Proactive Resource Allocation: Turning Predictable Behavior into Spectral Gain

Hesham El Gamal
Department of Electrical
and Computer Engineering
Ohio State University, Columbus, USA
helgamal@ece.osu.edu

John Tadrous
Wireless Intelligent Networks
Center (WINC)
Nile University, Cairo, Egypt
john.tadrous@nileu.edu.eg

Atilla Eryilmaz
Department of Electrical
and Computer Engineering
Ohio State University, Columbus, USA
eryilmaz@ece.osu.edu

***Abstract*—This paper introduces the novel concept of proactive resource allocation in which the predictability of user behavior is exploited to balance the wireless traffic over time, and hence, significantly reduce the bandwidth required to serve the wireless network at a given blocking/outage probability. We start with a simple model in which 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 under different arrival processes. This model is then generalized to incorporate the effect of prediction errors and the randomness in the prediction lookahead time T. Remarkably, we also show that, in the cognitive networking paradigm, the appropriate use of proactive resource allocation by the primary users results in more spectral opportunities for the secondary users at a marginal, or no, cost in the primary network outage.**

I. A New Paradigm for Resource Allocation

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 poses a challenge of designing 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.

1) While we are experiencing a severe shortage in the spectrum, it was well documented now that a significant fraction of the available spectrum is under-utilized [1]. This, in fact, is the main motivation for the cognitive networking paradigm where secondary users are allowed to use the spectrum in the off time, where the primary users are idle, in an attempt to maximize the spectrum utility [2]. Unfortunately, the cognitive radio approach is still facing significant regulatory and technological hurdles [3], [4] and, at best, will offer only a partial solution to the problem. This limitation of the cognitive radio approach is intimately 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. As an example, if we take a typical cellular network, one can easily see that the traffic demand in the peak hours is much higher than that at night time; which inspires the different rates offered by cellular operator at night and day times. Now, the cognitive radio approach assumes that the secondary users will be able to utilize the spectrum in the off peak times but, unfortunately, at those particular 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). As argued in the following, 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.
2) In the traditional approach, wireless networks were constructed assuming that the subscribers are equipped with *dumb terminals* with very limited computational power. It is obvious that the new generation of *smart phones* enjoy significantly enhanced capabilities in terms of both **processing power and available memory**, as compared with the older generation wireless devices. Moreover, according to Moore's law predictions, one should expect the computational and memory resources available at the typical wireless device to increase at an exponential rate. 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 first step in this direction.
3) The introduction of smart phones, most notably the

iPhone, has resulted in a paradigm shift in the dominant traffic in mobile cellular networks. Whereas the primary traffic source in the traditional paradigm 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.

4) The final piece of our puzzle relates to the observation that the usage of the wireless devices is **highly predictable**. For example, a particular user favorite source for the daily news 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 **safely predict** that the user will be interested in the CNN again in the following day. Coupled with the fact that the most websites are refreshed at a relatively slow rate, as compared with the dynamics of the underlying wireless network, 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.

The objective of this paper is to highlight the potential improvement in the spectral efficiency of wireless networks through the judicious exploitation of the predictable behavior of wireless users. More specifically, in the current paradigm, traffic requests are considered urgent, at the time scale of the application layer, and hence, have to be served upon initiation by the network users in order to satisfy the required QoS metrics. However, if the wireless devices can **predict** the requests to be generated by the corresponding users and submit them in advance, then the network will have the flexibility in scheduling these requests over an expanded time horizon as long as the imposed deadlines are not violated. When a **predictive** network serves a request before its deadline, the corresponding data is stored in cache memory of the wireless device and, when the request is actually initiated, the application pulls the information directly from the memory instead of accessing the wireless network. It is worth noting that, not all applications, although predictable, can be served prior to their time of initiation. For example, some multimedia traffic maybe predictable, but, can only be served on a real time basis as they are based on live interactions between users. However, predicting these type of requests can still be considered as an advantage, as the network may schedule other non-real-time requests while taking into account the predicted real-time requests in a way that enhances the QoS of all applications.

