Delay Gain Analysis of Wireless Multicasting for Content Distribution

Bahman Abolhassani, John Tadrous, Atilla Eryilmaz

***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.

To tackle this problem, several techniques have already been proposed such as WiFi offloading, proactive caching, and wireless multicasting. WiFi offloading is a straightforward approach that communicates some of the wireless cellulars data through WiFi networks (e.g., [1]). Different approaches to implement WiFi offloading and to improve its performance 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. Although arrival requests can be served by sending a separate unicast packet to each user, this approach suffers from poor performance. The situation is especially acute in delay tolerant networks (DTNs) [6].

To illustrate the potential gains of multicasting with an example, consider a football stadium full of people watching a game and after a goal, many of them may request (at different time offsets) the related footage to watch it on their smart device, giving the opportunity to broadcast content of common interest to multiple users with small delay, since in a short period of time, there will be a lot of requests for one content. There are so many similar real world scenarios where a large group of users are interested in a certain group of content during a small time window. These are the scenarios that potentially can have great delay gains using the proposed model. Since by utilizing the multicast nature of wireless communication on the edge, instead of sending

Manuscript received January 19, 2020; revised April 21, 2020 and August 7, 2020; accepted November 12, 2020; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor H. Jiang. Date of publication –, 2020; date of current version November 17, 2020. This research is funded primarily by the ONR Grant N00014-19-1-2621, and the NSF grants CNS-NeTS-2007231, CNS-ICN-WEN-1719371, and, in part by: NSF Grants: CNS-SpecEES-1824337, CNS-NeTS-1717045; and the DTRA grant: HDTRA1-18-1-0050.

B. Abolhassani and A. Eryilmaz are with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail:abolhassani.2@osu.edu; eryilmaz.2@osu.edu).

J. Tadrous is with the Department of Electrical and Computer Engineering, Gonzaga University, Spokane, WA 99202 (e-mail:tadrous@gonzaga.edu).

same content to multiple users separately, we can collect the requests of the content and then broadcast it using one service of that content over the wireless medium.

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. Our contributions, along with the organization of the paper, are as follows.

- In Section III, we present a tractable content distribution model for serving dynamically arriving demand over wireless broadcast channels.
- In Section IV-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 V-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 IV-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 V-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 VI, 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 VII. In the next section we provide a literature review of all the related works on multicast networks.

II. Related Work

There are massive amount of works in liturature that study the multicast in wireless networks [7], [8] and [9]. Some of these works focus on the delay performance of multicasting. In [10], [11] and [12], authors study the problem of multicasting in delay tolerant networks. In multi-hop wireless network, [13] and [14] show that cross-layer cooperation of different network layers is needed to efficiently utilize network resources. In [15], by incorporating delay differentiation into cross-layer framework, authors propose a novel Cross-Layer Control algorithm (CLC-DD) that takes into account different delay requirements of flows. The main idea of the proposed algorithm is to distribute delays among flows to achieve low delays for delay-sensitive flows at the expense of increasing the delays of other flows, while simultaneously guaranteeing maximum network utility. In [16] and [17], authors use network coding for reducing transmission delay of large files in multicasts. In [18], using the network coding approach, authors analyze the delay performance in multicasting systems and show that delay can be minimized by appropriate scheduling of data packets and appropriate size of the coding buffer. In [19], authors propose an efficient framework to model the statistical delay QoS guarantees and develop a set of optimal adaptive transmission schemes to minimize the resource consumption while satisfying the diverse QoS requirements under various scenarios, including video unicast/multicast. Traditional solution on multicasting over IP-based network rely on IP multicast which suffer from poor congestion control, as well as slow and complex group membership and multicast tree management on the control plane [20]. Even though multicasting is widely acknowledged to be a promising approach in IP-based wireless networks, the complex dynamic multicast tree building and maintenance, specially for large database sizes, increases the latency and have caused most network operators to eschew its use [21] and [22]. As datacenter size continues to grow, one approach is to deploy a high bandwidth network core for datacenters using optical communication technologies [23]. In [24], authors propose HyperOptics, a low latency optical multicast architecture for datacenters which eliminates the reconfiguration delay by using optical switches.

Transitioning from IP based networks to information-centric networks (see [25] and [26]) encourages us to rigorously investigate the multicasting gain in information-centric networks. Such networks allow us to group requests targeting the same content and serve all of them at once, eliminating the need for dynamic multicast tree building and maintenance. To achieve this multicasting gain, some requests will not be served instantly. In other words, requests from different users for the same content do not happen at the exact same time. Users with earlier requests have to wait until the content is scheduled for service. This introduces the delay which each incoming request needs to sustain before it can be served by database.

In real world scenarios, usually, requests are correlated and this will help the multicast to potentially have great delay gains by utilizing the multicast nature of the wireless communication which is already available on the edge. Base station will collect the requests and put them in dedicated queues to be broadcasted upon the availability of the service. Delay gains in queuing theory are well known and traditional queuing dynamics, under which requests are served one by one [27], have been investigated in various works (see [28] 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 [29] or fluid-limit [30] analysis techniques do not apply. We aim to reveal the stability conditions and the delay gains that multicasting can offer over its unicast counterpart using queuing dynamics. The multicasting scenario introduces new challenges which have not been studied before and to the best of our knowl-

edge, this is the first work to study the delay gains of work-conserving multicasting using queuing theory. In order to analyze the delay performance of work-conserving multicast in information-centric networks, we take a different novel approach based on the number of *active queues*. This work extends [31] to improve the bounds as well as obtain an exact asymptotic expression with numerical simulations.

III. 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, i.e., a single-server system. 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 a reasonable assumption.

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

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 $B_{tot}(t) := \sum_{k=1}^{n} B_k(t)$.

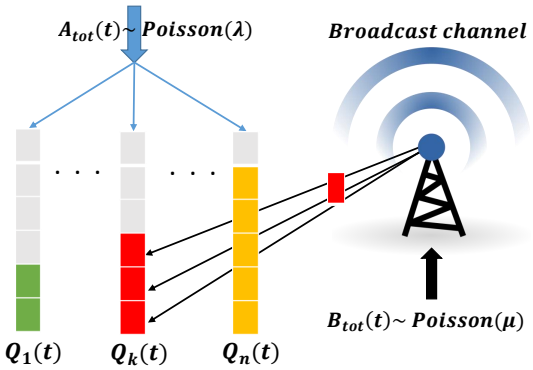


Fig. 1: Queuing system model

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 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 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.

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

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 $Q^U_{tot}(t)$ and $Q^M_{tot}(t)$, respectively, where $Q^o_{tot}(t) := \sum_{k=1}^{n} Q^o_k(t), \quad o \in \{U, M\}$.

