Proactive Resource Allocation: Harnessing the Diversity and Multicast Gains

John Tadrous, Atilla Eryilmaz, and Hesham El Gamal

***Abstract*—This paper introduces the novel concept of proactive resource allocation through which the predictability of user behavior is exploited to balance the wireless traffic over time, and hence, significantly reduce the bandwidth required to achieve a given blocking/outage probability. We start with a simple model in which the smart wireless devices are assumed to predict the arrival of new requests and submit them to the network T time slots in advance. Using tools from large deviation theory, we quantify the resulting prediction diversity gain to establish that the decay rate of the outage event probabilities increases with the prediction duration T. This model is then generalized to incorporate the effect of the randomness in the prediction look-ahead time T. Remarkably, we also show that, in the cognitive networking scenario, the appropriate use of proactive resource allocation by the primary users improves the diversity gain of the secondary network at no cost in the primary network diversity. We also shed lights on multicasting with predictable demands and show that the proactive multicast networks can achieve a significantly higher diversity gain that scales super-linearly with T. Finally, we conclude by a discussion of the new research questions posed under the umbrella of the proposed proactive (non-causal) wireless networking framework.**

***Index Terms*—Scheduling, large deviations, diversity gain, multicast alignment, predictive traffic.**

I. Introduction

IDeally, wireless networks should be optimized to deliver the best Quality of Service (in terms of reliability, delay, and throughput) to the subscribers with the minimum expenditure in resources. Such resources include transmitted power, transmitter and receiver complexity, and allocated frequency spectrum. Over the last few years, we have experienced an ever increasing demand for wireless spectrum resulting from the adoption of *throughput hungry* applications in a variety of civilian, military, and scientific settings.

Since the available spectrum is non renewable and limited, this demand motivates the need for efficient wireless networks that *maximally utilize* the spectrum. In this work, we focus our attention on the resource allocation aspect of the problem and propose a new paradigm that offers remarkable spectral gains in a variety of relevant scenarios. More specifically, our proactive resource allocation framework exploits the predictability of our daily usage of wireless devices to smooth out the traffic demand in the network, and hence, reduce the required resources to achieve a certain point on the Quality of Service (QoS) curve. This new approach is motivated by the following observations.

- While there is a severe shortage in the spectrum, it is well-documented that a significant fraction of the available spectrum is under-utilized [1]. In fact, this is the main motivation for the cognitive networking where secondary users are allowed to use the spectrum at the off time of the primary so as to maximize the spectral efficiency [2]. The cognitive radio approach, however, is still facing significant technological hurdles [3], [4] and, will offer only a partial solution to the problem. This limitation is tied to the main reason behind the under-utilization of the spectrum; namely *the large disparity between the average and peak traffic demand in the network*.

Actually, one can see that the traffic demand at the peak hours is much higher than that at night. Now, the cognitive radio approach assumes that the secondary users will be able to utilize the spectrum at the off-peak times, but at those times one may expect the secondary traffic characteristics to be similar to that of the primary users (e.g., at night most of the primary and secondary users are expected to be idle). Thereby, the overarching goal of the proactive resource allocation framework is to avoid this limitation, and hence, achieve a significant reduction in the peak to average demand ratio *without relying on out of network users*.

- In the traditional approach, wireless networks are constructed assuming that the subscribers are equipped with *dumb terminals* with very limited computational power. It is obvious that the new generation of *smart devices* enjoy significantly enhanced capabilities in terms of both processing power and available memory.This observation should inspire a similar paradigm shift in the design of wireless networks whereby the capabilities of the smart wireless terminals are leveraged to maximize the utility of the frequency spectrum, *a non-renewable resource that does not scale according to Moore's law*. Our proactive resource allocation framework is a significant step in this direction.
- The introduction of smart phones has resulted in a paradigm shift in the dominant traffic in mobile cellular networks. While the primary traffic source in traditional cellular networks was real-time voice communication, one can argue that a significant fraction of the traffic generated by the smart phones results from non-real-time data requests (e.g., file downloads). As demonstrated in the following, this feature allows for more degrees of freedom in the design of the scheduling algorithm.
- The final piece of our puzzle relates to the observation that the human usage of the wireless devices is *highly predictable*. This claim is supported by a growing body of evidence that ranges from the recent launch of **Google Instant** to the

Authors are with the Department of Electrical and Computer Engineering at the Ohio State University, Columbus, USA.

E-mail: {tadrousj,eryilmaz,helgamal}@ece.osu.edu

The work of John Tadrous is supported by QNRF grant number NPRP 09-1168-2-455. The work of Atilla Eryilmaz is also supported by NSF grants: CAREER-CNS-0953515 and CCF-0916664.

interesting findings on our predictable mobility patterns [5]. An example would be the fact that our preference for a particular news outlet is not expected to change frequently. So, if the smart phone observes that the user is downloading CNN, for example, in the morning for a sequence of days in a row then it can **anticipate** that the user will be interested in the CNN again the following day. One can now see the potential for scheduling early downloads of the predictable traffic to *reduce the peak to average traffic demand* by maximally exploiting the available spectrum in the network idle time.

These observations motivate us in this work to develop and analyze proactive resource allocation strategies in the presence of user predictability under various conditions, dynamics, and operational capabilities. In particular, our contributions along with their position in the rest of the paper are:

- In Section II we state the predictive network model and introduce the outage probability and the associated diversity gain for two main scaling regimes, namely, linear and polynomial scaling.
- In Section III, we establish the diversity gain of non-predictive and predictive networks, and analyze the effect of the random look-ahead window size, T. Our analysis reveals a minimum improvement factor of (1+T) in the diversity gain for both linear and polynomial scaling regimes.
- In Section IV, we investigate proactive scheduling in a two-QoS network,typical of a cognitive radio network, where we prove the existence of a proactive scheduling policy that can maintain the diversity gain level of the primary predictive network while strictly improving it for the secondary non-predictive network.
- In Section V, we analyze the robustness of the proactive resource allocation scheme to the prediction errors and determine the optimal choice of the look-ahead window size given an imperfect prediction mechanism to maximize the diversity gain, which is shown to be always strictly greater than that of the non-predictive network.
- In Section VI, we analyze the proactive multicasting with predictable demands, and show the significant gains that can be leveraged through the alignment property offered by predictable multicast traffic. More specifically, we show that the diversity gain of a proactive multicasting network is increasing super-linearly with the window size, T, for the linear scaling regime.
- In Section VIII, we conclude the paper and highlight other important research aspects that can be leveraged through predictive wireless communications.

The proactive wireless network can be viewed as an ordinary network with delay tolerant requests, that is, when the network predicts a request a head of time, the actual arrival time of that request can be considered as a hard deadline that the scheduler should meet. In [6], scheduling with deadlines was considered for a single packet under the objective of minimizing the expected energy consumed for transmission. In [7], the asymptotic performance of the error probability with the signal-to-noise ratio was analyzed when the bits of each codeword must be delivered under hard deadline constraints. In [8] and [9], scheduling with deadlines was also addressed from queuing theory point of view under different objectives and multiple priority classes while optimal scheduling policies were investigated for different scenarios.

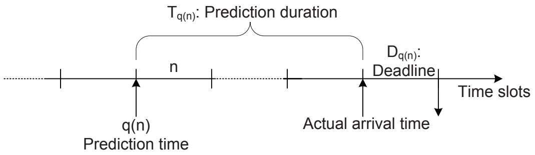


Fig. 1: Prediction Model: $q(n)$ is a request predicted at the beginning of time slot n, $T_{q(n)}$ is the prediction duration of request $q(n)$, and $D_{q(n)}$ is the actual slot of arrival for request $q(n)$ which can be considered as the deadline for such a request.

In this paper, however, we look at the scheduling problem with deadlines from a different perspective, where we define the outage probability as the probability of having a time slot suffering expiring requests, and we analyze the asymptotic decay rate of this outage probability with the system capacity, C, when the input traffic is increasing in C either linearly or polynomially, and C is approaching infinity. We call this metric the diversity gain of the network and show that its behavior can significantly be improved by exploiting the predictable behavior of the users. This metric and line of investigation are also motivated by the order-wise difference between the timescale of the prediction window lengths (typically of the order of tens of minutes, if not hours) and the timescale of application-based deadline-constraints (of the order of milliseconds) considered in other works.

II. System Model

We consider a simplified model of a single server, time-slotted wireless network where the requests arrive at the beginning of each slot. The number of arriving requests at time slot n is an integer-valued random variable denoted by $Q(n)$ that is assumed to be ergodic and Poisson distributed with mean λ. Each request is assumed to consume one unit of resource and is completely served in a single time slot. Moreover, the wireless network has a *fixed* capacity C per slot. We distinguish two types of wireless resource allocation: **reactive** and **proactive**. In reactive resource allocation, the wireless network responds to user requests right after they are initiated by the user, whereas in proactive resource allocation, the network can track, learn and then *predict* the user requests ahead of time, and hence possesses more flexibility in scheduling these requests before their actual time of arrival. We refer to the networks that perform reactive and proactive resource allocation, respectively, as **non-predictive** and **predictive** networks.

The predictive wireless network can anticipate the arrival of requests a number of time slots ahead. That is, if $q(n)$, $q \in \{1, \cdots, Q(n)\}$, is the identifier of a request predicted at the beginning of time slot n, the predictive network has the advantage of serving this request within the next $T_{q(n)}$ slots. Hence, when request $q(n)$ arrives at a predictive network, it has a deadline at time slot $D_{q(n)} = n + T_{q(n)}$ as shown in Fig. 1.

Conversely, in a non-predictive network, all arriving requests at the beginning of time slot n must be served in the same time slot n, i.e., if $q(n)$ is an unpredicted request, then $T_{q(n)} = 0$ and $D_{q(n)} = n$. At this point, we wish to stress the fact that the model operates as the time scale of the application layer at which 1) the current paradigm, i.e., non-predictive networking, treats all the requests as urgent, 2) each slot duration may be in the order of minutes and possibly hours, and 3) the system capacity is fixed since the channel fluctuation dynamics are averaged out at this time scale.

Definition 1: Let $N_0(n)$ be the number of requests in the system at the beginning of time slot n having a deadline of n. The outage event $\mathcal{O}$ is then defined as

$$\mathcal{O} \triangleq \{N_0(n) > C, n \gg 1\}, \tag{1}$$

The above definition states that an outage occurs at slot n if and only if at least one of the requests in the system expires in this slot. The term $N_0(n)$ coincides on $Q(n)$ when the network is non-predictive, and is different when the network is predictive.

Following the definition of the outage event, we denote the probability that the wireless network runs into an outage at slot $n > 0$ by $P(\mathcal{O})$. Throughout this paper, we will focus on analyzing the asymptotic decay rate of the outage probability with the system capacity C when it approaches infinity. We call this decay rate the **diversity gain** of the network. Moreover, in our analysis we assume that the mean input traffic λ scales with the system capacity in two different regimes as follows.

1) *Linear Scaling:* In this regime, the arrival process $\overline{Q}(n), n > 0$ is Poisson with rate $\bar{\lambda}$ that scales with C as

$$\bar{\lambda} = \overline{\gamma}C, \quad 0 < \overline{\gamma} < 1.$$

And with outage probability denoted by $P(\overline{\mathcal{O}})$, the associated diversity gain is defined as

$$\overline{d}(\overline{\gamma}) \triangleq \lim_{C\to\infty} -\frac{\log P(\overline{\mathcal{O}})}{C}.$$

2) *Polynomial Scaling:* In this regime, the arrival process $\widetilde{Q}(n), n > 0$, is also Poisson with rate $\tilde{\lambda}$, but the rate scales with the system capacity C polynomially as

$$\tilde{\lambda} = C^{\widetilde{\gamma}}, \quad 0 < \widetilde{\gamma} < 1.$$

And with outage probability $P(\widetilde{\mathcal{O}})$, the associated diversity gain is defined as

$$\widetilde{d}(\widetilde{\gamma}) \triangleq \lim_{C\to\infty} -\frac{\log P(\widetilde{\mathcal{O}})}{C\log C}.$$

We consider the linear scaling of the input traffic with the system resources because it is commonly used in networking literature where the parameter $\overline{\gamma}$ serves as bandwidth utilization factor. As $\overline{\gamma}$ approaches 1 the average input traffic approaches the capacity and the system becomes critically stable and more subject to outage events, whereas small values of $\overline{\gamma}$ imply underutilized resources but small probability of outage. The polynomial scaling regime is also introduced because under this type of scaling, the optimal prediction diversity gain can be fully determined through the asymptotic analysis of simple scheduling policies like earliest deadline first. Except for Section VI and its associated appendices, we consistently use the accents $\overline{\cdot}$ and $\widetilde{\cdot}$ to denote linear and polynomial scaling regimes respectively, while symbols without accents are used to denote a general case.

III. Diversity Gain Analysis

A. Diversity Gain of Reactive Networks

The reactive networks are supposed to have no prediction capabilities so they cannot serve any request prior to its time of actual arrival. Hence, the reactive network encounters an outage at time slot n if and only if $Q(n) > C$ as $N_0(n) = Q(n)$.

Theorem 1: Denote the outage probability and the diversity gain of the non-predictive network, respectively, by $P_N(\mathcal{O})$ and $d_N(\gamma)$, then

$$\overline{d}_N(\overline{\gamma}) = \overline{\gamma} - 1 - \log\overline{\gamma}, \quad 0 < \overline{\gamma} < 1, \tag{2}$$

and

$$\widetilde{d}_N(\widetilde{\gamma}) = 1 - \widetilde{\gamma}, \quad 0 < \widetilde{\gamma} < 1. \tag{3}$$

Proof: Please refer to Appendix A. ■

It can be noted that as $\overline{\gamma}$ and $\widetilde{\gamma}$ approach 1, the corresponding diversity gains $\overline{d}_N(\overline{\gamma})$ and $\widetilde{d}_N(\widetilde{\gamma})$ approach 0, as in this case the arrival rate in both regimes matches the system capacity, and hence the system becomes critically stable and the logarithm of the outage probability does not decay with C. However, the behavior of the the diversity gain is not the same when both $\overline{\gamma}$ and $\widetilde{\gamma}$ approach 0. As $\overline{\gamma} \to 0$, $\overline{d}_N(\overline{\gamma}) \to \infty$ because the arrival rate $\bar{\lambda} \to 0$, thus the resulting outage probability approaches 0 and the diversity gain approaches ∞. Whereas $\widetilde{\gamma} \to 0$ implies that $\widetilde{d}_N(\widetilde{\gamma}) \to 1$ which is the case when the input traffic is still positive but does not scale with the system capacity.

B. Diversity Gain of Proactive Networks

Unlike reactive networks, the proactive network has the flexibility to schedule the predicted requests in a window of time slots through some scheduling policy. Depending on the scheduling policy employed, the resulting outage probability (and of course the associated diversity gain) varies. By the term **optimal** prediction diversity gain, we mean the maximum diversity gain that can be achieved by the predictive network, which corresponds to the minimum outage probability denoted $P_P^*(\mathcal{O})$.

In our analysis, we consider, for simplicity, the earliest deadline first (EDF) scheduling policy, which has also been called in [13] shortest time to extinction (STE). This policy, as proved in [13], maximizes the number of served requests under a per-request deadline constraint. Further studies on this policies can be found in [8] and [14]. In the proposed predictive network, the EDF scheduling policy is defined as follows.

Definition 2 (Earlies Deadline First (EDF)): Let the maximum prediction interval for a request be denoted by T^*, i.e., $T^* = \sup_{q,n}\{T_{q(n)}\}$, and let $N_i(n), i = 0, 1, \cdots, T^*$ be the number of requests in the system at the beginning of time slot n and having a deadline of $n+i$. Then, at the beginning of slot n, the EDF policy sorts $\{N_i(n)\}_{i=0}^{T^*}$ in an ascending order with respect to i, and serves them in that order until either a total of C requests get served or the network completes the service of all existing requests in this slot.

It can be noted that EDF does not necessarily minimize the outage probability as it is only concerned with maximizing the number of served requests while the outage event does not take into account the number of dropped requests. However, EDF has two main characteristics that help in analysis. Namely, it always serves requests as long as there are any, i.e., it is a work conserving policy, and it serves requests in the order of their remaining time to deadline.

1) Deterministic Look-ahead Time: In this scenario, $T_{q(n)} = T$ for all $q(n), n > 0$ for some constant $T \geq 0$. Hence, assuming that the system employs EDF scheduling policy, we have $T^* = T$ and $N_T(n) = Q(n), n > 0$. Thus, the EDF policy will reduce to first-come-first-serve (FCFS). The outage probability in this case is denoted by $P_{PD}(\mathcal{O})$.

Lemma 1: Under EDF, let

$$\mathcal{U}_D \triangleq \left\{ \sum_{i=0}^{T} Q(n-T-i) > C(T+1), n \gg 1 \right\}$$

and

$$\mathcal{L}_D \triangleq \left\{ Q(n-T) > C(T+1), n \gg 1 \right\}.$$

Then, the events $\mathcal{U}_D$ and $\mathcal{L}_D$ constitute a necessary condition and a sufficient condition on the outage event, respectively. Hence, $P(\mathcal{L}_D) \leq P_{PD}(\mathcal{O}) \leq P(\mathcal{U}_D)$.
In the above lemma, we assume that $n \gg 1$ as we are interested in the steady state performance. The event $\mathcal{U}_D$ occurs when the number of arriving requests in consecutive $T+1$ slots is larger than the total capacity of $T+1$ slots, whereas the event $\mathcal{L}_D$ occurs when the number of arriving requests at any slot is larger than the total capacity of $T+1$ slots.