The rest of this paper is devoted to developing quantitative evidence that supports the previous qualitative discussion via analyzing certain asymptotic scenarios. More specifically, Section II describes a simplified system that will be the basis of our analytical results. The notion of **prediction diversity** is introduced in Section III and quantified under different assumption on the performance of the prediction algorithm. Our analysis is extended to the scenario where users require different QoS guarantees, e.g., primary and secondary users in Section IV. Here, we demonstrate a remarkable phenomenon whereby prediction at one user, i.e., *good citizen*, is shown to improve the performance of the other without compromising its own. Throughout the paper, our theoretical claims are supported by numerical results that clearly illustrate the potentially remarkable gains in spectral efficiency that can be achieved by our proactive resource allocation approach. Finally, the paper is concluded in Section V.

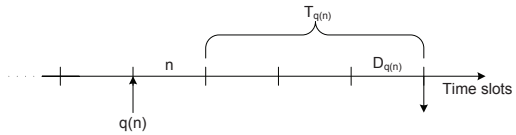

Fig. 1. Prediction Model

II. System Model

Unless otherwise stated, we adopt a simplified model where that time is divided into slots and the requests are allowed to arrive only at the beginning of each slot. The number of arriving requests at time slot $n > 0$ is denoted by $Q(n)$ which is assumed to be ergodic and to follow a Poisson distribution with rate λ. All requests are assumed to have the same amount of required resources which is taken to be unity. That is, each request has to be totally served in a single slot via a one unit of resources. Moreover, the wireless network has a fixed capacity C (total resources) per slot, i.e., the number of served requests per slot cannot exceed C.

Furthermore, we assume that a predictive wireless network can expect the arrival of each request by an integer number of time slots in advance. That is, if $q(n)$, $1 \leq q \leq Q(n)$, is the ID of a request predicted at the beginning of time slot n, the predictive network has the capability of serving this request no later than the next $T_{q(n)}$ slots. Hence, when a 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. In the *non-predictive* network, all arriving requests at the beginning of time slot n have to be served in the same time slot n, i.e., if $q(n)$ is a non-predicted request, its deadline is $D_{q(n)} = n$ meaning that $T_{q(n)} = 0$. Finally, we assume that an outage event occurs at a certain time slot if and only if at least one of the requests in the system expires in this slot.

III. Prediction Diversity

In this section we characterize the outage probability of both non-predictive and predictive networks, while assuming that both networks have the same arrival process $Q(n)$, $n > 0$ per slot. The difference only is in the deadlines of the arriving requests. The deadline for a request $q(n)$ is slot n when the network is non-predictive, and is $n + T_{q(n)}$ when the network is predictive with $T_{q(n)} = 1, 2, \cdots$. In general, as the

system capacity C grows, the outage probability is expected to decrease. In our analysis we assume that the arrival rate of requests scales with C as C^γ where $0 \leq \gamma \leq 1$ and then we use tools of large deviations [5], [6] to characterize the prediction diversity gain as the rate of decay of the outage probability with $C \log C$. So, if $P(\text{outage})$ is the outage probability, then the **prediction diversity gain** is defined as

$$d(\gamma) \triangleq \lim_{C\to\infty} \frac{-\log P(\text{outage})}{C \log C}.$$

In our formulation γ can be viewed as the multiplexing gain of the system and $d(\gamma)$ quantifies the diversity-multiplexing tradeoff achieved in each scenario.

A. The Non-predictive Network Benchmark

In the non-predictive network, all of the arriving requests at the beginning of time slot n, $Q(n)$, must be served in slot n, otherwise, slot n goes in outage. Hence, the probability that slot n suffers an outage is given by

$$Pr(Q(n) > C).$$

Since $Q(n)$, $n > 0$, is assumed to be ergodic, then the outage probability is independent of n. Consequently, the outage probability of the non-predictive system will be

$$P_N(\text{outage}) = \sum_{k=C+1}^{\infty} \frac{(C^\gamma)^k}{k!} e^{-C^\gamma}.$$

For large values of C, the above outage probability can be written using Stirling's approximation as follows,

$$P_N(\text{outage}) \approx \sum_{k=C+1}^{\infty} \frac{1}{\sqrt{2\pi k}} \frac{(C^\gamma e)^k}{k^k} e^{-C^\gamma}. \quad (1)$$