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 $\overline{Q}^U_{tot}$, $\overline{Q}^M_{tot}$, 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 [32], we have

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

which clearly shows that the system can be stabilized by unicasting only for $\rho < 1$. We can also observe that $\overline{Q}^U_{tot}$ 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 $\overline{Q}^M_{tot}$ and compare it to that of its unicast counterpart.

IV. 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 V. 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 when it is not feasible to receive feedback from the individual users.

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

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

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

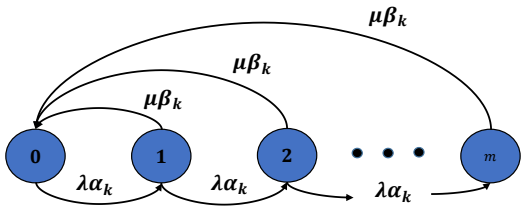


Fig. 2: Markov chain diagram of queue k under blind multicast operation.

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 $\overline{Q}^{M,B}_{tot}$ 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

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

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

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

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

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

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

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.

$$\mathbb{E}[D_{tot}^{M,F}] = \frac{\overline{Q}_{tot}^{M,F}}{\lambda_{tot}}.$$

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.

A Zipf distribution $\boldsymbol{\alpha}^z$ with parameter γ is written as

$$\alpha_k^z := \frac{\frac{1}{k^\gamma}}{\sum_{m=1}^n \frac{1}{m^\gamma}}, \quad k = 1, ..., n.$$

Remark 3: For large number of content items n and Zipf popularity distribution with parameter $\gamma = 2$,

$$\overline{Q}_{tot}^{M,B} \sim \rho(\log n)^2.$$

Proof. For $\gamma = 2$, the direct substitution of $\boldsymbol{\alpha}^z$ in (3) yields:

$$\begin{aligned}\overline{Q}_{tot}^{M,B} =& \rho||\boldsymbol{\alpha}^{\boldsymbol{z}}||_{\frac{1}{2}} = \rho\left(\sum_{k=1}^n \sqrt{\alpha_k^z}\right)^2 = \rho\left(\sum_{k=1}^n \sqrt{\frac{\frac{1}{k^2}}{\sum_{m=1}^n \frac{1}{m^2}}}\right)^2 \\ &= \rho\frac{(\sum_{k=1}^n \frac{1}{k})^2}{\sum_{m=1}^n \frac{1}{m^2}} \doteq \frac{6\rho}{\pi^2}(\log n)^2.\end{aligned}$$

Which is asymptotically true as $n \to \infty$. ■

Remark 4: For large number of content items n and Zipf popularity distribution with parameter $0 \leq \gamma < 2$,

$$\overline{Q}_{tot}^{M,B} \sim \rho n^{\min\{2-\gamma,1\}}.$$

Proof. By direct substitution of Zipf distribution with parameter γ in Equation (3), we have:

$$\begin{aligned}\overline{Q}_{tot}^{M,B} &= \rho\left(\sum_{k=1}^n \sqrt{\frac{\frac{1}{k^\gamma}}{\sum_{m=1}^n \frac{1}{m^\gamma}}}\right)^2 = \frac{\rho}{\sum_{m=1}^n \frac{1}{m^\gamma}}\left(\sum_{k=1}^n \frac{1}{k^{\frac{\gamma}{2}}}\right)^2 \\ &\leq \frac{\rho}{\sum_{m=1}^n \frac{1}{m^\gamma}}\left(\int_0^n \frac{1}{k^{\frac{\gamma}{2}}} dk\right)^2 = \frac{\rho}{\sum_{m=1}^n \frac{1}{m^\gamma}}\frac{4n^{2-\gamma}}{(2-\gamma)^2} \\ &\leq \frac{\rho}{\int_1^{n+1} \frac{1}{m^\gamma} dm}\frac{4n^{2-\gamma}}{(2-\gamma)^2} = \frac{4(\gamma-1)n^{\gamma-2}(n+1)^\gamma\rho}{(\gamma-2)^2((n+1)^\gamma-(n+1))}\end{aligned} \tag{5}$$

On the other hand we have:

$$\overline{Q}_{tot}^{M,B} \geq \frac{\rho}{1+\int_1^n \frac{1}{m^\gamma}dm}\left(\int_1^{n+1}\frac{1}{k^{\frac{\gamma}{2}}}dk\right)^2$$
$$= \frac{4\rho(\gamma-1)n^\gamma((n+1)^\gamma - 2(n+1)^{1+\frac{\gamma}{2}} + (n+1)^2)}{(\gamma-2)^2(n+1)^\gamma(\gamma n^\gamma - n)} \quad (6)$$

From equations (5) and (6) as n grows, we have the result. ■

Remark 5: 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 $\rho < 1$.

Remark 6: Assume uniform distribution of popularities and $\rho > 1$. As $n \to \infty$, the average expected number of requests per queue under multicast operation satisfies

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

That is, FCFS work-conserving multicasting attains 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 7: Assume uniform distribution of popularities and $\rho = 1$. As $n \to \infty$, the average expected number of requests per queue under multicast operation satisfies

$$\lim_{n\to\infty}\frac{\overline{Q}_{tot}^{M,o}}{n} = \begin{cases} \sqrt{\frac{2}{\pi n}}, & o = F, \\ 1, & o = B. \end{cases}$$

That is, FCFS work-conserving multicasting attains a delay that grows with $\sqrt{n}$ while blind multicasting delay grows with n. We emphasize that in the case of $\rho = 1$, FCFS multicasting has its most advantage compared to blind multicasting.

Remark 8: 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 $\rho \ll 1$; and (ii) blind-multicast strategy tends to suffer a delay performance loss within a factor of 2 under over-loaded conditions, i.e., when $\rho \gg 1$. 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.

Table I shows a summary of the main results. For the case of loading factor $\rho < 1$, our simaulatins shows that unicast performs as good as multicast. In other words for the case $\rho < 1$ there is no need to multicast and we will just drive an upper bound on the delay of FCFS multicasting for the case $\rho < 1$.

TABLE I: $\lim_{n\to\infty}\frac{\overline{Q}_{tot}}{n}$ for different strategies

	FCFS Multicast	Blind Multicast	Unicast
$\rho > 1$	$= \frac{1}{2}(\frac{\rho^2-1}{\rho})$	$= \rho$	$= \infty$
$\rho = 1$	$= \sqrt{\frac{2}{\pi n}}$	$= 1$	$= \infty$
$\rho < 1$	$\leq \frac{1}{2}\rho^3(1-\rho)^2$	$= \rho$	$= 0$

V. 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 V-A) is based on decomposing the system into parallel queues to optimize the delay. However, the proof of Theorem 2 (in Section V-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

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbb{E}[Q_k^{M,B}(t)]dt = \rho\frac{\alpha_k}{\beta_k}. \quad (7)$$

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:

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbb{E}[Q_k^{M,B}(t)]dt = \sum_{m=0}^{\infty} m p_m. \quad (8)$$

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 (8) 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 (7) 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 [28] 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 [29] or fluid-limit [30] 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 we characterize the behavior of the average number of requests in the system. Each proof is broken down into segments 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 describing 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 the limit of $N(t) \xrightarrow[t\to\infty]{d} \bar{N}(\rho, n)$, i.e., the steady state distribution of N(t) which is characterized next and studied subsequently[3].