Proof: Please refer to Appendix B. ■

It is obvious from the proof that the event $\mathcal{U}_D$ is related to the outage event $\mathcal{O}$ through the EDF scheduling policy, whereas the event $\mathcal{L}_D$ is independent of the scheduling policy employed.

Theorem 2: The optimal prediction diversity gain of a proactive network with deterministic prediction interval T, denoted $d_{PD}(\gamma)$, satisfies

$$\overline{d}_{PD}(\overline{\gamma}) \geq (1+T)(\overline{\gamma} - 1 - \log \overline{\gamma}), \quad 0 < \overline{\gamma} < 1, \tag{4}$$
$$\widetilde{d}_{PD}(\widetilde{\gamma}) = (1+T)(1-\widetilde{\gamma}), \quad 0 < \widetilde{\gamma} < 1. \tag{5}$$

The above result shows that proactive resource allocation offers a multiplicative diversity gain of at least $T+1$ for the linear scaling regime and exactly $T+1$ for the polynomial scaling regime.

Proof: Please refer to Appendix C. ■

Note that, an upper bound on $\overline{d}_{PD}(\overline{\gamma})$ can be established using $P(\overline{\mathcal{L}}_D) \leq P_{PD}(\overline{\mathcal{O}})$ and following the same approach of deriving the lower bound in the theorem. This upper bound will be given by

$$\overline{d}_{PD}(\overline{\gamma}) \leq (T+1)\left(\frac{\overline{\gamma}}{T+1} - 1 + \log\left(\frac{T+1}{\overline{\gamma}} \right) \right). \tag{6}$$

Comparing the right hand sides of (4), and (6) it can be seen that they match only when $T=0$, and in this case, they also match the non-predictive diversity gain obtained in (59). Otherwise, for positive values of T, the two bounds differ.

2) Random Look-ahead Time: We consider a more general scenario where $T_{q(n)}$, $0 \leq q(n) \leq Q(n)$, $n > 0$ is a sequence of IID non-negative integer-valued random variables defined over a finite support $\{T_*, \cdots T^*\}$, where $0 \leq T_* \leq T^* < \infty$. The random variable $T_{q(n)}$ has the following probability mass function (PMF),

$$P\left(T_{q(n)} = k\right) \triangleq \begin{cases} p_k, & T_* \leq k \leq T^*, \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where $\sum_{k=T_*}^{T^*} p_k = 1$ and $p_k \geq 0, \quad \forall k$, the cumulative distribution function (CDF) of $T_{q(n)}$ can be written as

$$P(T_{q(n)} \leq k) = F_k = \begin{cases} 1, & k > T^*, \\ \sum_{i=T_*}^{k} p_i, & T_* \leq k \leq T^*, \\ 0, & k < T_*. \end{cases} \tag{8}$$

Hence, the overall process $Q(n)$ can be decomposed to a superposition of independent Poisson processes, i.e.,

$$Q(n) = \sum_{k=T_*}^{T^*} Q_k(n)$$

where $Q_k(n)$, $n > 0$ is the process of requests predicted k slots ahead, $k = T_*, \cdots, T^*$. The arrival rate of $Q_k(n)$ is $p_k \lambda$.

In this scenario, we denote the outage probability under EDF by $P_{PR}(\mathcal{O})$ and the optimal diversity gain by $d_{PR}(\gamma)$. Unlike the case of deterministic look-ahead time, EDF here does not reduce to FCFS because the arriving requests at the subsequent slots can have earlier deadlines than some of those who have already arrived. Upper and lower bounds on $P_{PR}(\mathcal{O})$ are introduced in the following lemma.

Lemma 2: Let

$$\mathcal{I} \triangleq \left\{ \sum_{j=0}^{T^*} \sum_{i=T_*}^{T^*} Q_i(n-i-j) > C(T^*+1), n \gg 1 \right\},$$

$$\mathcal{J} \triangleq \bigcup_{k=T_*}^{T^*-1} \left\{ \sum_{j=T_*}^{k} \sum_{i=T_*}^{j} Q_i(n-j) > C(k+1), n \gg 1 \right\},$$

$$\mathcal{U}_R \triangleq \mathcal{I} \bigcup \mathcal{J}$$

and

$$\mathcal{L}_R \triangleq \bigcup_{k=T_*}^{T^*} \left\{ \sum_{j=T_*}^{k} Q_j(n-j) > C(k+1), n \gg 1 \right\},$$

then, the events $\mathcal{U}_R$ and $\mathcal{L}_R$ constitute necessary and sufficient conditions on the outage event, respectively. Hence $P(\mathcal{L}_R) \leq P_{PR}(\mathcal{O}) \leq P(\mathcal{U}_R)$.
Here also, we assume the system is at steady state.

Proof: Please refer to Appendix D. ■

Theorem 3: Let

$$\overline{v}_* \triangleq \min_{T_* \leq k \leq T^*-1} \left\{ (k+1)\left[\log\left(\frac{k+1}{\overline{\gamma}\sum_{i=0}^{k-T_*} F_{k-i}}\right) - 1\right] + \overline{\gamma}\sum_{i=0}^{k-T_*} F_{k-i} \right\},$$

the optimal diversity gain of a proactive wireless network with random prediction interval, $d_{PR}(\gamma)$, satisfies

$$\overline{d}_{PR}(\overline{\gamma}) \geq \min\{(T^*+1)(\overline{\gamma}-1-\log\overline{\gamma}), \overline{v}_*\}, \quad 0 < \overline{\gamma} < 1 \tag{9}$$

for the linear scaling regime, and

$$\widetilde{d}_{PR}(\widetilde{\gamma}) = (T_* + 1)(1 - \widetilde{\gamma}), \quad 0 < \widetilde{\gamma} < 1, \tag{10}$$

for the polynomial scaling regime.

Proof: Please refer to Appendix E. ■

Theorem 3 determines a lower bound on the optimal prediction diversity gain of the linear scaling regime and fully characterizes the optimal prediction diversity. It is obvious that the lower bound on $\overline{d}_{PR}(\overline{\gamma})$ depends on the distribution of $T_{q(n)}$, however, this lower bound is always larger than $\overline{d}_N(\overline{\gamma})$ as long as $T^* > 0$ and $p_{T^*>0}$. This can be viewed by considering the term $(T^* + 1)(\overline{\gamma} - 1 - \log\overline{\gamma})$ which is strictly larger than $\overline{d}_N(\overline{\gamma})$ and $\overline{v}_*$ where for any k such that $T_* \leq k \leq T^* - 1$,

$$\begin{aligned}
&(k+1)\left[\overline{\gamma}\left(\frac{\sum_{i=0}^{k-T_*} F_{k-i}}{k+1}\right) - \log\frac{\sum_{i=0}^{k-T_*} F_{k-i}}{k+1} - 1 - \log\overline{\gamma}\right] \\
&\overset{(a)}{>} (k+1)(\overline{\gamma} - 1 - \log\overline{\gamma}) \\
&\overset{(b)}{\geq} \overline{d}_N(\overline{\gamma}).
\end{aligned}$$

Inequality (a) follows as

$$0 < \frac{\sum_{i=0}^{k-T_*} F_{k-i}}{k+1} < 1$$

and $\overline{\gamma}x - \log x > \overline{\gamma}, \quad \forall x \in (0,1)$, while inequality (b) follows because $k \geq T_* \geq 0$. Hence, the proactive network in linear scaling regime with $T^* > 0$ and $p_{T^*} > 0$ always improves the diversity gain.

For the polynomial scaling regime, Theorem 3 shows that the prediction diversity gain of a proactive wireless network with random look-ahead interval is dominated by arrivals with $T_{q(n)} = T_*$. Hence, the main drawback of this is that, if $T_* = 0$ the prediction diversity becomes tantamount to that of the non-predictive scenario. However, even though $T_* = 0$, the outage probability of the predictive network is evaluated numerically in Section VII and shown to outperform the non-predictive case.

IV. Heterogenous QoS Requirements

We consider two types of users with different QoS requirements, the first is a primary user who has the priority to access the network, whereas the second is a secondary user that is allowed to access the primary network resources opportunistically. That is, it can use the primary resources at any time slot only when there is sufficient capacity to serve all primary requests at that slot with the remaining capacity assigned to the secondary user. This type of opportunistic access to the primary network adds more utilization to its resources while it may get paid by the secondary user for the offered service.

The primary and secondary requests arrive to the network following two Poisson processes $Q^p(n), n > 0$ and $Q^s(n), n > 0$ with arrival rates λ^p and λ^s respectively. We also assume that the network is stable and dominated by primary arrivals as follows.

Assumption 1:

$$\lambda^s + \lambda^p < C, \tag{11}$$

$$\lambda^s < \lambda^p. \tag{12}$$

The network is reactive to the secondary requests and hence each secondary request will expire if it is not served in the same slot of arrival. In the following subsection, we analyze the performance of the secondary outage probability and diversity gain when the primary network is also reactive, then we proceed to the proactive case.

A. Non-predictive Primary Network

At the beginning of time slot n the network has $Q^p(n) + Q^s(n)$ arrivals that should be served within the same slot, i.e., all have a deadline of n. The network typically serves the primary requests before the secondary. Hence, the diversity gain of the primary network in this scheme, denoted $d_N^p(\gamma^p)$, follows the same expressions obtained in Theorem 1, i.e.,

$$\overline{d}_N^p(\overline{\gamma}^p) = \overline{\gamma}^p - 1 - \log\overline{\gamma}^p, \quad 0 < \overline{\gamma}^p < 1 \tag{13}$$

$$\widetilde{d}_N^p(\widetilde{\gamma}^p) = 1 - \widetilde{\gamma}^p, \quad 0 < \widetilde{\gamma}^p < 1, \tag{14}$$

where $\bar{\lambda}^p = \overline{\gamma}^p C$ and $\tilde{\lambda}^p = C^{\widetilde{\gamma}^p}$.

The secondary user, therefore, suffers an outage at time slot n if and only if

$$Q^p(n) + Q^s(n) > C, \quad Q^s(n) > 0.$$

Theorem 4: The diversity gain of the secondary network, $d_N^s(\gamma^p, \gamma^s)$, when the primary network is non-predictive, satisfies

$$\overline{d}_N^s(\overline{\gamma}^p, \overline{\gamma}^s) \leq \overline{\gamma}^p - 1 - \log\overline{\gamma}^p, \tag{15}$$

$$\overline{d}_N^s(\overline{\gamma}^p, \overline{\gamma}^s) \geq \overline{\gamma}^p + \overline{\gamma}^s - 1 - \log(\overline{\gamma}^p + \overline{\gamma}^s), \tag{16}$$

$$\widetilde{d}_N^s(\widetilde{\gamma}^p, \widetilde{\gamma}^s) = 1 - \widetilde{\gamma}^p, \tag{17}$$

where $\bar{\lambda}^s = \overline{\gamma}^s C$, $\tilde{\lambda}^s = C^{\widetilde{\gamma}^s}$ and $0 < \overline{\gamma}^s < \overline{\gamma}^p < 1$, $\overline{\gamma}^p + \overline{\gamma}^s < 1$ and $0 < \widetilde{\gamma}^s < \widetilde{\gamma}^p < 1$.

Proof: Please refer to Appendix F. ■

Theorem 4 reveals that the diversity gain of the secondary user, under non-predictive network, is at most equal to the diversity gain of the primary network in the linear scaling regime and is exactly equal to it in the polynomial scaling regime although the secondary user has strictly less traffic rate than the primary. It can also be noted that $\widetilde{d}_N^s(\widetilde{\gamma}^p, \widetilde{\gamma}^s)$ is independent of $\widetilde{\gamma}^s$, that is, regardless of how small $\widetilde{\gamma}^s$ is, the diversity gain of the secondary user is kept fixed at $\widetilde{d}_N^p(\widetilde{\gamma}^p)$

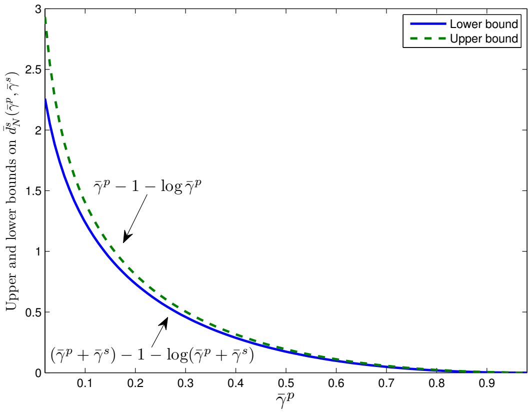


Fig. 2: The gap between the upper and lower bounds on $\overline{d}^s_N(\overline{\gamma}^p, \overline{\gamma}^s)$ declines when $\overline{\gamma}^s \ll \overline{\gamma}^p$. In this figure, $\overline{\gamma}^s = 0.02$ and $\overline{\gamma}^p \in (\overline{\gamma}^s, 1 - \overline{\gamma}^s)$.

as long as $\widetilde{\gamma}^s > 0$. The lower bound in (16), although does not match the upper bound in (15), it is always positive and approaches the upper bound when $\overline{\gamma}^s$ is much smaller than $\overline{\gamma}^p$ as shown in Fig. 2.

B. Predictive Primary Network

When the primary network is predictive, the arriving primary requests $Q^p(n)$, $n > 0$ are assumed to be predictable with a deterministic look-ahead time T. The secondary requests, $Q^s(n)$, conversely, are all urgent.

Let $N^p_i(n)$ be the number of all primary requests awaiting in the network at the beginning of time slot n with deadline $n + i$, $i = 0, \cdots, T$ and let $N^p(n) = \sum_{i=0}^{T} N^p_i(n)$.

1) Selfish Primary Scheduling: By a **selfish** primary behavior we mean the primary network has a dedicated capacity C per slot and no secondary request is served at the beginning of time slot n unless all primary requests $N^p(n)$ are served at this slot and $C - N^p(n) > 0$. The optimal prediction diversity gain and the outage probability of the primary network in this case are not affected by the presence of the secondary user. On the other hand, the selfish behavior of the primary predictive network cannot improve the outage probability of the secondary user. To show this, let $P_P(\mathcal{O}^s)$ denote the outage probability of the secondary user when the primary network is predictive. Then

$$\begin{aligned} P_P(\mathcal{O}^s) &= P(N^p(n) + Q^s(n) > C, Q^s(n) > 0) \\ &\geq P(Q^p(n) + Q^s(n) > C, Q^s(n) > 0) \qquad (18) \\ &= P_N(\mathcal{O}^s), \qquad (19) \end{aligned}$$

where inequality (18) follows since $N^p_T(n) = Q^p(n)$ and $N^p(n) \geq N^p_T(n)$. Here we note that the above result holds for any scheduling policy that serves all primary requests in the network at any slot before the secondary requests.

2) Cooperative Primary User: The predictive primary network, however, can act in a **less-selfish** manner without losing performance and, at the same time, enhance the diversity gain of the secondary user. This can be done by limiting the per-slot capacity dedicated to serve the primary requests in the system. One possible way to do so is to decide the capacity for the primary network dynamically at the beginning of each slot. We suggest the following less-selfish policy.

Definition 3: The number of primary requests to be served at slot n is denoted by $C^p(n)$ and given by

$$C^p(n) \triangleq \min \left\{ C, N^p_0(n) + f \times \sum_{i=1}^{T} N^p_i(n) \right\}, \qquad (20)$$

where $0 \leq f \leq 1$, and the primary requests are served according to EDF.
This scheme determines the maximum number of primary requests that the primary network can serve at the beginning of each slot depending on the number of primary requests with deadline at this slot as well as some factor of the number of other primary requests in the system. Hence, at the beginning of time slot n, arriving secondary requests will have the chance to get service if $C - C^p(n) > 0$, while the primary network has the capability to schedule the $C^p(n)$ requests according to a service policy that minimizes the primary outage probability (we address the EDF scheduling, however, for simplicity). In the above scheme, if f is chosen to be 1, the primary network will act selfishly, whereas $f = 0$ implies a performance of primary **non-predictive** network. In the following theorem we show that for some range of f, the diversity gain expressions for the primary network satisfy the same bounds of the selfish scenario.

