theory*. Wiley New York, 1975.

[9] Y. Cui, V. K. Lau, R. Wang, H. Huang, and S. Zhang, "A survey on delay-aware resource control for wireless systemslarge deviation theory, stochastic lyapunov drift, and distributed stochastic learning," *IEEE Transactions on Information Theory*, vol. 58, no. 3, pp. 1677–1701, 2012.

[10] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.

[11] J. G. Dai, "On positive harris recurrence of multiclass queueing networks: a unified approach via fluid limit models," *The Annals of Applied Probability*, pp. 49–77, 1995.

[12] G. Jameson, "The incomplete gamma functions," *The Mathematical Gazette*, vol. 100, no. 548, pp. 298–306, 2016.

[13] W. Chojnacki, "Some monotonicity and limit results for the regularised incomplete gamma function," in *Annales Polonici Mathematici*, vol. 94, no. 3, 2008, pp. 283–291.