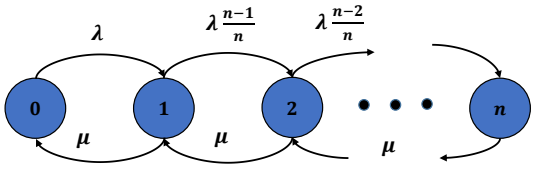


Fig. 3: Markov chain for active queues $N(t)$ under any work-conserving multicast

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

$$\pi_k = \pi_0 \prod_{m=0}^{k-1} (1 - \frac{m}{n})\rho, \forall k \geq 1, \tag{9}$$

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 (9). Then, setting the sum of probabilities to 1 gives the result for π_0. ■

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

$$s_i(\rho, n) := \sum_{k=1}^{n} k^i \prod_{m=0}^{k-1}(1 - \frac{m}{n})\rho, \tag{10}$$

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

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 (10),

$$s_i(\rho, n) = n(1 - 1/\rho)s_{i-1}(\rho, n) + \frac{n}{\rho}\sum_{m=2}^{i}(-1)^m \binom{i-1}{m-1} s_{i-m}(\rho, n) + n\delta(i-1) \tag{11}$$

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

Proof. We prove this by induction.

$$\begin{aligned} s_i(\rho, n) - n(1 - 1/\rho)s_{i-1}(\rho, n) &+ \frac{n}{\rho}\sum_{q=1}^{i-1}(-1)^q \binom{i-1}{q} s_{i-1-q}(\rho, n) \\ &= \sum_{k=1}^{n}\Bigg[k^i - n(1-1/\rho)k^{i-1} \\ &+ \frac{n}{\rho}\sum_{q=1}^{i-1}\binom{i-1}{q}(-1)^q k^{i-1-q}\Bigg]\prod_{m=0}^{k-1}(1 - \frac{m}{n})\rho. \end{aligned} \tag{12}$$

We have n terms on the right hand side of (12), 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:

$$\mathbb{E}[\bar{N}(\rho, n)] = \frac{n(1 - \frac{1}{\rho})s_0(\rho, n) + n}{1 + s_0(\rho, n)}, \tag{13}$$

$$\mathbb{E}[\bar{N}(\rho, n)^2] = \frac{(n^2(1 - \frac{1}{\rho})^2 + \frac{n}{\rho})s_0(\rho, n) + n^2(1 - \frac{1}{\rho})}{1 + s_0(\rho, n)}. \tag{14}$$

Proof.

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

From Lemma 3, writing $s_1(\rho, n)$ as a function of $s_0(\rho, n)$ gives Equation (13). Similarly,

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

Writing $s_2(\rho, n)$ in terms of $s_0(\rho, n)$ gives the result in Equation (14). ■

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

$$s_0(1, n) \doteq \sqrt{\frac{\pi}{2}n}, \tag{15}$$

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 $\Gamma(n) = \Gamma(n,0)$ [33], we can rewrite $s_0(1,n)$ as:

$$s_0(1,n) = \frac{n!}{n^n} e^n \frac{\Gamma(n,n)}{\Gamma(n)}. \tag{16}$$

Since from [34], $\lim_{n\to\infty} \frac{\Gamma(n,n)}{\Gamma(n)} = \frac{1}{2}$, 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:

$$\lim_{n\to\infty} f(\rho,n) = \begin{cases} \rho, & \rho < 1, \\ 1, & \rho > 1, \end{cases} \tag{17}$$

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

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

On the other hand we have:

$$\begin{aligned} s_0(\rho,n) &\geqslant \sum_{k=1}^{\sqrt{n}} \rho^k \prod_{m=0}^{k-1} (1 - \frac{m}{n}) \\ &\geqslant \sum_{k=1}^{\sqrt{n}} \rho^k (1 - \frac{k-1}{n})^k \geqslant \sum_{k=1}^{\sqrt{n}} \rho^k (1 - \frac{\sqrt{n}-1}{n})^k \\ &\geqslant \frac{\rho(1 - \frac{\sqrt{n}-1}{n})}{1 - \rho(1 - \frac{\sqrt{n}-1}{n})} \times (1 - \rho^{\sqrt{n}}(1 - \frac{\sqrt{n}-1}{n})^{\sqrt{n}}). \end{aligned} \tag{19}$$

From (18) and (19), 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.

$$\begin{aligned} s_0(\rho,n) &= \sum_{k=1}^{n} \prod_{m=0}^{k-1} (1 - \frac{m}{n})\rho = \sum_{k=1}^{n} \rho^k \prod_{m=0}^{k-1} (1 - \frac{m}{n}) \\ &= \frac{n!}{(\frac{n}{\rho})^n} \sum_{k=1}^{n} \frac{(\frac{n}{\rho})^{n-k}}{(n-k)!} = \frac{n!}{(\frac{n}{\rho})^n} \sum_{j=0}^{n-1} \frac{(\frac{n}{\rho})^j}{j!} \end{aligned} \tag{20}$$

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 specifies an expression 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 and $N_k(t)$ be the number of active queues in the system when queue k just becomes active at time t, then

$$\overline{Q}_{tot}^{M,F} = n \frac{\frac{\rho}{2n}\mathbb{E}[N_k(t)^2] + (1 + \frac{\rho}{2n})\mathbb{E}[N_k(t)]}{\mathbb{E}[N_k(t)] + \frac{n}{\rho}}. \tag{21}$$

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^{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 $N_k(t_i')$ be the number of active queues

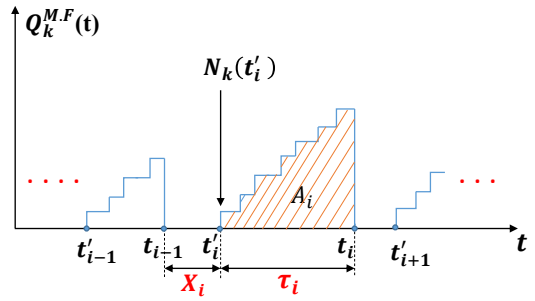


Fig. 4: Evolution of queue k under FCFS multicasting.