Theorem 5: Under the dynamic capacity assignment policy in Def. 3 with $f \in [0.5, 1]$, the diversity gain of the primary network satisfies

$$\overline{d}^p_P(\overline{\gamma}^p) \geq (T + 1)(\overline{\gamma}^p - 1 - \log \overline{\gamma}^p), \quad 0 < \overline{\gamma}^p < 1, \qquad (21)$$

$$\widetilde{d}^s_P(\widetilde{\gamma}^p) = (T + 1)(1 - \widetilde{\gamma}^p), \quad 0 < \widetilde{\gamma}^p < 1. \qquad (22)$$

Proof: Please refer to Appendix G. ■

The above theorem thus shows that the predictive primary network satisfies the same diversity gain bounds of the selfish behavior under the proposed dynamic capacity assignment policy as long as $f \in [0.5, 1]$. Moreover, it gives a potential for improvement in the outage performance of the secondary users by limiting the number of primary requests served per slot. The outage probability of the secondary network in this case is given by

$$\begin{aligned} P_P(\mathcal{O}^s) &= P\left(Q^s(n) + C^p(n) > C, Q^s(n) > 0\right) \\ &= P\Bigg(Q^s(n) + \min\Big\{C, N^p_0(n) \\ &\qquad + f \sum_{i=1}^{T} N^p_i(n)\Big\} > C, Q^s(n) > 0\Bigg). \qquad (23) \end{aligned}$$

To show that even the diversity gain of the secondary network is improved under such policy, we consider the case when $f = 0.5$, and $T = 1$ for simplicity. In this case, the per-slot capacity of the primary network turns out to be

$$C^p(n) = \min\left\{C, N^p_0(n) + 0.5 Q^p(n)\right\} \qquad (24)$$

with

$$N_0^p(n+1) = \begin{cases} Q^p(n), & \text{if } N_0^p(n) = C, \\ 0.5Q^p(n) + N_0^p(n) - C, & \text{if } N_0^p(n) < C,\ N_0^p(n) \\ & \quad +0.5Q^p(n) \geq C, \\ 0.5Q^p(n), & \text{if } N_0^p(n) \\ & \quad +0.5Q^p(n) < C. \end{cases} \tag{25}$$

It is clear from (25) that

$$\begin{aligned} P(N_0^p(n+1) = l | N_0^p(n) = i, \cdots, N_0^p(1) = k) \\ = P(N_0^p(n+1) = l | N_0^p(n) = i). \end{aligned}$$

That is, the discrete-time random process $N_0^p(n), n > 0$ satisfies the Markov property, and hence, it is a Markov chain. Moreover, it can be easily verified that $N_0^p(n), n > 0$ is irreducible and aperiodic as $P(Q^p(n) = q) > 0$ for all $q \geq 0$.

The drift of the chain can thus be obtained as

$$E[N_0(n+1) - N_0(n) | N_0(n) = i] \begin{cases} \leq -(1-\gamma^p)C, & \text{if } i \geq C, \\ \leq \frac{\gamma^p}{2} C, & \text{if } i < C. \end{cases} \tag{26}$$

Then, by Foster's theorem [15], the Markov chain is positive recurrent, and hence has a stationary state distribution.

Theorem 6: Suppose that the system is operating at the stationary distribution of $N_0^p(n), n > 0$, the diversity gain of the secondary network, $d_P^s(\gamma^p, \gamma^s)$, under the dynamic capacity allocation for the primary satisfies

$$\overline{d}_P^s(\overline{\gamma}^p, \overline{\gamma}^s) \geq -\overline{\gamma}^s(\overline{y}^2 - 1) - 2\overline{\gamma}^p(\overline{y} - 1) + 2\log(\overline{y}), \tag{27}$$

where

$$\overline{y} = -\frac{\overline{\gamma}^p}{2\overline{\gamma}^s} + \frac{\sqrt{(4\overline{\gamma}^s + \overline{\gamma}^{p2})}}{2\overline{\gamma}^s}$$

and

$$\widetilde{d}_P^s(\widetilde{\gamma}^p, \widetilde{\gamma}^s) \geq \begin{cases} (1 - \widetilde{\gamma}^p), & 1 + \widetilde{\gamma}^s \geq 2\widetilde{\gamma}^p, \\ \frac{1}{2}(1 - \widetilde{\gamma}^s), & 1 + \widetilde{\gamma}^s < 2\widetilde{\gamma}^p. \end{cases} \tag{28}$$

Proof: Please refer to Appendix H. ■

The right hand side of inequality (27) will be shown in Section VII to be strictly larger than the right hand side of (15) for a range of $\overline{\gamma}^s$, which implies a strict improvement in the diversity gain of the secondary network without any loss in the diversity gain of the primary. However, the right hand side of inequality (28) shows that if $1 + \widetilde{\gamma}^s < 2\widetilde{\gamma}^p$, then the diversity gain of the secondary network is at least equal to its non-predictive counterpart.

V. Robustness to Prediction Errors

In the previous sections we have assumed that the prediction mechanism is error free, that is, all predicted requests are true and will arrive in future after exactly the same look-ahead period of prediction. Under this assumption, we managed to treat the predicted arrival process with deterministic look-ahead time as a delayed version of the actual arrival process. However, in practical scenarios, this is not necessarily the case. In this section we provide a model for the imperfect prediction process and investigate its effect on the prediction diversity gain with fixed look-ahead interval T assuming a single class of QoS.

Let $Q(n)$, $n > 0$ be the actual arrival process that the network should predict T slots ahead. This process, as introduced in Section II, is Poisson with rate λ. Because the prediction mechanism employed by the network may cause errors, the predicted arrival process differs from the actual arrival process. The prediction mechanism is supposed to cause two types of errors:

1) It predicts false requests, those will not arrive actually in future, and serves them, resulting in a waste of resources.
2) It fails to predict requests and, as a consequence, the network encounters urgent arrivals (unpredicted requests that should be served in the same slot of arrival).

So, we model the predicted process as

$$Q^E(n) = Q'(n) + Q''(n) \tag{29}$$

where $Q'(n)$, $n > 0$ is the arrival process of the predicted requests. It represents the number of arriving requests at the beginning of time slot n with deadline $n + T$. The process $Q''(n)$, $n > 0$ represents the number of unpredicted requests that arrive at the beginning of time slot n and must be served in the same slot because the network has failed to predict them. We assume further that $Q'(n)$ and $Q''(n)$ are independently Poisson distributed with arrival rates λ' and λ'', respectively.

Since $Q''(n)$ is a part of the requests $Q(n)$, then

$$0 \leq \lambda'' < \lambda \tag{30}$$

where the second inequality is strict because we assume that $Q'(n)$ contains truly predicted requests as well as mistakenly predicted requests, which also implies

$$\lambda' + \lambda'' \geq \lambda \tag{31}$$

Moreover, the network is stable as long as

$$\lambda' + \lambda'' < C. \tag{32}$$

For the linear scaling regime, the arrival processes $\overline{Q}'(n)$ and $\overline{Q}''(n)$, $n > 0$ have arrival rates $\overline{\alpha}'\overline{\gamma}C$ and $\overline{\alpha}''\overline{\gamma}C$ respectively. Applying conditions (30)-(32) to $\overline{\alpha}'\overline{\gamma}C$ and $\overline{\alpha}''\overline{\gamma}C$ we obtain

$$\overline{\alpha}'' < 1 \tag{33}$$

and

$$1 \leq \overline{\alpha}' + \overline{\alpha}'' < \frac{1}{\overline{\gamma}} \tag{34}$$

So, if the prediction mechanism is perfect, then $\overline{\alpha}' = 1$ whereas $\overline{\alpha}'' = 0$.

The arrival process $\overline{Q}^E(n)$, $n > 0$, can be considered as a predicted process with random look-ahead interval that takes on values 0 and T. Hence, using the event $\mathcal{U}_R$ defined in Lemma 2, we obtain the following lower bound on the prediction diversity gain, $\overline{d}_P^E(\overline{\gamma})$,

$$\begin{aligned} \overline{d}_P^E(\overline{\gamma}) \geq \min\{(T+1)\left[(\overline{\alpha}' + \overline{\alpha}'')\overline{\gamma} - 1 - \log\left(\overline{\gamma}(\overline{\alpha}' + \overline{\alpha}'')\right)\right], \\ \overline{\alpha}''\overline{\gamma} - 1 - \log(\overline{\alpha}''\overline{\gamma})\} \end{aligned} \tag{35}$$

The best operating point (prediction window) that maximizes the right hand side of (35) is when both terms in the $\min\{.\}$ are equal, which implies

$$\bar{T}_{crit} = \frac{\overline{\alpha}''\overline{\gamma} - 1 - \log(\overline{\alpha}''\overline{\gamma})}{(\overline{\alpha}' + \overline{\alpha}'')\overline{\gamma} - 1 - \log(\overline{\gamma}(\overline{\alpha}' + \overline{\alpha}''))}. \tag{36}$$

Since $\overline{\alpha}'' < 1$, then for $\bar{T}_{crit}$ derived in (36), we obtain $\overline{d}_P^E(\overline{\gamma}) > \overline{d}_N(\overline{\gamma})$.

For the polynomial scaling regime, the processes $\widetilde{Q}'(n)$ and $\widetilde{Q}''(n)$, $n > 0$ have arrival rates $C^{\widetilde{\alpha}'\widetilde{\gamma}}$ and $C^{\widetilde{\alpha}''\widetilde{\gamma}}$ respectively. Applying conditions (30)-(32) to the arrival rates $C^{\widetilde{\alpha}'\widetilde{\gamma}}$ and $C^{\widetilde{\alpha}''\widetilde{\gamma}}$, we obtain,

$$\widetilde{\alpha}'' < 1, \tag{37}$$

$$C^{\widetilde{\alpha}'\widetilde{\gamma}} + C^{\widetilde{\alpha}''\widetilde{\gamma}} \geq C^{\widetilde{\gamma}}, \tag{38}$$

and

$$C^{\widetilde{\alpha}'\widetilde{\gamma}} + C^{\widetilde{\alpha}''\widetilde{\gamma}} < C. \tag{39}$$

If the prediction mechanism is perfect, then $\widetilde{\alpha}' = 1$ whereas $\widetilde{\alpha}'' = -\infty$.

We also use events $\mathcal{U}_R$ and $\mathcal{L}_R$ from Lemma 2 to determine the prediction diversity gain with imperfect prediction mechanism, $\widetilde{d}_P^E(\widetilde{\gamma})$, as

$$\widetilde{d}_P^E(\widetilde{\gamma}) = \min\{(T+1)\left[1 - \max\{\widetilde{\alpha}', \widetilde{\alpha}''\}\widetilde{\gamma}\right], 1 - \widetilde{\alpha}''\widetilde{\gamma}\}. \tag{40}$$

Nevertheless, since at $\widetilde{d}_P^E(\widetilde{\gamma})$ is at $C \to \infty$, then from (38), (39), as $C \to \infty$, we obtain, $1 \leq \widetilde{\alpha}' < \frac{1}{\widetilde{\gamma}}$. And from (37), $\max\{\widetilde{\alpha}', \widetilde{\alpha}''\} = \widetilde{\alpha}'$. Hence,

$$\widetilde{d}_P^E(\widetilde{\gamma}) = \min\{(T+1)(1 - \widetilde{\alpha}'\widetilde{\gamma}), 1 - \widetilde{\alpha}''\widetilde{\gamma}\}. \tag{41}$$

So, to obtain the maximum diversity gain, the best prediction window $\tilde{T}_{crit}$ should satisfy

$$\tilde{T}_{crit} = \frac{(\widetilde{\alpha}' - \widetilde{\alpha}'')\widetilde{\gamma}}{1 - \widetilde{\alpha}''\widetilde{\gamma}}, \tag{42}$$

and at this point, since $\widetilde{\alpha}'' < 1$, we have $\widetilde{d}_P^E(\widetilde{\gamma}) > \widetilde{d}_N(\widetilde{\gamma})$.

This section hence has shown theoretically that even under imperfect prediction mechanisms, the prediction window size is judiciously chosen to strike the best balance between the predicted traffic and the urgent one.

VI. Proactive Scheduling in Multicast Networks

This section sheds light on the predictive multicast networks and investigates the diversity gains that can be leveraged from efficient scheduling of multicast traffic. Typically, multicast traffic minimizes the usage of the network resources because the same data is sent to a group of users consuming the same amount of resources that serve only a single user which is taken to be unity [16]. So, even in the non-predictive case, the multicast traffic is expected to result in an improved diversity gain performance over its unicast counterpart, discussed in the previous sections.

Furthermore, when the multicast traffic is predictable, there is an additional gain that can be obtained from the ability to *align* the traffic in time. That is, the network can keep on receiving predictable requests that target the same data over time then serve them altogether as the earliest deadline approaches. In this case, the network will end up serving all the gathered requests in a window of time slots with the same resources required to serve one request, and hence will significantly improve the diversity gain of the network. We assume that there are L data sources available (e.g. files, packets, movies, podcasts, etc.) for multicast transmission. The number of multicast requests arriving at the beginning of time slot $n > 0$ is a random variable $Q^m(n)$ which is assumed to be Poisson distributed with mean λ^m.

Assuming that the data sources are demanded independently across time and requests, the process $Q^m(n), n > 0$ can be decomposed into

$$Q^m(n) = \sum_{l=1}^{L} Q^{m,[l]}(n), \quad \text{for all } n > 0,$$

where $Q^{m,[l]}(n)$ denotes the number of multicast requests for data source $l \in \{1, \cdots, L\}$ arriving in slot n, and is Poisson distributed with mean $\lambda^{m,[l]} \triangleq p^{[l]}\lambda^m$, where $\mathbf{p} \triangleq (p^{[l]})_{l=1}^{L}$ is a valid probability distribution[1] capturing the potentially asymmetric multicast demands over the pool of L data sources.

In this section we focus only on the analysis of the linear scaling regime where the potential improvement in the diversity gain is tangible [2]. The mean number of arriving multicast requests scales with C as $\lambda^m = \overline{\gamma^m}C$, $\overline{\gamma^m} \in (0, 1)$. The number of data sources L scales also linearly with C as $L = \overline{\theta}C$, $\overline{\theta} > 0$.

The binary parameter $X^{m,[l]}(n)$ for each multicast data source $l \in \{1, \cdots, L\}$ is defined as

$$X^{m,[l]}(n) \triangleq \left\{ \begin{array}{ll} 1, & \text{if } Q^{m,[l]}(n) > 0, \\ 0, & \text{if } Q^{m,[l]}(n) = 0, \end{array} \right. \quad l = 1, \cdots, L, \tag{43}$$

which gives the indicator of at least one multicast request for data source l arrives at slot n. And, under the aforementioned Poisson assumptions, $X^{m,[l]}(n)$ is a simple Bernoulli random variable with parameter

$$A^{m,[l]} = 1 - e^{-p^{m,[l]}\lambda^m}, \quad l \in \{1, \cdots, L\}. \tag{44}$$

We denote the total number of distinct multicast data requests arriving in slot n as $S^m(n)$, defined as

$$S^m(n) \triangleq \sum_{l=1}^{L} X^{m,[l]}(n). \tag{45}$$

Definition 4: Let $N_0^{m,[l]}(n)$ denote the indicator that there is *at least* one awaiting multicast request for data source $l \in \{1, \cdots, L\}$ that expires in slot n. Then, letting $N_0^m(n) \triangleq \sum_{l=1}^{L} N_0^{m,[l]}(n)$, the multicast outage event is defined as

$$\overline{\mathcal{O}}_m \triangleq \left\{N_0^m(n) > C, n \gg 1\right\}.$$

The pure multicast network will be investigated in the following subsection where the diversity gain of its non-predictive side will be shown to be larger than its unicast counterpart, furthermore, the alignment property of the predictive multicast will be proven to result in a significantly improved diversity gain, that scales super-linearly with the prediction interval T. Then, the subsequent subsection will address a composite network consisting of unicast and multicast traffics; the potential diversity gain will be investigated under different prediction scenarios.

[1] $\mathbf{p}$ is a valid distribution if $0 \leq p^{[l]} \leq 1$ and $\sum_{l=1}^{L} p^{[l]} = 1$.

[2] The additional multicast gains do not appear in the polynomial scaling regime because the traffic to each data source vanishes asymptotically, as $C \to \infty$, when the number of data sources L scales with C, implying that at most one request can target a data source at each slot, i.e., the multicast traffic will approach the unicast as $C \to \infty$.

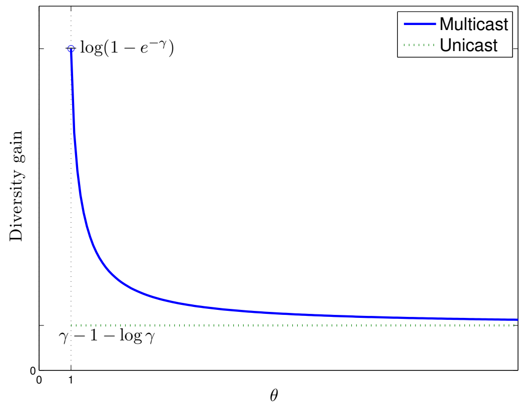


Fig. 3: Diversity gain of the non-predictive multicast network monotonically decreases with $\overline{\theta}$. However, it is lower bounded by the diversity gain of non-predictive unicast networks.

A. Symmetric Multicast Demands

Suppose that the number of data sources scales with C as $L=\overline{\theta}C$, $\overline{\theta}>0$. Then, $\overline{\theta}\leq 1$ implies zero outage probability and infinite diversity gain regardless of the value of $\overline{\gamma^m}$. This is the first gain improvement that can be leveraged from the nature of the multicast traffic. We now confine the analysis to the case when $\overline{\theta}>1$. Assume that the multicast demands are equally distributed on the available data sources, i.e.

$$p^{[l]}=p=\frac{1}{\overline{\theta}C},$$

$$A^{m,[l]}=A^m=1-e^{-\frac{\overline{\gamma^m}}{\overline{\theta}}},\quad \forall l\in\{1,\cdots,L\}.$$

1) Non-predictive Multicast Network: Under the above symmetric setup (and assuming $\overline{\theta}>1$), the random variable $S^m(n)$ turns out to have a binomial distribution with parameter A^m and the outage probability in this case, denoted by $P_N(\overline{\mathcal{O}}_m)$, is equal to $P(S^m(n)>C)$. In other words, the multicast outage occurs in slot n if and only if the number of distinct data sources requested at this slot is larger than the network capacity.

Theorem 7: The diversity gain of non-predictive multicasting, denoted by $\overline{d}_N(\overline{\gamma^m},\overline{\theta})$, is given by

$$\begin{aligned}\overline{d}_N(\overline{\gamma^m},\overline{\theta})&=(\overline{\theta}-1)\log(\overline{\theta}-1)-\overline{\theta}\log\overline{\theta}+\overline{\gamma^m}\left(\frac{\overline{\theta}-1}{\overline{\theta}}\right)\\&-\log\left(1-e^{-\frac{\overline{\gamma^m}}{\overline{\theta}}}\right),\quad 0<\overline{\gamma^m}<1,\quad \overline{\theta}>1.\end{aligned}\tag{46}$$

Proof: Please refer to Appendix I. ■

Theorem 7 and Fig. 3, which depicts the diversity gains of non-predictive multicast (46) and unicast (2) networks with $\overline{\gamma^m}=\overline{\gamma}$, show that $\overline{d}_N(\overline{\gamma^m},\overline{\theta})$ is monotonically decreasing in $\overline{\theta}$. As $\overline{\theta}$ increases, the number of data sources in the network grows faster with C, and hence, from (46),

$$\lim_{\overline{\theta}\to\infty}\overline{d}_N(\overline{\gamma^m},\overline{\theta})=\overline{\gamma^m}-\log\overline{\gamma^m}-1=\overline{d}_N(\overline{\gamma^m}).\tag{47}$$

That is, multicast diversity gain $\overline{d}_N(\overline{\gamma^m},\overline{\theta})$ is strictly greater than its unicast counterpart $\overline{d}_N(\overline{\gamma^m})$, and converges to it in the limit as $\overline{\theta}\to\infty$. In fact, a much stronger result is that, when $\overline{\gamma^m}=\overline{\gamma}$,

$$\begin{aligned}\lim_{\overline{\theta}\to\infty}LA^m&=\lim_{\overline{\theta}\to\infty}\overline{\theta}C\left(1-e^{-\frac{\overline{\gamma^m}}{\overline{\theta}}}\right)\\&=\overline{\gamma^m}C,\quad 0<\overline{\gamma^m}<1,\end{aligned}\tag{48}$$

we have also $A^m\to 0$ and $L=\overline{\theta}C\to\infty$ as $\overline{\theta}\to\infty$. Therefore, $S^m(n)$ converges in distribution to $\overline{Q}(n)$, and consequently, $P_N(\overline{\mathcal{O}}_m)\to P_N(\mathcal{O})$, $\overline{\theta}\to\infty$.

In this subsection, we have highlighted the extra diversity gain achieved through one of the multicast properties, that is all the requests arriving to the network at time slot n and demanding a certain data source are all served with one unit resources exactly as if only one request demands that data source.

2) Predictive Multicast Network: Now suppose that the symmetric multicast network has predictable demands with a prediction window of $T>0$ slots. The traffic alignment in this case appears in the following sense, the resource serving a group of requests arriving at slot n also serves all other requests in the system (that have arrived withing the previous T slots) requesting the same data source. So, the resource value is extendable across time. The prediction capability of the network is thus equal to infinity as long as $\overline{\theta}\leq(T+1)$, which implies a multiplicative gain of $T+1$ in the number of data sources that the network can support with zero outage probability, as compared to the non-predictive case.

Consider then the other range of $\overline{\theta}$, that is $\overline{\theta}>(T+1)$. The network now is subject to outage events and efficient scheduler has to be employed. Because of the symmetric demands, we focus the analysis on the EDF scheduling. Let the optimal diversity gain in this predictive scenario be denoted by $\overline{d}_P(\overline{\gamma^m},\overline{\theta})$, in [17], we have shown that $\overline{d}_P(\overline{\gamma^m},\overline{\theta})\geq(T+1)\overline{d}_N(\overline{\gamma^m},\overline{\theta})$ which is consistent with the results of Subsection III-B as the predictability multiplies the diversity gain by a factor of at least $T+1$. However, we show now that the alignment property can even improve the diversity gain and result in a super-linear scaling of $\overline{d}_P(\overline{\gamma^m},\overline{\theta})$ with T.

Theorem 8: The optimal diversity gain of the predictive multicast network with symmetric demands, $\overline{d}_P(\overline{\gamma^m},\overline{\theta})$, satisfies

$$\begin{aligned}\overline{d}_P(\overline{\gamma^m},\overline{\theta})&\geq(T+1)\log\left(\frac{(1-\xi_T^m)(T+1)}{\xi_T^m(\overline{\theta}-(T+1))}\right)\\&-\overline{\theta}\log\left(1-\xi_T^m+\frac{(1-\xi_T^m)(T+1)}{\overline{\theta}-(T+1)}\right),\\&\triangleq\mathbb{L}_{sym}.\end{aligned}\tag{49}$$

where

$$\xi_T^m=1-\exp\left(-\frac{(T+1)\overline{\gamma^m}}{\overline{\theta}}\right).$$

Proof: Please refer to Appendix J. ■

The new lower bound $\mathbb{L}_{sym}$ takes into account the alignment property of the predictable multicast traffic, and thus shows significant increase in the diversity gain with T as compared to the older bound $(T+1)\overline{d}_N(\overline{\gamma^m},\overline{\theta})$ in Fig. 4.

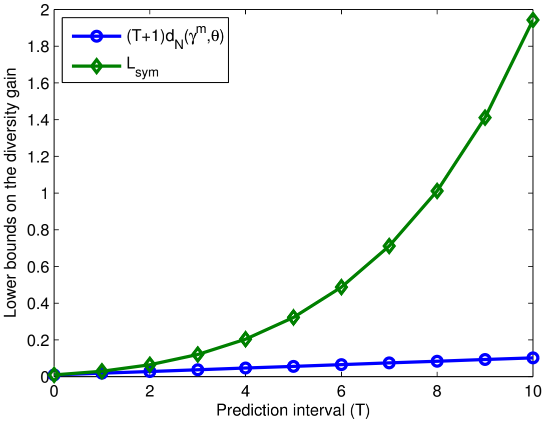


Fig. 4: Superlinear increase in the diversity gain of the multicast network with the prediction interval T because of the alignment property. In this figure $\overline{\gamma^m} = 0.9$ and $\overline{\theta} = 15$.

B. *Multicast and Unicast Traffic*

Generally, wireless networks support both types of traffic: multicast and unicast. For instance, a smart phone user my receive unicast data such as e-mail or electronic bank statement as well as multicast data such as movies or podcasts. In this subsection we investigate the potential diversity gain of wireless networks encompassing both types of traffic under different predictability conditions.

The multicast traffic model adopted here is exactly as defined in the beginning of this section, with the only difference is we assume that $L = \overline{\theta} C$, where $\overline{\theta} \in (0, 1)$. The multicast data sources are also equally demanded, each with probability

$$A^m = 1 - \exp\left(-\frac{\overline{\gamma^m}}{\overline{\theta}}\right).$$

The unicast traffic arrives at the beginning of each slot n according to $Q^u(n)$ which is Poisson distributed with mean $\lambda^u = \overline{\gamma^u} C$, $\overline{\gamma^u} \in (0, 1)$. Each of the unicast requests consumes one unit of the system capacity. The stability condition of the non-predictive network necessitates that

$$A^m \overline{\theta} + \overline{\gamma^u} < 1. \tag{50}$$

Definition 5: Letting $N_0^u(n)$ denote the number of unicast requests in the system at the beginning of time slot n, the combined outage event of the wireless network with unicast and multicast traffic is defined as

$$\overline{\mathcal{O}}_A = \left\{N_0^m(n) + N_0^u(n) > C, n \gg 1\right\}.$$

In [17], we have addressed the case when only on multicast data source exists in the network an consumes μC, $\mu \in (0, 1)$ of the available resources to supply data. This data source shares the network with unicast traffic. We have shown the impact of the multicast traffic alignment on the diversity gain where more gains can be leveraged by gathering more of the predictable multicast traffic and serving them altogether in a single slot. Alternatively, in this subsection we address the scenario of multiple data sources each consumes one unit of the available resources. We will investigate the diversity gain of the network in the following four scenarios of demand predictability:

1) Both unicast and multicast traffics are non-predictive.
2) Unicast is non-predictive but multicast is predictive.
3) Both unicast and multicast traffics are predictive.
4) Unicast is predictive but multicast is non-predictive.

1) Scenario 1: Both Unicast and Multicast Traffics are Non-predictive: In this scenario, all of the arriving requests are urgent and hence, $N_0^m(n) = S^m(n)$ and $N_0^u(n) = Q^u(n)$.

Theorem 9: Let the outage probability in Scenario 1 be denoted by $P_1(\overline{\mathcal{O}}_A)$ and the associated diversity gain be denoted by $\overline{d}_1(\overline{\gamma^u}, \overline{\gamma^m}, \overline{\theta})$, then

$$\begin{aligned}\overline{d}_1(\overline{\gamma^u}, \overline{\gamma^m}, \overline{\theta}) =& \log(y_1) + \gamma^u(1 - y_1) \\ & - \overline{\theta} \log\left(e^{-\frac{\overline{\gamma^m}}{\overline{\theta}}} + y_1\left(1 - e^{-\frac{\overline{\gamma^m}}{\overline{\theta}}}\right)\right),\end{aligned} \tag{51}$$

where

$$\begin{aligned}y_1 =& \frac{1}{2\overline{\gamma^u}\left(e^{\frac{\overline{\gamma^m}}{\overline{\theta}}} - 1\right)}\Bigg[\bigg((\overline{\theta}^2 - 2\overline{\theta} + 1)e^{\frac{2\overline{\gamma^m}}{\overline{\theta}}} \\ & + \left(-2\overline{\theta}^2 + 2\overline{\theta}(\overline{\gamma^u} + 2) + 2(\overline{\gamma^u} - 2)\right)e^{\frac{\overline{\gamma^m}}{\overline{\theta}}} + \overline{\theta}^2 \\ & - 2\overline{\theta}(\overline{\gamma^u} + 1) + \overline{\gamma^u}^2 - 2\overline{\gamma^u} + 1\bigg)^{\frac{1}{2}} + (1 - \overline{\theta})e^{\frac{\overline{\gamma^m}}{\overline{\theta}}} \\ & + \overline{\theta} - \overline{\gamma^u} - 1\Bigg].\end{aligned}$$

Proof: Please refer to Appendix K. ■

Theorem 9 thus tightly characterizes the diversity gain of the network in Scenario 1. The expression of $\overline{d}_1(\overline{\gamma^u}, \overline{\gamma^m}, \overline{\theta})$, however, is not insightful, so it will be compared graphically to the results of the other scenarios.

2) Scenario 2: Unicast is Non-predictive but Multicast is Predictive: In this scenario, the network can predict the multicast requests T slots ahead, whereas the unicast traffic is urgent. We consider a scheduling policy π_2 to establish a lower bound on the optimal diversity gain, denoted $\overline{d}_2(\overline{\gamma^u}, \overline{\gamma^m}, \overline{\theta})$, of this scenario.

Definition 6 (Scheduling Policy π_2): At each slot n, the scheduling policy π_2 serves as much as possible of the existing requests in the system in the following order:

1) Multicast data sources demanded by urgent requests, $N_0^u(n)$.
2) Unicast requests, $Q^u(n)$.
3) The rest of the multicast data sources according to EDF.

The policy π_2 is a slightly modified version of EDF with priority given to urgent multicast requests.

Theorem 10: Let the outage probability in Scenario 2 under the scheduling policy π_2 be denoted $P_2(\overline{\mathcal{O}}_A)$ and the optimal diversity gain be denoted by $\overline{d}_2(\overline{\gamma^u}, \overline{\gamma^m}, \overline{\theta})$, then

$$\begin{aligned}\overline{d}_2(\overline{\gamma^u}, \overline{\gamma^m}, \overline{\theta}) \leq& \min\Big\{\overline{d}_N(\overline{\gamma^u}), (T + 1)\log y_2 \\ & - (T + 1)\overline{\gamma^u}(y_2 - 1) \\ & - \overline{\theta}\log(1 - \xi_T^m + \xi_T^m y_2)\Big\}, \\ \triangleq& \, \mathbb{L}_2.\end{aligned} \tag{52}$$

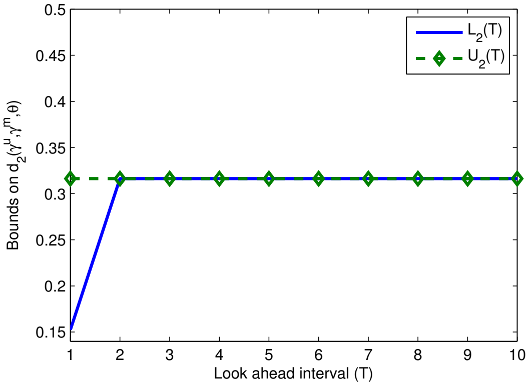


Fig. 5: As T increases, the system attains the same diversity gain of the non-predictive unicast network. In this figure, $\overline{\theta} = 0.7$, $\overline{\gamma^m} = 0.9$ and $\overline{\gamma^u} = 0.4$.

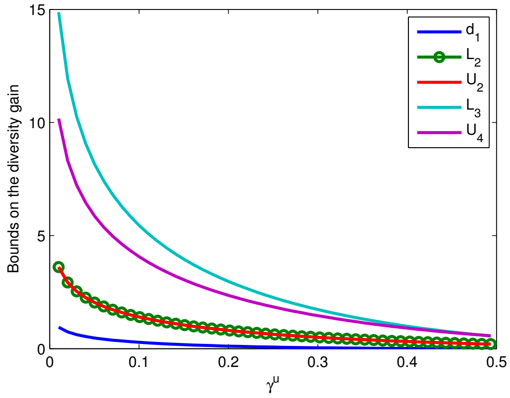


Fig. 6: Bounds on the optimal diversity gain versus the unicast traffic factor $\overline{\gamma^u}$. In this figure, $\overline{\gamma^m} = 0.9$, $\overline{\theta} = 0.7$ and $T = 4$ for any predictive network.

and

$$\overline{d}_2(\overline{\gamma^u}, \overline{\gamma^m}, \overline{\theta}) \leq \overline{d}_N(\overline{\gamma^u}) \triangleq \mathbb{U}_2, \tag{53}$$

where $\overline{d}_N(\overline{\gamma^u})$ is as derived in (2) with $\overline{\gamma} = \overline{\gamma^u}$, and

$$\begin{aligned} y_2 =& \frac{1}{2\xi_T^m \overline{\gamma^u}(T+1)} \Bigg[\Bigg(\bigg((1-\xi_T^m)^2 \overline{\gamma^u}^2 + 2\xi_T^m \overline{\gamma^u}(1-\xi_T^m) \\ &+ \xi_T^{m\,2} \bigg)^2 T^2 + \Big([2\xi_T^m \overline{\gamma^u}(1-\xi_T^m) - 2\xi_T^{m\,2}]\overline{\theta} \\ &+ 2\xi_T^{m\,2}(1-\xi_T^m)^2 + 4\xi_T^m \overline{\gamma^u}(1-\xi_T^m) + 2\xi_T^{m\,2} \Big) T \\ &+ [2\xi_T^m \overline{\theta}(1-\xi_T^m) - 2\xi_T^{m\,2}]\overline{\theta} + \overline{\gamma^u}^2(1-\xi_T^m)^2 \\ &+ 2\xi_T^m \overline{\theta}(1-\xi_T^m) + \xi_T^{m\,2}(1+\overline{\theta})^2 \Bigg)^{\frac{1}{2}} \\ &+ \Big((\xi_T^m - 1)\overline{\gamma^u} \Big) T - \xi_T^m \overline{\theta} + \overline{\gamma^u}(\xi_T^m - 1) + \xi_T^m \Bigg]. \end{aligned}$$

Proof: Please refer to Appendix L. ■

The upper and lower bounds on $\overline{d}_2(\overline{\gamma^u}, \overline{\gamma^m}, \overline{\theta})$ established in Theorem 10 match each other as T increases. In fact, the second term in $\min\{.,.\}$ of expression (52) is monotonically increasing in T, and hence $\exists t$ such that $T \geq t$ implies $\overline{d}_2(\overline{\gamma^u}, \overline{\gamma^m}, \overline{\theta}) = \overline{d}_N(\overline{\gamma^u})$. This result means that, efficient scheduling of the predictable multicast traffic results in the same diversity gain that will be obtained if the system sees only the unicast traffic. This result is clarified in Fig. 5 where the lower bound $\mathbb{L}_2$ increases in T until it becomes dominated by $\overline{d}_N(\overline{\gamma^u})$ at $T = 2$, and from this point on, $\mathbb{L}_2$ and $\mathbb{U}_2$ coincide and the diversity gain of the network is only determined by the non-predictive unicast traffic.

3) Scenario 3: Both Unicast and Multicast Traffics are Predictive: In this scenario we assume that both traffics are predictable with the same look-ahead interval of T slots. The scheduling policy we consider is EDF where requests are served in the order of their arrival.

Theorem 11: Let the outage probability of the network in Scenario 2 under EDF scheduling policy be denoted by $P_3(\overline{\mathcal{O}}_A)$ and the optimal diversity gain of this scenario be denoted by $\overline{d}_3(\overline{\gamma^u}, \overline{\gamma^m}, \overline{\theta})$, then

$$\begin{aligned} \overline{d}_3(\overline{\gamma^u}, \overline{\gamma^m}, \overline{\theta}) \geq& (T+1)\log y_2 - (T+1)\overline{\gamma^u}(y_2 - 1) \\ &- \overline{\theta}\log(1 - \xi_T^m + \xi_T^m y_2) \\ &\triangleq \mathbb{L}_3. \end{aligned} \tag{54}$$

Proof: Please refer to Appendix M. ■

In Scenario 3 one should expect that the optimal diversity gain should be the largest amongst the other three scenarios. To highlight this intuition, an upper bound will be established on the diversity gain of Scenario 4.