in the system, given that queue k has just become active at time t_i'. We should note that $N_k(t_i')$ includes queue k itself. That is, $N_k(t_i') \geq 1$. Let τ_i denote the duration queue k must wait while being active in order to be fully serviced for the i^{th} time. So $t_i = t_{i-1} + X_i + \tau_i$. Fig. 4 demonstrates the evolution of queue k under FCFS multicasting. Since we are interested in the average number of requests in queue k, based on Fig. 4 we claim that:

$$\mathbb{E}\left[Q_k^{FCFS}\right] = \lim_{i\to\infty} \frac{1}{t_i} \sum_{j=0}^{i-1} A_j = \frac{\mathbb{E}\left[A_i\right]}{\mathbb{E}\left[X_i\right] + \mathbb{E}\left[\tau_i\right]},$$

where A_i is the area shown in the figure and based on the figure is equal to $E\left[A_i\right] = E\left[\tau_i\right] + \frac{\lambda}{n}\frac{E[\tau_i^2]}{2}$. The first term comes from the fact that at time t_i' there is one arrival to queue k that makes it active and the second term is the area of the triangle knowing that the rate at which queue k receives arrival is $\frac{\lambda}{n}$. Now we will provide a rigorous proof for the claim we just presented.

We define $T_i := X_i + \tau_i$. Given $N_k(t_i') = v, \tau_i = \sum_{j=1}^{v} Y_j$, 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 $K_T = \max\{i \geq 0 | t_{i-1} \leq T\}$, which is the number of times that queue k has received service by time T. We have:

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

where $M_k[i] = \int_{t_i}^{t_{i+1}} \mathbb{E}[Q_k^F(t)]dt$ 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 $\frac{K_T}{T} \to \frac{1}{\mathbb{E}[T_i]}$ as $T \to \infty$, both with probability 1. We thus have

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

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

$$\mathbb{E}[M_k[i]] = \mathbb{E}[\int_{t_i}^{t_i+1} \mathbb{E}[Q_k^F(t)]dt] =$$

$$\mathbb{E}_{X_i,\tau_i}\left[\tau + \int_{t_i+x}^{t_i+x+\tau}\int_{t_i+x}^{t_i+x+\tau} \mathbb{E}[A_k(t)]dldt | x_i = x, \tau_i = \tau\right]$$

$$= \mathbb{E}_{X_i,\tau_i}\left[\tau + \frac{\lambda}{2n}\tau^2\right] = \frac{\mathbb{E}[N_k(t)]}{\mu} + \frac{\lambda}{2n\mu^2}\mathbb{E}[N_k(t)^2 + N_k(t)],$$

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

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

Which $N_k(t)$ is the number of active queue in the system when queue k turns active at time t. Substituting in the previous inequality, we get

$$\overline{Q}_k^{M,F} = \frac{\frac{\mathbb{E}[N_k(t)]}{\mu} + \frac{\lambda(\mathbb{E}[N_k(t)^2] + 2\mathbb{E}[N_k(t)])}{2n\mu^2}}{\frac{n}{\lambda} + \frac{\mathbb{E}[N_k(t)]}{\mu}}.$$

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

Lemma 8: Let $N_k(t)$ be the number of active queues in the system, given that queue k has just become active at time t, then:

$$\mathbb{E}[N_k(t)] \doteq \begin{cases} \mathbb{E}[\bar{N}(\rho, n)], & \rho > 1, \\ \mathbb{E}[\bar{N}(1, n)] + \frac{2}{\pi}, & \rho = 1. \end{cases} \quad (22)$$

$$\mathbb{E}[N_k(t)^2] \doteq \begin{cases} \mathbb{E}[\bar{N}(\rho, n)^2], & \rho > 1, \\ \mathbb{E}[\bar{N}(1, n)^2] + E[\bar{N}(1, n)], & \rho = 1. \end{cases} \quad (23)$$

Proof. Define $N_k^c(t)$ which is the number of active queues at time t given that queue k is not active, as:

$$N_k^c(t) = \mathbb{1}(Q_k^{M,F}(t) = 0)\sum_{i=1}^{n} \mathbb{1}(Q_i^{M,F}(t) > 0), \quad (24)$$

where $\mathbb{1}(x) = \begin{cases} 1, & x > 0, \\ 0, & x = 0. \end{cases}$

Let t_i' be the time when queue k just became active for the i^{th} time. Therefore $N_k(t_i')$ is the number of active queues when queue k became active for the ith time. Then we can write:

$$N_k(t_i') = N_k^c(t_i' - \delta) + 1, \quad (25)$$

For small enough δ. In other words, at time $t_i' - \delta$ that queue k is not still active, number of active queues is given by $N_k^c(t_i' - \delta)$. After small enough time δ when queue k just becomes active, the number of active queues is given by $N_k(t_i')$. If δ is small enough, queue k turning active is the only event ocurring in the tiny interval $(t_i' - \delta, t_i')$, so the number of active queues will increase exactly by one at the moment when queue k just becomes active. Since $A_k(t)$ is the request arrival process for item k which under uniform popularity distribution is a Poisson with rate $\frac{\lambda}{n}$ independent of all the arrival processes to other queues. Let δ be infinitesimally small and define the event

$\mathbb{A}_k(t) := \{A_k(t) - A_k(t-\delta) = 1\}$ to show the time t that queue k just receives an arrival. Since the arrivals to queues are independently distributed, $\mathbb{A}_k(t)$ is independent of $N_k^c(t)$. In other words at any time t, whether there is a arrival to queue k or not is independent of how many active queues are in system given that queue k is not active. We thus have:

$$P(N_k^c(t) = k) = P(N_k^c(t) = k|\mathbb{A}_k(t)) = P(N_k^c(t_i{}') = k)\forall i,$$

since the time that queue k have an arrival given that it was not already active is shown with $t_i{}'$. Under any work-conserving multicast strategy, $N_k^c(t)$ is a Markov process and we are interested in the steady state distribution of $N_k^c(t) \xrightarrow[t\to\infty]{d} \bar{N}_k^c$. As we started with general t, $P(N_k^c(t) = k) = P(\bar{N}_k^c = k)$, which gives the distribution of $N_k(t_i{}')$ as:

$$P(N_k(t_i{}') = k) = P(\bar{N}_k^c = k-1), \tag{26}$$

Using the Markov chain for $N_k^c(t)$ under any work-conserving multicast, let $\pi_k{}' = P(\bar{N}_k^c = k)$, then:

$$\pi_k{}' = \pi_0{}' \prod_{m=1}^{k} (1 - \frac{m}{n})\rho, \forall k \geq 1, \tag{27}$$

setting the sum of probabilities to 1 gives $\pi_0{}' = \frac{\rho}{s_0(\rho,n)}$. Then:

$$\mathbb{E}[\bar{N}_k^c] = \pi_0{}' \sum_{k=0}^{n-1} k \prod_{m=1}^{k} \left(1 - \frac{m}{n}\right)\rho = \frac{s_1(\rho,n) - s_0(\rho,n)}{s_0(\rho,n)},$$

$$\mathbb{E}\left[(\bar{N}_k^c)^2\right] = \pi_0{}' \sum_{k=0}^{n-1} k^2 \prod_{m=1}^{k} \left(1 - \frac{m}{n}\right)\rho = \frac{s_2(\rho,n) - 2s_1(\rho,n) + s_0(\rho,n)}{s_0(\rho,n)}$$