4) Scenario 4: Unicast is Predictive but Multicast is Non-predictive: Assuming that the unicast traffic is predictable with a look-ahead window of T slots, and the multicast traffic is urgent.

Theorem 12: Let the optimal diversity gain of Scenario 4 be denoted by $\overline{d}_4(\overline{\gamma^u}, \overline{\gamma^m}, \overline{\theta})$ and the minimum possible outage probability be denoted by $P_4^*(\overline{\mathcal{O}}_A)$, then

$$\begin{aligned} \overline{d}_4(\overline{\gamma^u}, \overline{\gamma^m}, \overline{\theta}) \leq& \overline{d}_1(\overline{\gamma^u}, \overline{\gamma^m}, \overline{\theta}) + T\Big[2\log y_4 - \overline{\gamma^u}(y_4 - 1) \\ &- 2\overline{\theta}\log(1 - A^m + A^m y_4) \Big] \\ &\triangleq \mathbb{U}_4, \end{aligned} \tag{55}$$

where

$$\begin{aligned} y_4 = \frac{1}{2\overline{\gamma^u} A^m} \Bigg[& \Big((4\overline{\theta}^2 - 4\overline{\theta}(\overline{\gamma^u} + 2) + (2 - \overline{\gamma^u})^2) A^{m\,2} + \overline{\gamma^u}^2 \\ &+ (4\overline{\gamma^u}\overline{\theta} - 2\overline{\gamma^u}^2 + 4\overline{\gamma^u}) A^m \Big)^{\frac{1}{2}} \\ &+ (-2\overline{\theta} + \overline{\gamma^u} + 2) A^m - \overline{\gamma^u} \Bigg]. \end{aligned}$$

Proof: Please refer to Appendix N. ■

To collectively compare the obtained bounds on the optimal diversity gain of the discussed scenarios, Fig.6 plots the different bounds obtained in the last four theorems versus $\overline{\gamma^u}$, where the range of $\overline{\gamma^u}$ ensures that (50) is satisfied, and hence the non-predictive network always sees a positive diversity gain. It is clear from the figure that the totally predictive

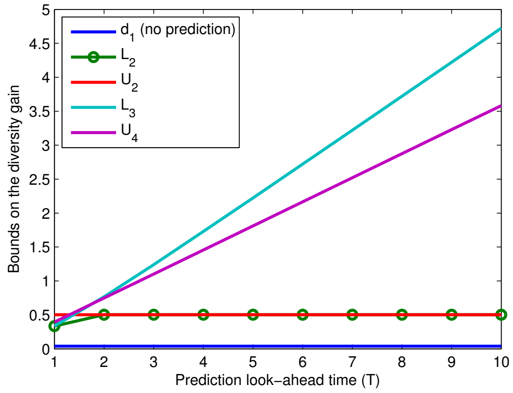


Fig. 7: Bounds on the optimal diversity gain versus the prediction look-ahead time T. In this figure, $\overline{\gamma^u} = 0.4$, $\overline{\gamma^m} = 0.9$ and $\overline{\theta} = 0.7$.

network (of Scenario 3) has the highest possible diversity gain as the lower bound $\mathbb{L}_3$ even exceeds the upper bound $\mathbb{U}_4$ on the entire range of plotted $\overline{\gamma^u}$. Also, it shows that $\mathbb{L}_2$ and $\mathbb{U}_2$ are coinciding at $\overline{d}_N(\overline{\gamma^u})$, and of course this is the best diversity gain that the network can achieve with unpredictable unicast traffic.

Also, Fig. 7 demonstrates the effect of the prediction look-ahead period T on the derived bounds. It shows that both $\mathbb{L}_3$ and $\mathbb{U}_4$ are both increasing in T, and that as T increases $\mathbb{L}_3$ exceeds $\mathbb{U}_4$ and $\mathbb{L}_2$ matches $\mathbb{U}_2$.

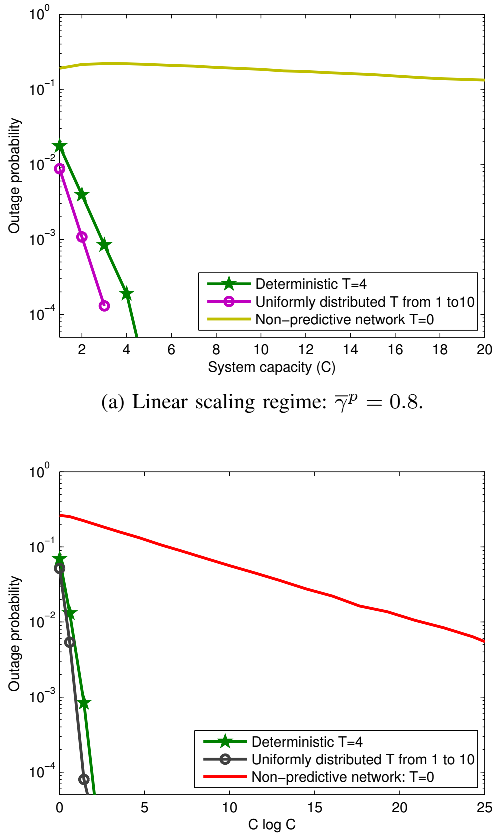
VII. Simulation Results

The analytical results obtained in this paper are demonstrated through numerical simulations in this section. The outage probability is quantified as the ratio of the number of slots that suffer expired requests to the total number of simulated slots. Each simulation result is obtained by averaging a 100 sample paths each contains a 1000 slots.

A. Diversity Gain of Deterministic and Random T Scenarios

Fig. 8 compares the outage probability of proactive networks with different look-ahead schemes to the non-predictive network. The results obtained for the linear scaling regime are plotted versus C in Fig. 8a and for the polynomial scaling regime are plotted versus $C \log C$ in Fig. 8b. It is obvious from both figures that being proactive significantly enhances the outage probability performance at a given capacity, or it considerably reduces the required capacity to satisfy a given level of outage performance. This ascribes to the more flexibility given to the predictive network that allows it to schedule the arriving requests over a longer time horizon compared to the urgent demand of the non-predictive network. The effect of the distribution of random look-ahead prediction interval is demonstrated in Fig. 9 for both the linear and polynomial scaling regimes.

The predictive network in each regime is assumed to anticipate requests by a random period which varies between T_* and T^* where $T_* = 0$ and $T^* = 5$. We consider a general binomial distribution with parameter p, $0 \leq p \leq 1$ to represent the



(a) Linear scaling regime: $\overline{\gamma}^p = 0.8$.

(b) Polynomial scaling regime: $\widetilde{\gamma}^p = 0.8$.

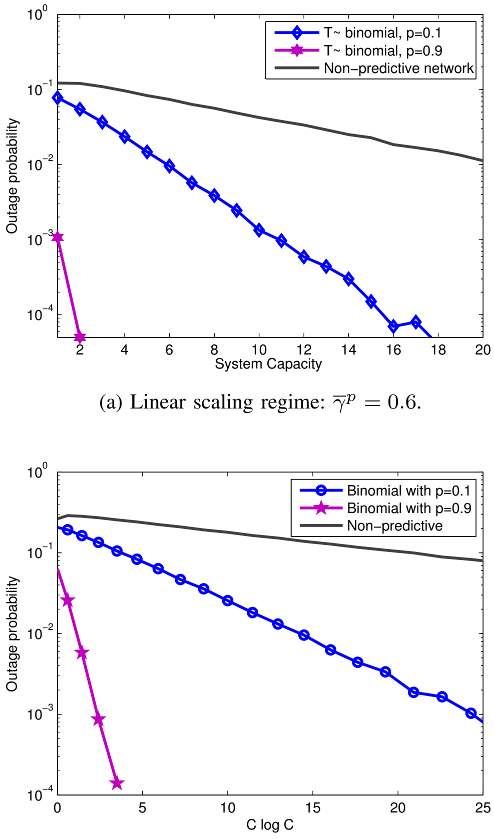
Fig. 8: Outage probability is significantly improved by proactive networks.

PDF of the look-ahead interval. Hence, the probability that an arriving request at the beginning of time slot n has a deadline at slot $n + T$, $T_* \leq T \leq T^*$, is given by

$$P(T_{q(n)} = T) = p_T = \binom{T^*}{T} p^T (1-p)^{T^*-T}. \tag{56}$$

We consider different values of p in each regime in addition to the non-predictive network scenario. The obtained results for the linear scaling regime are shown in Fig. 9a where at $p = 0.1$, $\overline{d}_{PR}(\overline{\gamma}) \geq \overline{\gamma} p_0 - \log(\overline{\gamma} p_0) - 1$, and $\overline{d}_{PR}(\overline{\gamma}) = (T^*{+}1)(\overline{\gamma}{-}1{-}\log\overline{\gamma})$ at $p = 0.9$. The results of the polynomial scaling regime are shown in Fig. 9b. Although the diversity gain is tantamount to that of the non-predictive network, it is clear from the figure that the outage probability is significantly improved. Here, we want to point out that diversity gain represents the asymptotic decay rate of the outage probability with the system capacity (or $C \log C$), but it does not capture the relative difference between the outage probability curves themselves. This is why the curves show different trends at small values of C. After all, the figure shows that even if $T_* = 0$ the network achieves a significantly better outage performance when it follows a proactive resource allocation technique.

Finally, from Figs. 9a, 9b, we can roughly infer that as p in creases, it is more likely to have arriving requests with larger prediction interval and hence the network gets more degrees



(a) Linear scaling regime: $\overline{\gamma}^p = 0.6$.

(b) Polynomial scaling regime: $\widetilde{\gamma}^p = 0.9$.

Fig. 9: Outage probability is significantly improved by proactive networks.

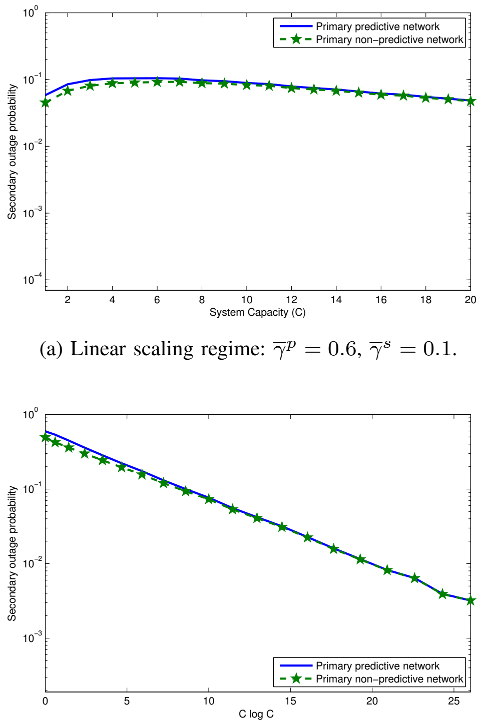
of freedom in scheduling such requests in an efficient way that reduces the number of outage events.

B. Two-QoS Network

Fig. 10 demonstrates the result (19) for both the linear scaling and polynomial scaling regimes. The simulation is run assuming 10^3 time slots and averaged over 10^2 sample paths. For the selfish predictive primary network, we assume that $T = 4$ and the primary requests are served according to EDF. The results of the linear scaling regime are depicted in Fig. 10a, whereas that of the polynomial scaling regime are depicted in Fig. 10b.

Figure 11 shows the potential improvement in the diversity gain of the secondary network by efficient use of prediction at the primary side only. Also, simulation results and analytical results are plotted together on the same figure to show the relative differences.

The performance of the dynamic-primary-capacity scheme, has been evaluated numerically and plotted in Fig. 12 for different values of f and under the two scaling regimes, namely, the linear scaling in Fig. 12a and the polynomial scaling in Fig. 12b. The prediction interval is chosen to be $T = 4$ and at each slot n, the primary network is assumed to serve the $C^p(n)$ primary requests according to EDF policy. For the two schemes, the selfish primary network, at $f = 1$, results in the smallest primary outage probability, while at $f = 0.5$, the primary outage probability is slightly increased beyond the selfish case, but the secondary outage probability outperforms its counterpart of the non-predictive primary network obtained at $f = 0$. It is clear from the figures that at $f = 0.5$ the secondary outage probability achieves the primary outage probability of the primary non-predictive network at $f = 0$ in the linear scaling regime, and is even better in the polynomial



(a) Linear scaling regime: $\overline{\gamma}^p = 0.6$, $\overline{\gamma}^s = 0.1$.

(b) Polynomial scaling regime: $\widetilde{\gamma}^p = 0.75$, $\widetilde{\gamma}^s = 0.05$.

Fig. 10: Selfish primary predictive network cannot improve the outage probability of the secondary.

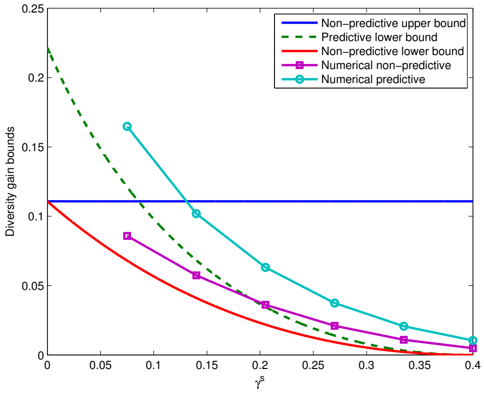
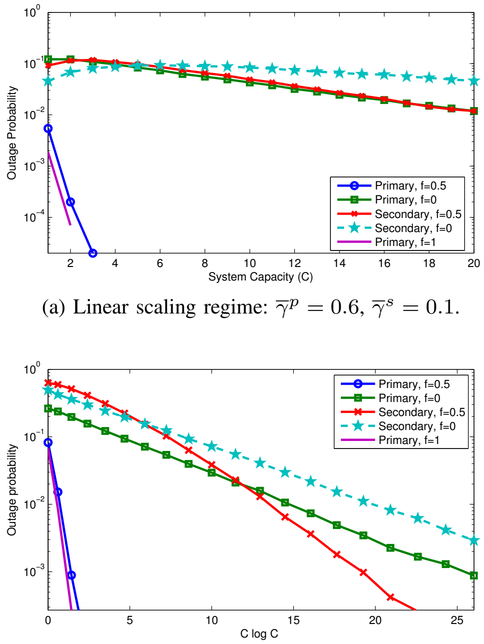


Fig. 11: Improvement in the diversity gain of the secondary network under predictive primary with $T = 1$ and dynamic capacity assignment. Considered in the figure is the linear scaling regime with $\overline{\gamma}^p = 0.6$. The lower bound on $\overline{d}_P(\overline{\gamma}^p, \overline{\gamma}^s)$ is shown in red, and obviously it strictly exceeds the upper bound on $\overline{d}_N(\overline{\gamma}^p, \overline{\gamma}^s)$ determined in Theorem 4 plotted in blue.



(a) Linear scaling regime: $\overline{\gamma}^p = 0.6$, $\overline{\gamma}^s = 0.1$.

(b) Polynomial scaling regime: $\widetilde{\gamma}^p = 0.75$, $\widetilde{\gamma}^s = 0.05$.

Fig. 12: Primary predictive network can tolerate a trivial loss in outage probability at a significant improvement in the secondary outage probability.

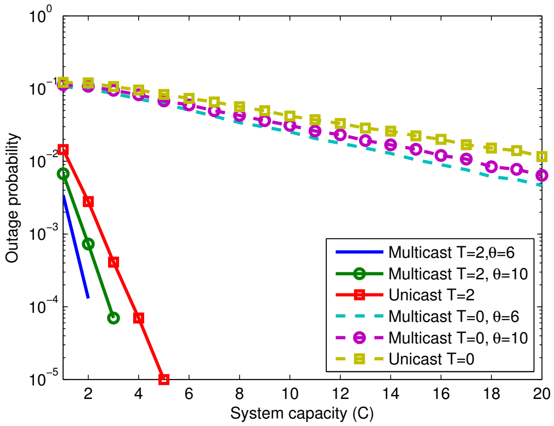


Fig. 13: Outage probability versus C. In this simulation, $\overline{\gamma^u} = 0.6$.

scaling regime. The simulation is for 10^3 time slots averaged over 10^2 sample paths.

C. Proactive Multicasting with Symmetric Demands

The outage probability of the predictive multicast and unicast networks of the symmetric input traffic is compared numerically to that of non-predictive network and is plotted in Fig. 13. The figure shows the significant enhancement to the outage probability of the multicast network when prediction is employed. Moreover, we can see that the outage probability of the unicast predictive network is better than that of the multicast non-predictive network. The impact of $\overline{\theta}$ also appears clearly, as it can easily be noticed that as the $\overline{\theta}$ decreases, the outage performance is enhanced even for the same value of T. When $\overline{\theta} \to \infty$ the multicast curves coincide on the unicast as shown in Section VI.

VIII. Conclusion and Discussion

We have proposed a novel paradigm for wireless resource allocation which exploits the predictability of user behavior to minimize the spectral resources (e.g., bandwidth) needed to achieve certain QoS metrics. Unlike the traditional reactive resource allocation approach in which the network can only start serving a particular user request upon its initiation, our proposed scheme anticipates future requests. This grants the network more flexibility in scheduling those potential requests over an extended period of time. By adopting the outage (blocking) probability as our QoS metric, we have established the potential of the proposed framework to achieve significant spectral efficiency gains in several interesting scenarios.

More specifically, we have introduced the notion of prediction diversity gain and used it to quantify the gain offered by the proposed resource allocation algorithm under different assumption on the performance of the traffic prediction technique. Moreover, we have shown that, in the cognitive network scenario, prediction at one side only does not only enhance its diversity gain, but it also improves the diversity gain performance of the other user class. On the multicasting front, we have shown that the diversity gain of predictive multicast network scales super-linearly with the prediction window. Our theoretical claims were supported by numerical results that demonstrate the remarkable gains that can be leveraged from the proposed techniques.

We believe that this work has only scratched the surface of a very interesting research area which spans several disciplines and could potentially have a significant impact on the design of future wireless networks. In fact, one can immediately identify a multitude of interesting research problems at the intersection of information theory, machine learning, behavioral science, and networking. For example, the analysis have focused on the case of fixed supply and variable demand. Clearly, the same approach can be used to match demand with supply under more general assumptions on the two processes.

Appendix A
Proof of Theorem 1

Let $\Lambda_Q(r)$ denote the log moment generating function [12] of a Poisson random variable $Q(n), n > 0$ with mean λ, i.e.,

$$\Lambda_Q(r) = \lambda(e^r - 1), \quad r \in \mathbb{R}.$$

For the linear scaling regime, let $\overline{X}_i$, $i = 1, 2, \cdots$ be a sequence of independent and identically distributed (IID) random variables, each with a Poisson distribution with mean $\overline{\gamma}$, and define

$$\overline{S}_C \triangleq \sum_{i=1}^{C} \overline{X}_i.$$

The outage probability, $P_N(\overline{\mathcal{O}})$, can then be written as

$$\begin{aligned} P_N(\overline{\mathcal{O}}) &= P(\overline{Q}(n) > C) \\ &= P\left(\frac{\overline{S}_C}{C} > 1\right) \end{aligned} \tag{57}$$

Applying Cramer's theorem [12] to (57), we get

$$\lim_{C\to\infty}\frac{1}{C}\log P\left(\frac{\overline{S}_C}{C}>1\right)=\inf_{r>0}\left\{\Lambda_X(r)-r\right\}, \tag{58}$$

where $\Lambda_X(r)=\overline{\gamma}(e^r-1)$. By the convexity of the log moment generating function, we obtain

$$\inf_{r>0}\left\{\Lambda_X(r)-r\right\}=1-\overline{\gamma}+\log\overline{\gamma}.$$

Then, it follows that

$$\begin{aligned}\overline{d}_N(\overline{\gamma}) &= -\lim_{C\to\infty}\frac{\log(P(\overline{\mathcal{O}}))}{C}\\ &=\overline{\gamma}-1-\log\overline{\gamma}, \quad 0<\overline{\gamma}<1.\end{aligned}\tag{59}$$

For the polynomial scaling regime, we determine the diversity gain using tight lower and upper bounds. First, the outage probability is given by

$$\begin{aligned}P_N(\widetilde{\mathcal{O}}) &= P(\widetilde{Q}(n)>C)\\ &=\sum_{k=C+1}^{\infty}\frac{(C^{\widetilde{\gamma}})^k}{k!}e^{-C^{\widetilde{\gamma}}}\\ &\geq\frac{(C^{\widetilde{\gamma}})^{(C+1)}}{(C+1)!}e^{-C^{\widetilde{\gamma}}}.\end{aligned}\tag{60}$$

Using Stirling's formula to approximate the factorial function, we have

$$(C+1)!\doteq\sqrt{2\pi(C+1)}\left(\frac{C+1}{e}\right)^{(C+1)},$$

where $\doteq$ means that the left hand side approaches the right hand side in the limit as $C\to\infty$. Hence,

$$\lim_{C\to\infty}-\frac{\log P_N(\widetilde{\mathcal{O}})}{C\log C}\leq$$

$$\lim_{C\to\infty}-\frac{1}{C\log C}\log\left(\frac{e^{-C^{\widetilde{\gamma}}}}{\sqrt{2\pi(C+1)}}\left(\frac{C^{\widetilde{\gamma}}e}{C+1}\right)^{C+1}\right)$$

Therefore,

$$\widetilde{d}_N(\widetilde{\gamma})\leq 1-\widetilde{\gamma}. \tag{61}$$

Second, applying tightest Chernoff bound [12] on (60), we have

$$P(\widetilde{Q}(n)>C)\leq\inf_{r>0}e^{\Lambda_{\widetilde{Q}}(r)-rC} \tag{62}$$

where $\Lambda_{\widetilde{Q}}(r)=C^{\widetilde{\gamma}}(e^r-1)$. And since $\Lambda_{\widetilde{Q}}(r)-r$ is convex in r, by simple differentiation, we get

$$P_N(\widetilde{\mathcal{O}})\leq e^{C-C^{\widetilde{\gamma}}-(1-\widetilde{\gamma})C\log C}. \tag{63}$$

Now, taking the logarithm of both sides of (63), dividing by $-C\log C$, and letting $C\to\infty$, it follows that

$$\widetilde{d}_N(\widetilde{\gamma})\geq 1-\widetilde{\gamma}. \tag{64}$$

By (61), (64),

$$1-\widetilde{\gamma}\leq\widetilde{d}_N(\widetilde{\gamma})\leq 1-\widetilde{\gamma},$$

then

$$\widetilde{d}_N(\widetilde{\gamma})=1-\widetilde{\gamma}, \quad 0<\widetilde{\gamma}<1. \tag{65}$$

Appendix B
Proof of Lemma 1

For $\mathcal{U}_D$, we need to show that the outage occurring at time slot n implies $\sum_{i=0}^{T}Q(n-T-i)>C(T+1)$. To see this, assume there is an outage at slot n. Since in our scenario EDF reduces to FCFS, then: 1) the outage at slot n occurs only on the arrivals of slot $n-T$ and 2) during the interval of slots $n-T, n-T+1,\cdots,n$, the system does not serve any of the arriving requests at slots beyond $n-T$. Let $N(m), m>0$ denote the number of requests in the system at the beginning of slot m, then having an outage at slot n implies $N(n-T)>C(T+1)$. And since at any slot $m>0$, there are no requests in the system arriving at slots prior to $m-T$, it follows that $\sum_{i=0}^{T}Q(n-T-i)\geq N(n-T)>C(T+1)$.

For $\mathcal{L}_D$, we need to show that $Q(n-T)>C(T+1)$ implies an outage at slot n. This is straightforward as the arrivals at slot $n-T$ can not remain in the system at any slot beyond n, furthermore, since $Q(n-T)>C(T+1)$, the capacity of the system at the slot of arrival in addition to the next T slots is not sufficient to serve the $Q(n-T)$ requests, hence the system encounters an outage at slot n.

Appendix C
Proof of Theorem 2

For the linear scaling regime, we have from Lemma 1, $P(\overline{\mathcal{O}})_{PD}\leq P(\overline{\mathcal{U}_D})$, hence,

$$P_{PD}(\overline{\mathcal{O}})\leq P\left(\sum_{i=0}^{T}\overline{Q}(n-T-i)>C(T+1)\right). \tag{66}$$

Using the same definition of the sequence of IID random variables $X_i, i>0$ as in the proof of Theorem 1, we have $\overline{S}_{C(T+1)}=\sum_{i=1}^{C(T+1)}X_i$ and

$$P\left(\sum_{i=0}^{T}\overline{Q}(n-T-i)>C(T+1)\right)=P\left(\frac{\overline{S}_{C(T+1)}}{C(T+1)}>1\right). \tag{67}$$

Using Cramer's theorem,

$$\lim_{C\to\infty}-\frac{\log P(\overline{\mathcal{U}_D})}{C(T+1)}=\overline{\gamma}-1-\log\overline{\gamma}. \tag{68}$$

Since $P_{PD}^{*}(\overline{\mathcal{O}})\leq P_{PD}(\overline{\mathcal{O}})\leq P(\overline{\mathcal{U}_D})$, we have

$$\begin{aligned}\lim_{C\to\infty}-\frac{\log P_{PD}^{*}(\overline{\mathcal{O}})}{C} &\geq\lim_{C\to\infty}-\frac{\log P(\overline{\mathcal{U}_D})}{C}\\ &=(T+1)(\overline{\gamma}-1-\log\overline{\gamma}),\end{aligned}\tag{69}$$

for which (4) follows.

For the polynomial scaling regime, first we use the upper bound $P_{PD}(\widetilde{\mathcal{O}})\leq P(\widetilde{\mathcal{U}_D})$ to establish a lower bound on the optimal diversity gain $\widetilde{d}_{PD}(\widetilde{\gamma})$ as follows. Using Chernoff bound on $P(\widetilde{\mathcal{U}_D})$,

$$\begin{aligned}P_{PD}(\widetilde{\mathcal{O}}) &\leq P\left(\sum_{i=0}^{T}\widetilde{Q}(n-T-i)>C(T+1)\right)\\ &\leq\inf_{r>0}e^{(T+1)\Lambda_{\widetilde{Q}}(r)-C(T+1)r},\end{aligned}\tag{70}$$

where $\Lambda_{\widetilde{Q}}(r) = C^{\widetilde{\gamma}}(e^r - 1)$. Then, using differentiation,

$$P_{PD}(\widetilde{\mathcal{O}}) \le e^{(T+1)(C-C^{\widetilde{\gamma}})-(T+1)(1-\widetilde{\gamma})C\log C}. \tag{71}$$

And since $P^*_{PD}(\widetilde{\mathcal{O}}) \le P_{PD}(\widetilde{\mathcal{O}})$, we get

$$\widetilde{d}_{PD}(\widetilde{\gamma}) \ge (1+T)(1-\widetilde{\gamma}). \tag{72}$$

Second, we use the lower bound $P_{PD}(\widetilde{\mathcal{O}}) \ge P(\widetilde{\mathcal{L}_D})$ to establish an upper bound on $\widetilde{d}_{PD}(\widetilde{\gamma})$.

$$\begin{aligned}P(\widetilde{\mathcal{L}_D}) &= P(\widetilde{Q}(n-T) > C(T+1)) \\ &= \sum_{k=C(T+1)+1}^{\infty} \frac{(C^{\widetilde{\gamma}})^k}{k!} e^{-C^{\widetilde{\gamma}}} \\ &\ge \frac{(C^{\widetilde{\gamma}})^{(C(T+1)+1)}}{(C(T+1)+1)!} e^{-C^{\widetilde{\gamma}}} \\ &\doteq \frac{e^{-C^{\widetilde{\gamma}}}}{\sqrt{2\pi(C(T+1)+1)}} \left(\frac{C^{\widetilde{\gamma}} e}{C(T+1)+1}\right)^{C(T+1)+1}\end{aligned}$$

And since $P^*_{PD}(\widetilde{\mathcal{O}}) \le P_{PD}(\widetilde{\mathcal{O}})$, we obtain

$$\widetilde{d}_{PD}(\widetilde{\gamma}) \le (1+T)(1-\widetilde{\gamma}). \tag{73}$$

By (72), (73), it follows that

$$\widetilde{d}_{PD}(\widetilde{\gamma}) = (1+T)(1-\widetilde{\gamma}), \quad 0 < \widetilde{\gamma} < 1.$$

Appendix D
Proof of Lemma 2

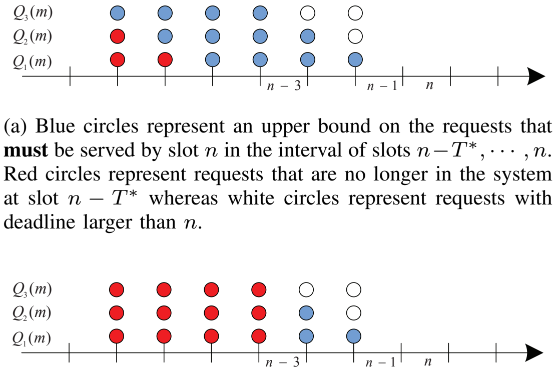
First, we show that $\mathcal{U}_R$ is a necessary condition for the outage event, that is, if an outage occurs at slot n, then $\mathcal{U}_R = \mathcal{I} \cup \mathcal{J}$ occurs. Suppose there is an outage at slot n. This outage occurs on the arrivals, $Q_k(n-k), \quad k = T_*, \cdots, T^*$, hence, $\sum_{i=0}^{T_*} N_i(n-T_*) > C(T_*+1)$, i.e., in the interval $n-T_*, \cdots, n$ the system is serving requests with deadlines not exceeding n.

Event $\mathcal{I}$ represents the case when at slot $n-T^*$, the number of requests in the system in addition to the requests that will arrive with deadlines not larger than n is larger than $C(T^*+1)$, i.e., larger than the maximum number of requests that the system can serve in the subsequent T^*+1 slots (Fig. 14a shows the requests considered in event $\mathcal{I}$ as blue circles for $T_* = 1$, $T^* = 3$.). However, event $\mathcal{I}$ alot is not a necessary condition for an outage as, for instance, we may have $Q_{T_*}(n-T_*) > C(T_*+1)$ but $\sum_{j=0}^{T^*} \sum_{i=T_*}^{T^*} Q_i(n-i-j) < C(T^*+1)$.

Now, suppose that I did not occur because of the outage at slot n, then there exists at least one slot $n-l$, $T_* < l \le T^*$ such that $\sum_{i=0}^{l} N_i(n-l) \le C$ (Otherwise, the system will be serving requests with deadline of at most n in slots $n-T^*, \cdots, n-T_*$ which implies $n \in \mathcal{I}$.). In other words, at slot l, the system will be empty of all requests that have deadlines not beyond slot n. Let

$$l^* = \min\left\{ l : \sum_{i=0}^{l} N_i(n-l) \le C, \quad T_* < l \le T^* \right\},$$

then $\sum_{j=T_*}^{l^*-1} \sum_{i=T_*}^{j} Q_i(n-j) > Cl^*$, hence J occurs. That is, all of the arriving requests in slots $n-l^*+1, \cdots, n-T_*$ with deadlines not beyond n are more than Cl^* (Fig. 14b shows the case when event $\mathcal{I}$ is not occurring while $l^* = 3$.).



(a) Blue circles represent an upper bound on the requests that **must** be served by slot n in the interval of slots $n-T^*, \cdots, n$. Red circles represent requests that are no longer in the system at slot $n-T^*$ whereas white circles represent requests with deadline larger than n.

(b) Here, event $\mathcal{I}$ is not satisfied. At slot $n-3$ the system has managed to serve all requests with deadlines not exceeding n. However, $l^* = 3$, meaning that all of the next arrivals with deadlines not exceeding n will consume the whole system capacity till slot n inclusive.

Fig. 14: An outage occurs at slot n where $T_* = 1$, $T^* = 3$. At the beginning of any time slot, arriving requests with the same deadline are represented by a circle.

Second, we show that $\mathcal{L}_R$ is a sufficient condition on the outage event. The proof is straightforward as for every k, $T_* \le k \le T^*$, the event that $\sum_{i=T_*}^{k} Q_i(n-i) > C(k+1)$ means the number of requests that must be served in the interval $n-k, \cdots, n$ is larger than $C(k+1)$ which is sufficient to cause an outage at slot n. Then, taking the union over all $k \in \{T_*, \cdots, T^*\}$ is also a sufficient condition for an outage at slot n.

Appendix E
Proof of Theorem 3

$$\begin{aligned}P_{PR}(\mathcal{O}) &\le P(\mathcal{U}_R) \\ &\le P\left(\sum_{j=0}^{T^*} \sum_{i=T_*}^{T^*} Q_i(n-j-i) > C(T^*+1)\right) \\ &\quad + \sum_{k=T_*}^{T^*-1} P\left(\sum_{j=T_*}^{k} \sum_{i=T_*}^{j} Q_i(n-j) > C(k+1)\right). \\ &\le \inf_{r_I>0} e^{\Lambda_{Q_I}(r_I) - r_I C(T^*+1)} \\ &\quad + \sum_{k=T_*}^{T^*-1} \inf_{r_k>0} e^{\Lambda_{Q_k}(r_k) - r_k C(k+1)}\end{aligned}$$

where $\Lambda_{Q_I}(r_I) = \lambda(T^*+1)(e^{r_I}-1)$ and $\Lambda_{Q_k}(r_k) = \lambda \sum_{i=0}^{k-T_*} F_{k-i}$, $T_* \le k \le T^*-1$.

For the linear scaling regime,

$$\begin{aligned}P_{PR}(\overline{\mathcal{O}}) &\le e^{(1-\overline{\gamma})C(T^*+1) + C(T^*+1)\log\overline{\gamma}} \\ &\quad + \sum_{k=T_*}^{T^*-1} e^{C(k+1)\left[1 - \frac{\overline{\gamma}\sum_{i=0}^{k-T_*} F_{k-i}}{k+1} + \log\frac{\overline{\gamma}\sum_{i=0}^{k-T_*} F_{k-i}}{k+1}\right]}.\end{aligned}$$

Let

$$\overline{v}(C) \triangleq \max_{T_* \leq k \leq T^*-1} \left\{ C(k+1) \left[1 - \frac{\overline{\gamma} \sum_{i=0}^{k-T_*} F_{k-i}}{k+1} + \log \frac{\overline{\gamma} \sum_{i=0}^{k-T_*} F_{k-i}}{k+1} \right] \right\}$$

and

$$\overline{m}(C) = \max \left\{ C(T^*+1)(1-\overline{\gamma}+\log \overline{\gamma}), \overline{v}(C) \right\},$$

then

$$\begin{aligned} \overline{d}_{PR}(\overline{\gamma}) &\geq \lim_{C\to\infty} -\frac{1}{C} \log e^{\overline{m}(C)} \\ &= \lim_{C\to\infty} -\frac{\overline{m}(C)}{C} \\ &= \min\{(T^*+1)(\overline{\gamma}-1-\log\overline{\gamma}), \overline{v}_*\} \end{aligned} \qquad (74)$$

which proves (9).