According to Equation (26) and letting t be a general time that queue k just became active, we have $\mathbb{E}[N_k(t)] = 1 + \mathbb{E}[\bar{N}_k^c]$ and $\mathbb{E}[N_k(t)^2] = \mathbb{E}[(\bar{N}_k^c)^2] + 2\mathbb{E}[\bar{N}_k^c] + 1$. Using the Equations (11), we expand $s_2(\rho,n)$ and $s_1(\rho,n)$ as a function of $s_0(\rho,n)$. Then comparing the results with Equations (13) and (14) and using the behaviour of $s_0(\rho,n)$ which is given at Equation (15) for $\rho = 1$ and in (17) for $\rho > 1$, we have the results. ■

Lemma 9: Let $N_k(t)$ be the number of active queues in the system, given that queue k has just become active at time t, then for $\rho < 1$:

$$\mathbb{E}[N_k(t)] \leq \mathbb{E}[\bar{N}(\rho,n)] + 1,$$

$$\mathbb{E}[N_k(t)^2] \leq \mathbb{E}[\bar{N}(\rho,n)^2] + 2\mathbb{E}[\bar{N}(\rho,n)] + 1.$$

Proof. First we show that $\pi_k{}' \leq \pi_k \ \forall k \geq 1$ which π_k and $\pi_k{}'$ are given in Equations (9) and (27) respectively. Assuming that the inequality $\pi_k{}' \leq \pi_k \ \forall k \geq 1$ holds, gives:

$$\frac{\rho}{s_0(\rho,n)} \prod_{m=1}^{k} (1 - \frac{m}{n})\rho \leq \frac{1}{1 + s_0(\rho,n)} \prod_{m=0}^{k-1} (1 - \frac{m}{n})\rho,$$

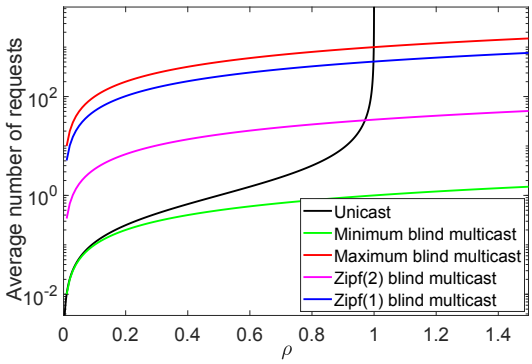


Fig. 5: Comparison between unicast and blind multicast for $n = 1000$ items.

removing the terms from both sides and by replacement we have:

$$(n-k)\rho \leq \frac{s_0(\rho,n)}{1 + s_0(\rho,n)} n.$$

Replacement of $\frac{s_0(\rho,n)}{1+s_0(\rho,n)} \doteq \rho$ which is given in Lemma 6 for $\rho < 1$ gives $(n-k) \leq n$ that is always true for all values of $k \leq n$, so the assumption holds.

Then using (26) for arbitrary time t that queue k just becomes active, gives:

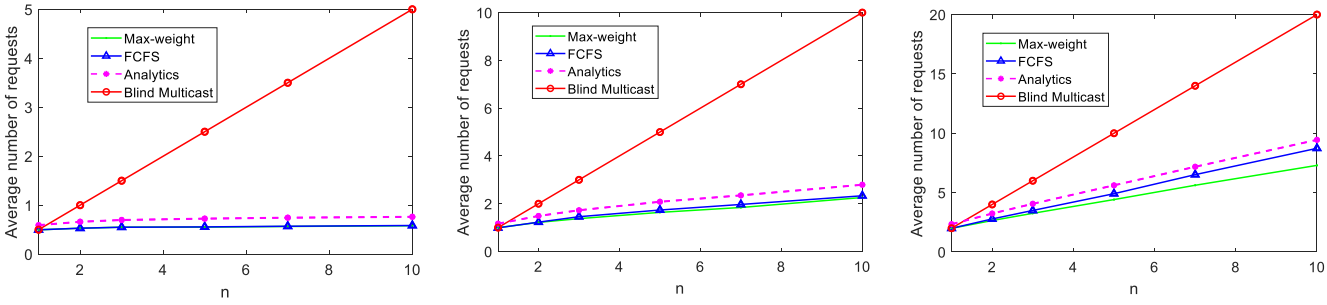
$$P(N_k(t) = k) = \pi_{k-1}{}' \leq \pi_{k-1} = P(\bar{N}(\rho,n) = k-1),$$

which hold for $\forall k \geq 1$. Recall that $N_k(t) \geq 1$, since it at least includes queue k which just became active at time t. Taking expectation of both sides gives, $\mathbb{E}[N_k(t)] \leq \mathbb{E}[\bar{N}(\rho,n)] + 1$ and similarly $\mathbb{E}[N_k(t)^2] \leq \mathbb{E}[\bar{N}(\rho,n)^2] + 2\mathbb{E}[\bar{N}(\rho,n)] + 1$. ■

Now for $\rho = 1$ and $\rho > 1$, using the results for $\mathbb{E}[N_k(t)]$ and $\mathbb{E}[N_k(t)^2]$ given in Equations (22) and (23) which is a function of ρ and n from the analysis of active queues and substituting the results in Equation (21), we have the exact expression for the total average number of requests in the system operating under FCFS multicasting as a function of ρ and n. For $\rho < 1$, using the results of Lemma 9 and analysis for statistics of active queues and using the behaviour of $s_0(\rho,n)$ which is given in Equation (17) for $\rho < 1$, we derive the upper bound for the average number of requests in the system working under FCFS multicasting. Letting $n \to \infty$ gives the result of Theorem 2.

VI. 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 distribution, and then provide more numerical results for non-uniform popularity distributions (such as the 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. We should note that our analysis is conducted under the



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

Fig. 6: Average number of requests in the system over n for different loading factors ρ.

assumption of large database size and uniform popularity distribution. We have investigated the average number of requests in the system, proposed an upper bound for the case of $\rho < 1$ and derived the exact asymptotic expression for the two cases of $\rho = 1$ and $\rho > 1$. Also, since the analysis has been provided for asymptotic as n grows, we simulate the system's behavior when the number of data contents, n, is large, we set it to $n = 1000$, unless stated otherwise.

A. *Validation of Main Results under Uniform Popularities*

In Fig. 5, 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 $\rho \geq 0$ as can be seen in Fig. 5. We can also observe from Fig. 5 that unicast outperforms blind multicast for $\rho < 1$ under the considered instances of Zipf distributed popularity with parameter $\gamma \leq 2$ which is consistent with the insights of Remark 5. Notice the degenerate distribution is a Zipf distribution with parameter $\gamma = \infty$ which results in the minimum average delay, shown in Fig. 5 as the minimum blind multicast.

Fig. 6 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 from Fig. 6(c) that our upper-bound for the case of $\rho < 1$ is very accurate even for small number of queues and we expect that as n increases, our upper bound becomes tighter. For the case of $\rho = 1$ and $\rho > 1$ our analysis for the average number of active queues

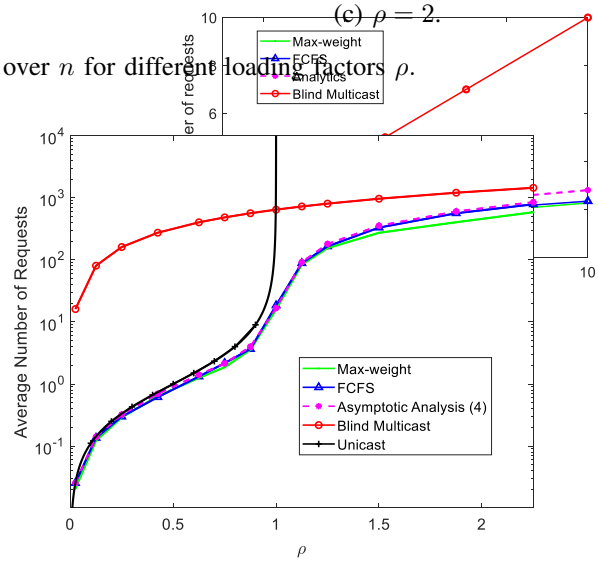


Fig. 7: Average number of requests in the system for different policies under uniform popularity distribution and $n = 1000$.

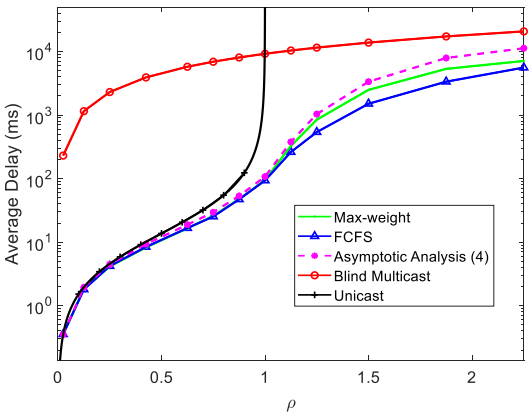


Fig. 8: Average delay of month june under the assumption of Zipf(0.86) popularity distribution and $n = 10000$.

that we did under large database size n, is close to simulation values even though that $n = 1000$ is not large here.

Fig. 7 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, analytical results that we derived for FCFS in Equation (4) is very exact. Moreover, performance of FCFS is very close to that of Max-Weight. According to Fig. 7, for small loading factor ρ, uni-

cast 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 $\rho \gg 1$, work-conserving multicast still outperforms blind multicast by a factor of 2 as it has been noted in Remark 6.

B. Performance Comparison for Non-uniform Popularities

In this section, to illustrate the possibility of practical application of the proposed content multicasting schemes and to verify the validity of research conclusions, we aim to investigate the performance of our blind and work-conserving multicast policies under non-uniform popularity distributions and compare their performance to a heuristic Max-Weight multicast policy.

To show the practicality of our analysis, we use an extensive set of real-world data, namely the data set of the BBC iPlayer [35], [36], [37], to obtain realistic video demand distributions. The BBC iPlayer is a video streaming service from BBC that provides video content for a number of BBC channels without charge. Content on the BBC iPlayer is available for up to 30 days depending on the policies. We consider the dataset covering June, 2014, which include 192,120,311 recorded access sessions, resulting in request rates $\lambda = 74.1205$ requests per second. The number of files according to the iPlayer database is larger than $n = 10000$. We consider multicast over 802.11 (Wi-Fi) wireless networks to stream video files. The 802.11 standard allows for multicast transmission as part of asynchronous services. According to [38], the popularity distribution of video files of the BBC iPlayer requested by the users in June 2014 can be approximated by the Zipf distributions with parameter $\gamma = .86$.

Fig. 8 shows the average delay of the system for different policies in the month of June under the approximation of Zipf popularity distribution with parameter $\gamma = 0.86$. As it can be seen from the figure, the performance of FCFS is very close to Max-Weight and our analysis which we derived for FCFS in Theorem 2, under uniform popularity distribution, is also reasonable upper bound for more practical systems and under non-uniform popularity distributions like the Zipf distribution.

Fig. 9 shows the delay gain of FCFS multicast compared to blind multicast as a function of parameter of zipf distribution under different loading factors ρ. Under uniform popularity distribution with $s = 0$ given that $\rho > 1$, the gain is $2(\frac{\rho^2}{\rho^2-1})$ as it has been noted in Remark 6 which is large for ρ close to 1 and the gain decreases as ρ increases. When zipf parameter s increases, content popularity distribution will become more degenerate and blind multicasting will assign most the service to queue with largest arrival rate, resulting in same performance of FCFS multicasting. So we expect that gain will pick somewhere between and as it can be seen from figure, for loading factor $\rho = 1$, gain picks under Zipf distribution with parameter 1. Also, this

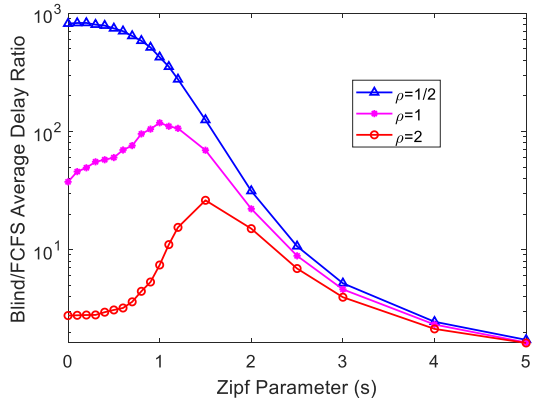


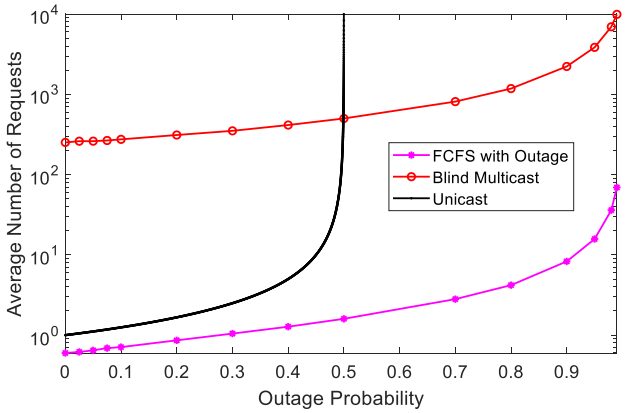
Fig. 9: Blind/FCFS average delay ratio over Zipf parameter s and $n = 1000$.

figure confirms that when ρ increases, the gain decreases which also agrees with the results of Remarks 6 and 7 that FCFS has the most advantage when ρ is close to 1 ans as ρ increases, the gain decreases. Note that, as we see in Fig. 7, unicast is a reasonable policy when $\rho \ll 1$, but it becomes unstable for $\rho \geq 1$, revealing the benefits of our proposed FCFS work-conserving policy.