For the polynomial scaling regime,

$$P_{PR}(\widetilde{\mathcal{O}}) \leq e^{(T^*+1)(C-C^{\widetilde{\gamma}}-C\log C^{1-\widetilde{\gamma}})} + \sum_{k=T_*}^{T^*-1} e^{C(k+1)\left[1-\frac{\sum_{i=0}^{k-T_*} F_{k-i}}{C^{1-\widetilde{\gamma}}(k+1)}+\log \frac{\sum_{i=0}^{k-T_*} F_{k-i}}{C^{1-\widetilde{\gamma}}(k+1)}\right]}.$$

Let

$$\widetilde{v}(C) = \max_{T_* \leq k \leq T^*-1} \left\{ C(k+1) \left[1 - \frac{\sum_{i=0}^{k-T_*} F_{k-i}}{C^{1-\widetilde{\gamma}}(k+1)} + \log \frac{\sum_{i=0}^{k-T_*} F_{k-i}}{C^{1-\widetilde{\gamma}}(k+1)} \right] \right\} \qquad (75)$$

and

$$\widetilde{m}(C) \triangleq \lim_{C\to\infty} -\frac{\widetilde{m}(C)}{C\log C},$$

for large values of C, the terms in the $\max\{.\}$ of (75) are decreasing in k, hence

$$\widetilde{d}_{PR}(\widetilde{\gamma}) \geq (T_*+1)(1-\widetilde{\gamma}). \qquad (76)$$

Then, we use the event $\mathcal{L}_R$ with the polynomial scaling as follows.

$$\begin{aligned} P_{PR}(\widetilde{\mathcal{O}}) &\geq P(\widetilde{\mathcal{L}_R}) \\ &\geq \max_{T_* \leq k \leq T^*} \left\{ P\left(\sum_{i=T_*}^{k} \widetilde{Q}_i(n-i) > C(k+1) \right) \right\} \\ &\geq \max_{T_* \leq k \leq T^*} \frac{\left(F_k C^{\widetilde{\gamma}}\right)^{C(k+1)+1}}{C(k+1)+1!} e^{-F_k C^{\widetilde{\gamma}}} \\ &\doteq \max_{T_* \leq k \leq T^*} \left(\frac{F_k C^{\widetilde{\gamma}} e}{C(k+1)+1} \right)^{C(k+1)+1} \\ &\quad \times \frac{e^{-F_k C^{\widetilde{\gamma}}}}{\sqrt{2\pi(C(k+1)+1)}} \\ &= \left(\frac{F_k C^{\widetilde{\gamma}} e}{C(T_*+1)+1} \right)^{C(T_*+1)+1} \\ &\quad \times \frac{e^{-p_{T_*} C^{\widetilde{\gamma}}}}{\sqrt{2\pi(C(T_*+1)+1)}}. \end{aligned}$$

Hence,

$$\widetilde{d}_{PR}(\widetilde{\gamma}) \leq (T_*+1)(1-\widetilde{\gamma}). \qquad (77)$$

From (76) and (77), result (10) follows.

Appendix F
Proof of Theorem 4

Let the outage probability of the secondary user while the primary network is non-predictive be denoted by $P_N(\mathcal{O}^s)$, then

$$P_N(\mathcal{O}^s) = P(Q^p(n) + Q^s(n) > C, Q^s(n) > 0). \qquad (78)$$

Since $Q^p(n) + Q^s(n)$ and $Q^s(n)$ are two dependent random variables, we use upper and lower bounds on $P_N^s(\mathcal{O})$ to characterize $d_N^s(\gamma^p, \gamma^s)$ as follows.

$$\begin{aligned} P_N(\mathcal{O}^s) &= P(Q^p(n) + Q^s(n) > C | Q^s(n) > 0) P(Q^s(n) > 0) \\ &\overset{(a)}{\geq} P(Q^p(n) > C | Q^s(n) > 0) P(Q^s(n) > 0) \\ &\overset{(b)}{=} P(Q^p(n) > C) P(Q^s(n) > 0), \end{aligned} \qquad (79)$$

where (a) follows from the fact that $Q^s(n) \geq 0$ and (b) follows as $Q^p(n)$ and $Q^s(n)$ are independent. Moreover, since $P(\mathcal{A}, \mathcal{B}) \leq P(\mathcal{A})$, then, from (78), we can write

$$P_N(\mathcal{O}^s) \leq P(Q^p(n) + Q^s(n) > C). \qquad (80)$$

For the linear scaling regime, we have $\bar{\lambda}^p = \overline{\gamma}^p C$ and $\bar{\lambda}^s = \overline{\gamma}^s C$. From (11), (12) we obtain $0 < \overline{\gamma}^s < \overline{\gamma}^p < 1$ and $\overline{\gamma}^s + \overline{\gamma}^p < 1$. From (79),

$$\begin{aligned} P_N(\overline{\mathcal{O}}^s) &\geq P(\overline{Q}^p(n) > C) P(\overline{Q}^s(n) > 0) \\ &= P(\overline{Q}^p(n) > C) \left(1 - e^{-\overline{\gamma}^s C}\right). \end{aligned}$$

Hence

$$\begin{aligned} \overline{d}_N^s(\overline{\gamma}^p, \overline{\gamma}^s) &\leq \lim_{C\to\infty} \frac{-\log P(\overline{Q}^p(n) > 0)}{C} - \frac{\log\left(1 - e^{-\overline{\gamma}^s C}\right)}{C} \\ &\overset{(c)}{=} \overline{\gamma}^p - 1 - \log(\overline{\gamma}^p), \end{aligned} \qquad (81)$$

where (c) follows by Cramer's theorem. This proves (15). Since $\overline{Q}^p(n)$, $\overline{Q}^s(n)$ are independent Poisson random variables, then $\overline{Q}^p(n) + \overline{Q}^s(n)$ is a Poisson process with rate $(\overline{\gamma}^p + \overline{\gamma}^s)C$. Applying Cramer's theorem to (80), we obtain

$$\overline{d}_N^s(\overline{\gamma}^p, \overline{\gamma}^s) \geq (\overline{\gamma}^p + \overline{\gamma}^s) - 1 - \log(\overline{\gamma}^p + \overline{\gamma}^s)$$

which proves (16).

For the polynomial scaling regime, $\tilde{\lambda}^p = C^{\widetilde{\gamma}^p}$, $\tilde{\lambda}^s = C^{\widetilde{\gamma}^s}$. From (11), (12), we get $0 < \widetilde{\gamma}^s < \widetilde{\gamma}^p < 1$. From (80),

$$\begin{aligned} P_N(\widetilde{\mathcal{O}}^s) &\geq P(\widetilde{Q}^p(n) > C) P(\widetilde{Q}^s(n) > 0) \\ &= P(\widetilde{Q}^p(n) > C)(1 - e^{-C^{\widetilde{\gamma}^s}}) \\ &\geq \frac{C^{\widetilde{\gamma}^p (C+1)}}{(C+1)!} e^{-C^{\widetilde{\gamma}^p}} \left(1 - e^{-C^{\widetilde{\gamma}^s}}\right) \\ &\doteq \left(\frac{C^{\widetilde{\gamma}^p} e}{C+1} \right)^{C+1} \frac{e^{-C^{\widetilde{\gamma}^p}}}{\sqrt{(2\pi(C+1))}} \left(1 - e^{-C^{\widetilde{\gamma}^s}}\right). \end{aligned}$$

Hence,

$$\widetilde{d}^s_N(\widetilde{\gamma}^p, \widetilde{\gamma}^s) \leq \lim_{C \to \infty} \frac{-\log P(\widetilde{Q}^p(n) > C)}{C \log C} - \frac{\left(1 - e^{-C^{\widetilde{\gamma}^s}}\right)}{C \log C}$$
$$\leq 1 - \widetilde{\gamma}^p. \tag{82}$$

From (80), we obtain, using tightest Chernoff bound,

$$P_N(\widetilde{\mathcal{O}}^s) \leq \inf_{r>0} e^{\Lambda_{\widetilde{Q}^s + \widetilde{Q}^p}(r) - rC}, \tag{83}$$

where $\Lambda_{\widetilde{Q}^p+\widetilde{Q}^r}(r) = (C^{\widetilde{\gamma}^p} + C^{\widetilde{\gamma}^s})(e^r - 1)$. Then it follows that,

$$\begin{aligned}\widetilde{d}^s_N(\widetilde{\gamma}^p, \widetilde{\gamma}^s) &\geq 1 - \max\{\widetilde{\gamma}^p, \widetilde{\gamma}^s\} \\ &= 1 - \widetilde{\gamma}^p.\end{aligned} \tag{84}$$

From (82) and (84), the result (17) follows.

Appendix G
Proof of Theorem 5

Let the outage probability of the primary network under the dynamic scheduling policy be denoted by $P_P(\mathcal{O}^p)$. To upper bound this outage probability, it suffices to show that $f \in [0.5, 1]$ implies $P_P(\mathcal{O}^p) \leq P(\mathcal{U}_D)$, where $\mathcal{U}_D$ is as defined in Lemma 1. So, suppose that there is an outage at slot n, hence, according to the dynamic policy, $C^p(n) = C$ as $N^p_0(n) > C$. Moreover, that outage is occurring on $Q^p(n - T)$.

Now, at time slot $n - 1$, assume towards contradiction that $C^p(n - 1) < C$, then $fN_1(n - 1) < C$. This must lead to $N_0(n) \leq (1 - f)N_1(n - 1) < C$ as $1 - f \leq f$, $f \in [0.5, 1]$, which is a contradiction. Therefore, $C^p(n - 1) = C$.

Since the EDF nature of the dynamic policy implies that the network resources are only dedicated to serve primary requests that arrived prior to slot $n - T + 1$, then $C^p(n - 1)$ and $C^p(n)$ represent the served requests that arrived at slots $n - T - 1$ and $n - T$. But, $C^p(k) \leq \min\{C, f(C^p(n - 1) + C^p(n))\}$, $k = n - T, \cdots, n$. Hence, $C^p(k) = C$ for all $k = n - T, \cdots, n$ as $f \in [0.5, 1]$.

Therefore, an outage at slot n implies $\sum_{i=0}^{T} Q^p(n - i - T) > C(T + 1)$, and consequently, we obtain the lower bounds on $\overline{d}^p_P(\overline{\gamma}^p)$ and $\widetilde{d}^s_P(\widetilde{\gamma}^p)$ in the same manner as in Theorem 2.

Also, it is straightforward to see that the event $\mathcal{L}_D$ of Lemma 1 satisfies $P(\mathcal{L}_D) \leq P_P(\mathcal{O}^p)$. So the diversity gain of the polynomial scaling regime is fully determined.

Appendix H
Proof of Theorem 6

We will show the result for the linear scaling regime while its polynomial scaling regime counterpart is obtained through the same approach by taking into account the difference in the diversity gain definitions.

From (23) and (24), we can upper bound $P_P(\mathcal{O}^s)$ by

$$\begin{aligned}P_P(\mathcal{O}^s) \leq\ & P(Q^s(n) + N^p_0(n) + 0.5Q^p(n) > C, \\ & C^p(n - 1) < C) \\ & + P(Q^s(n) + N^p_0(n) + 0.5Q^p(n) > C, \\ & C^p(n - 1) = C).\end{aligned}$$

But $C^p(n-1) < C$ implies $N^p_0(n) = 0.5Q^p(n-1)$ and hence the joint event $Q^s(n) + N^p_0(n) + 0.5Q^p(n) > C, C^p(n - 1) < C$ implies $Q^s(n) + 0.5Q^p(n - 1) + 0.5Q^p(n) > C$. Therefore,

$$\begin{aligned}P_P(\mathcal{O}^s) \leq\ & P(Q^s(n) + 0.5Q^p(n - 1) + 0.5Q^p(n) > C) \\ & + P(C^p(n - 1) = C).\end{aligned} \tag{85}$$

Now, we show that the decay rate of the second term on the right hand side of (85) with C is larger than the first. We start with the second term $P(C^p(n - 1) = C)$ which can be upper bounded by

$$\begin{aligned}P(C^p(n - 1) = C) \leq\ & P(N^p_0(\eta) + 0.5Q^p(\eta) > C, \\ & C^p(\eta - 1) < C \text{ for some } \eta \leq n - 1) \\ & + P(N^p_0(m) + 0.5Q^p(m) > C, \\ & C^p(m - 1) = C \text{ for all } m \leq n - 1) \\ \leq\ & P(0.5Q^p(\eta - 1) + 0.5Q^p(\eta) > C) \\ & + P(C^p(m) = C, \text{ for all } m \leq n - 1).\end{aligned} \tag{86}$$

Fix $0 \leq M \leq n - 1$. The last term on the right hand side of (86) satisfies

$$P(C^p(m) = C, \text{ for all } m \leq n - 1) \leq P(C^p(1) = \cdots = C^p(M) = C),$$

where

$$\begin{aligned}P(C^p(1) = \cdots = C^p(M) = C) \leq\ & \\ & P(C^p(1) = \cdots = C^p(M) = C, \text{ No outages in } 1, \cdots, M) \\ & + \sum_{l=1}^{M} P(C^p(1) = \cdots = C^p(M) = C, l \text{ outages in } 1, \cdots, M)\end{aligned}$$

implying

$$\begin{aligned}P(C^p(1) = \cdots = C^p(M) = C) \leq\ & \\ & P(C^p(1) = \cdots = C^p(M) = C, \text{ No outages in } 1, \cdots, M) \\ & + (2^M - 1)P^p_P(\mathcal{O}^p).\end{aligned}$$

Since M is constant, the term $(2^M - 1)P^p_P(\mathcal{O}^p)$ decays with the system capacity as $d^p_P(\gamma^p)$. The joint event $C^p(1) = \cdots = C^p(M) = C$ and no outage in $1, \cdots, M$ implies

$$\begin{aligned}N^p_0(M) &= N^p_0(1) - (M - 1)C + \sum_{i=1}^{M-1} Q^p(i) \\ &\leq -(M - 1)C + \sum_{i=0}^{M-1} Q^p(i).\end{aligned}$$

and hence,

$$\begin{aligned}& P(C^p(1) = \cdots = C^p(M) = C, \text{ No outage in } 1, \cdots, M) \\ \leq\ & P\left(-(M - 1)C + \sum_{i=0}^{M-1} Q^p(i) + 0.5Q^p(M) > C\right) \\ \leq\ & P\left(\sum_{i=0}^{M} Q^p(i) > MC\right) \\ \leq\ & \inf_{r>0} \left\{e^{\Lambda(r) - rMC}\right\},\end{aligned}$$

where, for the linear scaling regime,

$$\overline{\Lambda}(r) = (M + 1)\overline{\gamma}^p C(e^r - 1).$$

Hence,

$$\lim_{C\to\infty} -\frac{1}{C}\log P\Big(\overline{C}^p(1) = \cdots = \overline{C}^p(M) = C, \text{ No outage in } 1,\cdots,M\Big) \geq (M+1)\overline{\gamma}^p - M + M\log\left(\frac{M}{(M+1)\overline{\gamma}^p}\right) \quad (87)$$

with the right hand side of (87) monotonically increasing in M as long as $\frac{M}{M+1} > \overline{\gamma}^p$. Then, M can be chosen sufficiently large[3] so that

$$\lim_{C\to\infty} -\frac{1}{C}\log P\left(\overline{C}^p(m) = C \text{ for all } m \leq n-1\right) \geq \overline{d}_P^p(\overline{\gamma}^p) = 2(\overline{\gamma}^p - 1 - \log\overline{\gamma}^p).$$

Also, the first term on the right hand side of (86) can be written as

$$P(0.5Q^p(\eta-1) + 0.5Q^p(\eta) > C) = P(Q^p(\eta-1) + Q^p(\eta) > 2C) \geq P_P^p(\mathcal{O}^p),$$

where $T = 1$. Hence,

$$\lim_{C\to\infty} -\frac{\log P\left(\overline{C}^p(n-1) = C\right)}{C} \geq \overline{d}_P^p(\overline{\gamma}^p) = 2(\overline{\gamma}^p - 1 - \log\overline{\gamma}^p). \quad (88)$$

Now, comparing the two terms $P(Q^s(n) + 0.5Q^p(n-1) + 0.5Q^p(n) > C)$ in (85) and $P(0.5Q^p(\eta-1)+0.5Q^p(\eta) > C)$ in (86), we have by the stationarity of $Q^p(n), n > 0$ and the non-negativity of $Q^s(n), n > 0$,

$$P(Q^s(n) + 0.5Q^p(n-1) + 0.5Q^p(n) > C) \geq P(0.5Q^p(\eta-1) + 0.5Q^p(\eta) > C).$$

This implies that the asymptotic decay rate of $\log P_P(\mathcal{O}^s)$ with C is lower bounded by the decay rate of $P(Q^s(n) + 0.5Q^p(n-1) + 0.5Q^p(n) > C)$ with C.