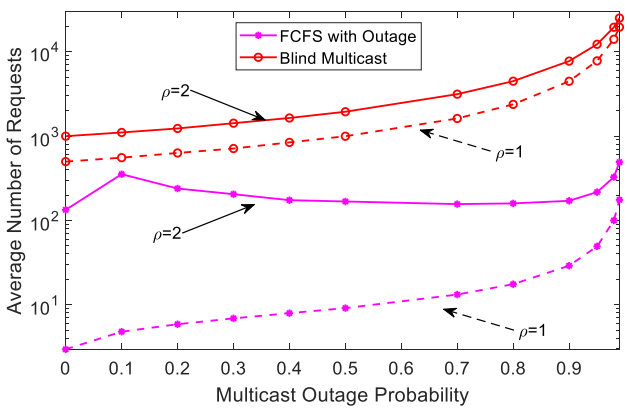
C. Effect of Error on Delay Gains

In this section, we consider the transmission failures over the wireless channel as a practical issue and address the effect of multicasting error on delay gains to demonstrate that the reaped gains are still substantial even in the systems with high multicasting outage. In practical systems, it is not possible to multicast to all users at once and due to multicast error, some users may not receive the multicasting content properly. So, we need to send that content again until it is successfully received. We should note that in the presence of outage, we adjust the FCFS multicasting policy so that, if a multicast transmission fails to be received by a user, we treat that request as a new request. We call such a policy FCFS with outage.

Fig. 10 shows the effect of channel failure on the delay performance of different policies for different loading factors ρ. For this figure, we assume $n = 1000$ and Zipf(1) popularity distribution. We can see from Fig 10(a) that, as the outage probability increases, unicast is the first policy to become unstable. Also, the proposed FCFS work-conserving multicast always outperforms both unicast and blind multicast. For $\rho \geq 1$, when unicast becomes unstable, Fig. 10(b) shows the effect of multicast outage probability on the average delays. As it can be seen from the figure, even in the presence of multicast error, both blind multicast and FCFS multicast with outage are stable in contrary to the unicast which is unstable for both $\rho = 1$ and $\rho = 2$. Also, the delay gains of FCFS multicast with outage compared to blind multicast is still substantial especially if ρ is close to 1. As the multicast outage probability approaches 1, the system becomes unstable independent of the multicast policy. We can conclude from this figure that under practical operational conditions with multicasting error and Zipf-like popularity



(a) $\rho = 1/2$.



(b) $\rho = 1, \rho = 2$.

Fig. 10: The effect of channel failure on the performance for different loading factors under Zipf(1) popularity distribution.

distributions, the gain of FCFS work-conserving multicast is significant compared to other multicast policies like blind multicast. Recall that for large loading factors $\rho \geq 1$, unicast is always unstable and is not a possible service policy.

VII. 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] K. Fall, "A delay-tolerant network architecture for challenged internets," in *Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications*, 2003, pp. 27–34.

[7] U. Varshney, "Multicast over wireless networks," *Communications of the ACM*, vol. 45, no. 12, pp. 31–37, 2002.

[8] S.-J. Lee, W. Su, and M. Gerla, "On-demand multicast routing protocol in multihop wireless mobile networks," *Mobile networks and applications*, vol. 7, no. 6, pp. 441–453, 2002.

[9] S.-J. Lee, W. Su, J. Hsu, M. Gerla, and R. Bagrodia, "A performance comparison study of ad hoc wireless multicast protocols," in *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No. 00CH37064)*, vol. 2. IEEE, 2000, pp. 565–574.

[10] W. Zhao, M. Ammar, and E. Zegura, "Multicasting in delay tolerant networks: semantic models and routing algorithms," in *Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, 2005, pp. 268–275.

[11] M. Mongiovì, A. K. Singh, X. Yan, B. Zong, and K. Psounis, "Efficient multicasting for delay tolerant networks using graph indexing," in *2012 Proceedings IEEE INFOCOM*. IEEE, 2012, pp. 1386–1394.

[12] W. Gao, Q. Li, B. Zhao, and G. Cao, "Multicasting in delay tolerant networks: a social network perspective," in *Proceedings of the tenth ACM international symposium on Mobile ad hoc networking and computing*, 2009, pp. 299–308.

[13] X. Lin and N. B. Shroff, "Joint rate control and scheduling in multihop wireless networks," in *2004 43rd IEEE Conference on Decision and Control (CDC)(IEEE Cat. No. 04CH37601)*, vol. 2. IEEE, 2004, pp. 1484–1489.

[14] M. J. Neely, E. Modiano, and C.-P. Li, "Fairness and optimal stochastic control for heterogeneous networks," *IEEE/ACM Transactions On Networking*, vol. 16, no. 2, pp. 396–409, 2008.

[15] A. Zhou, M. Liu, Z. Li, and E. Dutkiewicz, "Cross-layer design for proportional delay differentiation and network utility maximization in multi-hop wireless networks," *IEEE transactions on wireless communications*, vol. 11, no. 4, pp. 1446–1455, 2012.

[16] E. Ahmed, A. Eryilmaz, M. Medard, and A. E. Ozdaglar, "On the scaling law of network coding gains in wireless networks," in *MILCOM 2007-IEEE Military Communications Conference*. IEEE, 2007, pp. 1–7.

[17] A. Eryilmaz, A. Ozdaglar, and M. Medard, "On delay performance gains from network coding," in *2006 40th Annual Conference on Information Sciences and Systems*. IEEE, 2006, pp. 864–870.