Now, we can use Chernoff bound to lower bound $\overline{d}_P^s(\overline{\gamma}^p, \overline{\gamma}^s)$ as follows

$$P(\overline{Q}^s(n)+0.5\overline{Q}^p(n-1)+0.5\overline{Q}^p(n) > C) \leq \inf_{r>0}\left\{e^{\overline{\Lambda}_{tot}(r)-rC}\right\}, \quad (89)$$

where

$$\overline{\Lambda}_{tot}(r) = \overline{\gamma}^s C(e^r - 1) + 2\overline{\gamma}^p C(e^{0.5r} - 1).$$

By differentiation, the optimal value of r, denoted r^*, satisfies

$$\overline{\gamma}^s e^{r*} + \overline{\gamma}^{p0.5r^*} - 1 = 0.$$

Let $\overline{y} \triangleq e^{0.5r^*}$, we obtain

$$\overline{y} = -\frac{\overline{\gamma}^p}{2\overline{\gamma}^s} + \frac{\sqrt{4\overline{\gamma}^s + \overline{\gamma}^{p2}}}{2\overline{\gamma}^s}$$

and

$$r^* = 2\log\overline{y}.$$

Substituting with r^* in (89), taking $-\log$ of both sides, dividing by C and sending $C \to \infty$, the diversity gain of the secondary network in the linear scaling regime satisfies

$$\overline{d}_P^s(\overline{\gamma}^p, \overline{\gamma}^s) \geq -\overline{\gamma}^s(\overline{y}^2 - 1) - 2\overline{\gamma}^p(\overline{y} - 1) + 2\log(\overline{y}).$$

Appendix I
Proof of Theorem 7

$$\begin{aligned} P_N(\overline{\mathcal{O}}_m) &= P\left(S^m(n) > C\right) \\ &= P\left(\frac{S^m(n)}{\overline{\theta}C} > \frac{1}{\overline{\overline{\theta}}}\right) \\ &= P\left(\frac{\sum_{l=1}^{\overline{\theta}C} X^{[l]}}{\overline{\theta}C} > \frac{1}{\overline{\overline{\theta}}}\right). \end{aligned}$$

Applying Cramer's Theorem [12],

$$\overline{d}_N(\overline{\gamma^m}, \overline{\theta}) = -\inf_{r>0}\{\overline{\theta}\Lambda_{X^{[l]}}(r) - \overline{\theta}r\}, \quad (90)$$

but

$$\begin{aligned} \Lambda_{X^{[l]}}(r) &= \log(1 - A^m + A^m e^r) \\ &= \log\left(e^{-\frac{\overline{\gamma^m}}{\overline{\theta}}} + \left(1 - e^{-\frac{\overline{\gamma^m}}{\overline{\theta}}}\right)e^r\right), \end{aligned}$$

Then,

$$r^* = \log\left(\frac{e^{-\frac{\overline{\gamma^m}}{\overline{\theta}}}}{(\overline{\theta}-1)\left(1 - e^{-\frac{\overline{\gamma^m}}{\overline{\theta}}}\right)}\right).$$

The conditions $0 < \overline{\gamma} < 1$, $\overline{\theta} > 1$ ensure that $r^* > 0$. Substituting with r^* in (90), we obtain (46).

Appendix J
Proof of Theorem 8

Under EDF scheduling, an outage occurs in slot $n \gg 1$ if and only if $N^m(n-T) > C(T+1)$, where $N^m(n-T)$ is the number of distinct multicast data sources targeted by existing requests in the system at slot $n - T$. Hence

$$P_P^*(\overline{\mathcal{O}}_m) \leq P_P(\overline{\mathcal{O}}_m) = P(N^m(n-T) > C(T+1)).$$

Let $Z_T^m(n-T)$ be the number of distinct data sources that were requested in the window of slots $[n-2T, \cdots, n-T]$, then according to EDF,

$$N^m(n-T) \leq Z_T^m(n-T). \quad (91)$$

Therefore $P(N^m(n-T) \leq C(T+1)) \leq P(Z_T^m(n-T) > C(T+1))$.

Since each data source is requested independently of the others at each slot and from slot to another, then the probability that a data source is requested at least once in a window of $T+1$ slots, denoted ξ_T^m, is equal to

$$\begin{aligned} \xi_T^m &= 1 - (1 - A^m)^{T+1} \\ &= 1 - \exp\left(-\frac{(T+1)\overline{\gamma^m}}{\overline{\theta}}\right), \end{aligned}$$

hence

$$P(Z_T^m(n-T) = k) = \begin{cases} \binom{\overline{\theta}C}{k}\xi_T^{m\,k}(1 - \xi_T^m)^{\overline{\theta}C-k}, & k = 0,\cdots,\overline{\theta}C \\ 0, & \text{otherwise.} \end{cases}$$

[3] The system is assumed to operate in the steady state, i.e., $n \gg 1$.

Now we can upper-bound $P_P^*(\overline{\mathcal{O}}_m)$ using Chernoff bound as

$$\begin{aligned}P_P^*(\overline{\mathcal{O}}_m) &\leq P_P(\overline{\mathcal{O}}_m)\\ &\leq P(Z_T^m(n-T) > C(T+1))\\ &\leq \inf_{r>0}\{e^{\Lambda_Z(r)-rC(T+1)}\},\end{aligned}$$

where $\Lambda_Z(r) = \overline{\theta}C\log\left(1-\xi_T^m+\xi_T^m e^r\right)$. Solving for $r^* > 0$ that minimizes $e^{\Lambda_Z(r)-rC(T+1)}$, we obtain

$$r^* = \log\left(\frac{(1-\xi_T^m)(T+1)}{\xi_T^m(\overline{\theta}-(T+1))}\right).$$

Now, taking $-\log P_P^*(\overline{\gamma^m},\overline{\theta})$, dividing by C and taking the limit as $C \to \infty$, we obtain (49).

Appendix K
Proof of Theorem 9

We have by the definition of $\overline{\mathcal{O}}_A$ in Scenario 1 that

$$P(\overline{\mathcal{O}}_A) = P(S^m(n) + Q^u(n) > C).$$

By Cramer's theorem, we have

$$\overline{d}_1(\overline{\gamma^u},\overline{\gamma^m},\overline{\theta}) = \inf_{r>0}\{r - \Lambda_{m+u}(r)\}, \tag{92}$$

where

$$\Lambda_{m+u}(r) = \overline{\gamma^u}(e^r-1) + \overline{\theta}\log\left(e^{-\frac{\overline{\gamma^m}}{\overline{\theta}}} + e^r - e^{r-\frac{\overline{\gamma^m}}{\overline{\theta}}}\right).$$

Differentiating $r - \Lambda_{m+u}(r)$ with respect to r and equating with 0, we obtain

$$\overline{\gamma^u}\left(e^{\frac{\overline{\gamma^m}}{\overline{\theta}}}-1\right)e^{2r^*}+\left((\overline{\theta}-1)e^{\frac{\overline{\gamma^m}}{\overline{\theta}}}-\overline{\theta}+\overline{\gamma^u}+1\right)e^{r^*}-1=0. \tag{93}$$

Set $y_1 = e^{r^*}$, then (93) is a quadratic equation in y_1, that can be solve analytically for two possible roots. Choosing the root $y_1 > 1$ for $r^* > 0$, we get

$$\begin{aligned}y_1 =&\frac{1}{2\overline{\gamma^u}\left(e^{\frac{\overline{\gamma^m}}{\overline{\theta}}}-1\right)}\Bigg[\Big((\overline{\theta}^2-2\overline{\theta}+1)e^{\frac{2\overline{\gamma^m}}{\overline{\theta}}}\\ &+\left(-2\overline{\theta}^2+2\overline{\theta}(\overline{\gamma^u}+2)+2(\overline{\gamma^u}-2)\right)e^{\frac{\overline{\gamma^m}}{\overline{\theta}}}+\overline{\theta}^2\\ &-2\overline{\theta}(\overline{\gamma^u}+1)+\overline{\gamma^u}^2-2\overline{\gamma^u}+1\Big)^{\frac{1}{2}}+(1-\overline{\theta})e^{\frac{\overline{\gamma^m}}{\overline{\theta}}}\\ &+\overline{\theta}-\overline{\gamma^u}-1\Bigg].\end{aligned}$$

Substitution with $y_1 = e^{r*}$ into (92), we obtain (51).

Appendix L
Proof of Theorem 10

Under the policy π_2, suppose that an outage event has occurred in slot $n \gg 1$, then $N_0^m(n) + Q^u(n) > C$, which can be decomposed to either of the following to events: 1) $Q^u(n) > C$ or 2) $Q^u(n) \leq C$ but $N_0^m(n) > 0$ so that $N_0^m(n) + Q^u(n) > C$. Now, focus on the second event, specifically, $N_0^m(n) > 0$. To each data source of the $N_0^m(n)$, at least one corresponding request has already arrived at slot $n-T$. Since $N_0^m(n) > 0$ and $N_0^m(n)+Q^u(n) > C$, then the system is operating at full capacity in the slots $[n-T,\cdots,n]$. That is,

$$N^m(n-T) + \sum_{i=0}^{T} Q^u(n-i) > C(T+1),$$

where $N^m(n-T)$ is the number of distinct multicast data sources demanded by at least one request existing in the system at slot $n-T$.

From (91), $N^m(n-T) \leq Z_T^m(n-T)$, where $Z_T^m(n-T)$ is as defined in Appendix J, then we can now write

$$\begin{aligned}P_2(\overline{\mathcal{O}}_A) \leq\ & P(Q^u(n) > C)\\ &+ P\Bigg(\sum_{i=0}^{T} Q^u(n-i) + Z_T^m(n-T) > C(T+1),\\ &Q^u(n) < C\Bigg)\\ \leq\ & P(Q^u(n) > C)\\ &+ P\left(\sum_{i=0}^{T} Q^u(n-i) + Z_T^m(n-T) > C(T+1)\right).\end{aligned}$$

We have from Theorem 1 that

$$\lim_{C\to\infty} -\frac{\log P(Q^u(n) > C)}{C} = \overline{\gamma^u} - 1 - \log\overline{\gamma^u}. \tag{94}$$

Also, Cramer's theorem can be used in the same way of Appendix K to show that

$$\begin{aligned}&\lim_{C\to\infty} -\frac{1}{C}\log P\Bigg(\sum_{i=0}^{T} Q^u(n-i) + Z_T^m(n-T)\\ &\qquad > C(T+1)\Bigg)\\ &= (T+1)\log y_2 - (T+1)\overline{\gamma^u}(y_2-1)\\ &\quad - \overline{\theta}\log(1-\xi_T^m+\xi_T^m y_2),\end{aligned} \tag{95}$$

where

$$\begin{aligned}y_2 =&\frac{1}{2\xi_T^m\overline{\gamma^u}(T+1)}\Bigg[\Bigg(\Big((1-\xi_T^m)^2\overline{\gamma^u}^2+2\xi_T^m\overline{\gamma^u}(1-\xi_T^m)\\ &+\xi_T^{m2}\Big)^2T^2+\Big([2\xi_T^m\overline{\gamma^u}(1-\xi_T^m)-2\xi_T^{m2}]\overline{\theta}\\ &+2\xi_T^{m2}(1-\xi_T^m)^2+4\xi_T^m\overline{\gamma^u}(1-\xi_T^m)+2\xi_T^{m2}\Big)T\\ &+[2\xi_T^m\overline{\theta}(1-\xi_T^m)-2\xi_T^{m2}]\overline{\theta}+\overline{\gamma^u}^2(1-\xi_T^m)^2\\ &+2\xi_T^m\overline{\theta}(1-\xi_T^m)+\xi_T^{m2}(1+\overline{\theta})^2\Bigg)^{\frac{1}{2}}\\ &+\Big((\xi_T^m-1)\overline{\gamma^u}\Big)T-\xi_T^m\overline{\theta}+\overline{\gamma^u}(\xi_T^m-1)+\xi_T^m\Bigg].\end{aligned}$$

Therefore, from (94) and (95), (52) follows.

To see (53), it suffices to note that $Q^u(n) > C$ is a sufficient condition for an outage at slot n independently of the service policy used. Hence, $P_2(\overline{\mathcal{O}}_A) \geq P(Q^u(n) > C)$, therefore, $\overline{d}_2(\overline{\gamma^u},\overline{\gamma^m},\overline{\theta}) \leq \overline{d}_N(\overline{\gamma^u})$.

Appendix M
Proof of Theorem 11

An outage event at slot n implies $N^u(n-T)+N^m(n-T) > C(T+1)$ where $N^u(n-T)$ is the number of unicast requests existing in the network at time slot $n-T$. Hence

$$P_3(\overline{\mathcal{O}}_A) \leq P(N^u(n-T)+N^m(n-T) > C(T+1)),$$

but

$$N^u(n-T) \leq \sum_{i=0}^{T} Q^u(n-i-T),$$

and

$$N^m(n-T) \leq Z_T^m(n-T).$$

Therefore

$$P_3(\overline{\mathcal{O}}_A) \leq P\left(\sum_{i=0}^{T} Q^u(n-i-T)+Z_T^m(n-T) > C(T+1)\right).$$

Since $\{Q^u(i)\}_i$ are IID random variables, then from (95), we obtain (54).

Appendix N
Proof of Theorem 12

Regardless of the scheduling policy used, the following event is sufficient for an outage at slot n.

$$Q^u(n-i-T) > 2C - S^m(n-2i) - S^m(n-2i+1), \qquad i = 1, \cdots, T,$$

and

$$Q^u(n-T) > C - S^m(n).$$

The above event ensures that the number of delayed unicast requests is increasing over the window of slots $[n-2T, \cdots, n-T]$ where at slot $n-T$, the network will end up having

$$\sum_{i=0}^{T} Q^u(n-i-T) + S^m(n-i) > C(T+1),$$

implying that the total number of resources that have to be consumed by slot n inclusive is greater than the aggregate available capacity $C(T+1)$ which would cause an outage.

Noting that $\{S^m(i)\}_i$ are IID, we can write

$$\begin{aligned} P_4^*(\overline{\mathcal{O}}_A) \geq & P(Q^u(n-T)+S^m(n) > C) \\ & \times P\Big(Q^u(n-T+1)+S^m(n-2) \\ & + S^m(n-1) > 2C\Big)^T, \end{aligned}$$

which, using Chernoff bound, leads to (55).

References

[1] FCC. Spectrum policy task force report, FCC 02-155. Nov. 2002.

[2] J. Mitola III,"Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio" Doctor of Technology Dissertation, Royal Institute of Technology (KTH), Sweden, May, 2000

[3] I. Akyildiz, W. Lee, M. Vuran, and S. Mohanty, "NeXt generation/dynamic spectrum access/cognitive radio wireless networks: A survey," *Computer Networks Journal (Elsevier)*, September 2006.

[4] S. A. Jafar, S. Srinivasa, I. Maric, and A. Goldsmith, "Breaking spectrum gridlock with cognitive radios: an information theoretic perspective", *Proceedings of the IEEE*, May 2009.

[5] C. Song, Z. Qu, N. Blumm, and A. Barabas, "Limits of Predictability in Human Mobility", *Science*, vol. 327, pp. 1018-1021, Feb. 2010.

[6] J. Lee, and N. Jindal, "Asymptotically optimal policies for hard-deadline scheduling over fading channels", Submitted to *IEEE Transactions on Information Theory*, June 2009.

[7] S. Kittipiyakul, P. Elia, and T. Javidi, "High-SNR analysis of outage-limited communications with bursty and delay-limited information", *IEEE Transactions on Information Theory*, vol.55, no.2, pp.746-763, Feb. 2009.

[8] P. Bhattacharya, A. and Ephremides, "Optimal scheduling with strict deadlines", *IEEE Transactions on Automatic Control*, vol.34, no.7, pp.721-728, Jul. 1989.

[9] D. Pandelis, and D. Teneketzis, "Stochastic scheduling in priority queues with strict deadlines," *Probability in the Engineering and Informational Sciences*, pp. 273-289, 1993.

[10] R. G. Gallager, "Discrete Stochastic Processes", Kluwer, Boston, 1996.

[11] Peter W. Glynn, "Upper bounds on Poisson tail probabilities", *Operations Research Letters*, Vol. 6, pp. 9-14, March 1987.

[12] A. Ganesh, N. O'Connell and D. Wischik, "Big queues", Lecture Notes in Mathematics, vol. 1838, 2004.

[13] S. S. Panwar, D. Towsley and J. K. Wolf, "Optimal scheduling policies for a class of queues with customer deadlines to the beginning of service", *ACM SIGMETRICS Performance Evaluation Review*, vol. 18, no.e 3, Nov. 1990.

[14] M. Kargahi, and A. Movaghar, "Non-preemptive earliest-deadline-first scheduling policy: a Performance study," *Proceedings of The 13th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS '05)*.

[15] F. G. Foster, On the stochastic matrices associated with certain queueing processes, *Ann. Math Statist.*, vol. 24, pp. 355-360, 1953.

[16] H. Won, H. Cai, D. Y. Eun, K. Guo, A. Netraveli, I. Rhee, and K. Sabnani, "Multicast scheduling in cellular data networks," *IEEE Transactions on Wireless Communications*, vol.8, no.9, pp.4540-4549, Sept. 2009.

[17] J. Tadrous, A. Eryilmaz, and Hesham El Gamal, "Proactive Multicasting with Predicable Demands," *IEEE Internation Symposium on Information Theory (ISIT) 2011*, vol., no., pp.239-243, Jul. 2011.