[18] W.-L. Yeow, A. T. Hoang, and C.-K. Tham, "Minimizing delay for multicast-streaming in wireless networks with network coding," in *IEEE INFOCOM 2009*. IEEE, 2009, pp. 190–198.
[19] Q. Du and X. Zhang, "Statistical qos provisionings for wireless unicast/multicast of multi-layer video streams," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 3, pp. 420–433, 2010.
[20] D. Li, Y. Li, J. Wu, S. Su, and J. Yu, "Esm: Efficient and scalable data center multicast routing," *IEEE/ACM Transactions on Networking*, vol. 20, no. 3, pp. 944–955, 2011.
[21] X. Li and M. J. Freedman, "Scaling ip multicast on datacenter topologies," in *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies*, 2013, pp. 61–72.
[22] Y. Vigfusson, H. Abu-Libdeh, M. Balakrishnan, K. Birman, R. Burgess, G. Chockler, H. Li, and Y. Tock, "Dr. multicast: Rx for data center communication scalability," in *Proceedings of the 5th European conference on Computer systems*, 2010, pp. 349–362.
[23] M. V. Jamali, A. Mirani, A. Parsay, B. Abolhassani, P. Nabavi, A. Chizari, P. Khorramshahi, S. Abdollahramezani, and J. A. Salehi, "Statistical studies of fading in underwater wireless optical channels in the presence of air bubble, temperature, and salinity random variations," *IEEE Transactions on Communications*, vol. 66, no. 10, pp. 4706–4723, 2018.
[24] D. Wu, X. Sun, Y. Xia, X. Huang, and T. S. E. Ng, "Hyperoptics: A high throughput and low latency multicast architecture for datacenters," in *8th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 16)*. Denver, CO: USENIX Association, Jun. 2016. [Online]. Available: https://www.usenix.org/conference/hotcloud16/workshop-program/presentation/wu
[25] G. M. De Brito, P. B. Velloso, and I. M. Moraes, *Information-centric Networks: A New Paradigm for the Internet*. John Wiley & Sons, 2013.
[26] M. F. Bari, S. R. Chowdhury, R. Ahmed, R. Boutaba, and B. Mathieu, "A survey of naming and routing in information-centric networks," *IEEE Communications Magazine*, vol. 50, no. 12, pp. 44–53, 2012.
[27] 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.
[28] Y. Cui, V. K. Lau, R. Wang, H. Huang, and S. Zhang, "A survey on delay-aware resource control for wireless systems—large deviation theory, stochastic lyapunov drift, and distributed stochastic learning," *IEEE Transactions on Information Theory*, vol. 58, no. 3, pp. 1677–1701, 2012.
[29] 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.
[30] 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.
[31] B. Abolhassani, J. Tadrous, and A. Eryilmaz, "Wireless multicasting for content distribution: Stability and delay gain analysis," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 1–9.
[32] L. Kleinrock, *Queueing systems, Volume 1: Theory*. Wiley New York, 1975.
[33] G. Jameson, "The incomplete gamma functions," *The Mathematical Gazette*, vol. 100, no. 548, pp. 298–306, 2016.
[34] 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.
[35] D. Karamshuk, N. Sastry, M. Al-Bassam, A. Secker, and J. Chandaria, "Take-away tv: Recharging work commutes with predictive preloading of catch-up tv content," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 8, pp. 2091–2101, 2016.
[36] G. Nencioni, N. Sastry, G. Tyson, V. Badrinarayanan, D. Karamshuk, J. Chandaria, and J. Crowcroft, "Score: Exploiting global broadcasts to create offline personal channels for on-demand access," *IEEE/ACM Transactions on Networking*, vol. 24, no. 4, pp. 2429–2442, 2015.
[37] M.-C. Lee, A. F. Molisch, N. Sastry, and A. Raman, "Individual preference probability modeling and parameterization for video content in wireless caching networks," *IEEE/ACM Transactions on Networking*, vol. 27, no. 2, pp. 676–690, 2019.
[38] K. P. Gummadi, R. J. Dunn, S. Saroiu, S. D. Gribble, H. M. Levy, and J. Zahorjan, "Measurement, modeling, and analysis of a peer-to-peer file-sharing workload," in *Proceedings of the nineteenth ACM symposium on Operating systems principles*, 2003, pp. 314–329.

Bahman Abolhassani received the B.Sc. and M.Sc. degrees in electrical engineering from Sharif University of Technology (SUT), Tehran, Iran, in 2015 and 2017, respectively. He is currently pursuing the Ph.D. degree with the Electrical and Computer Engineering Department, The Ohio State University, Columbus, OH, USA. Between 2015 and 2017, he was a researcher at the Optical Networks Research Laboratory, SUT. His research interests include communication networks, optimization theory, caching and algorithm design.

John Tadrous (S'10–M'15–SM'20) received the B.Sc. degree from the EE Department, Cairo University, the M.Sc. degree in wireless communications from the Center of Information Technology, Nile University, in 2010, and the Ph.D. degree in electrical engineering from the ECE Department, The Ohio State University, in 2014. He was a Research Assistant at the Wireless Intelligent Networks Center, Nile University, from 2008 to 2010, where he worked on resource allocation and power control for cognitive radio networks. From 2010 to 2014, he was a Research Associate at the Information Processing Systems Laboratory, where he worked on proactive resource allocation and scheduling, smart data pricing, and information theory. From 2014 to 2016, he was with the Center for Multimedia Communication, Rice University, as a Post-Doctoral Research Associate, where he worked on modeling and analysis of interactive data traffic, full-duplex communications, and beamforming design for massive MIMO systems. Since 2016, he has been an Assistant Professor at the Department of Electrical and Computer Engineering, Gonzaga University. He received Gonzaga University's Faculty Award for Professional Contributions in May 2020.

Atilla Eryilmaz (S'00 / M'06 / SM'17) received his M.S. and Ph.D. degrees in Electrical and Computer Engineering from the University of Illinois at Urbana-Champaign in 2001 and 2005, respectively. Between 2005 and 2007, he worked as a Postdoctoral Associate at the Laboratory for Information and Decision Systems at the Massachusetts Institute of Technology. Since 2007, he has been at The Ohio State University, where he is currently a Professor and the Graduate Studies Chair of the Electrical and Computer Engineering Department. Dr. Eryilmaz's research interests span optimal control of stochastic networks, machine learning, optimization, and information theory. He received the NSF-CAREER Award in 2010 and two Lumley Research Awards for Research Excellence in 2010 and 2015. He is a co-author of the 2012 IEEE WiOpt Conference Best Student Paper, subsequently received the 2016 IEEE Infocom, 2017 IEEE WiOpt, 2018 IEEE WiOpt, and 2019 IEEE Infocom Best Paper Awards. He has served as: a TPC co-chair of IEEE WiOpt in 2014 and of ACM Mobihoc in 2017; an Associate Editor (AE) of IEEE/ACM Transactions on Networking between 2015 and 2019; and is an AE of IEEE Transactions on Network Science and Engineering since 2017.