The denominator of the kth term in the above summation scales as k^k, hence, in the asymptotic scenario where $C \to \infty$ the dominant term in (1) is at $k = C + 1$. Thus,

$$P_N(\text{outage}) \doteq \frac{1}{\sqrt{2\pi(C+1)}} \frac{(C^\gamma e)^{C+1}}{(C+1)^{(C+1)}} e^{-C^\gamma}, \quad (2)$$

where $\doteq$ means $C \to \infty$. The diversity gain of the non-predictive system, $d_N(\gamma)$, can then be obtained as

$$\begin{aligned} d_N(\gamma) &= \lim_{C\to\infty} -\frac{\log P_N(\text{outage})}{C \log C} \\ &= \lim_{C\to\infty} \frac{1}{2C \log C} \log\left(2\pi(C+1)\right) - \gamma.\frac{C+1}{C} \\ &\quad - \frac{C+1}{C \log C} + \frac{C+1}{C} \frac{\log(C+1)}{\log C} + \frac{C^\gamma}{C \log C} \\ &= 1 - \gamma. \end{aligned} \quad (3)$$

B. Predictive Network

In order to minimize the outage probability, we consider the following service policy:

Definition 1: Service Policy 1 (SP1):
Let $N(n)$ be the number of requests in the system at the beginning of time slot n (for the non-predictive network $N(n) = Q(n)$ but for the predictive network $N(n) \geq Q(n)$ as there may exist some unserved requests from the previous slots whose deadlines are not passed yet), then at time slot n the network scheduler sorts the $N(n)$ requests in an ascending order with respect to their deadlines then starts serving them in order from the request with the smallest deadline till either all the $N(n)$ requests are served or some requests are delayed to the next slot if $N(n) > C$ and the deadlines of the remaining requests are beyond n.

To characterize the outage probability and the corresponding prediction diversity gain, we start with a special case where all of the arriving requests are all predicted by the same amount of time slots in advance. That is, for all $n > 0$ and $0 \leq q(n) \leq Q(n)$, $T_{q(n)} = T$ where T is deterministic. Obtaining a closed form expression for the outage probability in this case, $P_P(\text{outage})$, is quite complicated, so we use upper and lower bounds on the outage probability which are shown to result in **a sharp characterization** of the prediction diversity gain.

1) Upper Bound: We define the event $\mathcal{U}_d(n)$ as the event that the number of arrivals in the slots $n - 2T, n - 2T + 1, \cdots, n - T$ is strictly greater than $C(T + 1)$ in the steady state, i.e.,

$$\mathcal{U}_d(n) \triangleq \left\{ \sum_{i=n-2T}^{n-T} Q(i) > C(T+1) \right\}.$$

Lemma 2: If an outage event occurs at time slot n, then $\sum_{i=n-2T}^{n-T} Q(i) > C(T + 1)$, and the converse is not necessarily true . i.e., $P_P(\text{outage}) \leq Pr(\mathcal{U}_d(n))$.

Proof: Since any request $q(m)$ does not exist in the system at any slots beyond $m + T$, hence, the outage at slot n is on the requests arriving at the beginning of slot $n - T$.

Since there is an outage at time slot n, in the interval of time slots $n-T, n-T+1, \cdots, n$, the system is only serving the arrivals of time slots $n - 2T, n - 2T + 1, \cdots, n - T$.

The maximum number of requests that the system can serve in an interval of $T + 1$ slots is $C(T + 1)$. Therefore, to have an outage at slot n, $\sum_{i=n-2T}^{n-T} Q(i) > C(T + 1)$.

Since the arrivals per slot form a stationary process, then $Pr(\mathcal{U}_d(n))$ is independent of the slot n, and for simplicity, we use $Pr(\mathcal{U}_d) = Pr(\mathcal{U}_d(n))$.

To show that $\sum_{i=n-2T}^{n-T} Q(i) > C(T + 1)$ is not sufficient for an outage event at slot n, consider the following counter example, $Q(n - 2T) > C(T + 1)$ but $Q(n - 2T + 1)$, $Q(n - 2T + 2)$, $\cdots$, $Q(n - T) < C$. In this case, $\sum_{i=n-2T}^{n-T} Q(i) > C(T + 1)$ but the arriving requests at the beginning of slot $n - 2T$ will go in outage at slot $n - T$, then no more outages will occur till slot n inclusive. Hence, $P_P(\text{outage}) \leq Pr(\mathcal{U}_d)$.

Note that, the above result is valid for any i.i.d. arrival process not necessarily Poisson distributed. ■

Let $d_{\mathcal{U}_d}(\gamma) \triangleq \lim_{C\to\infty} -\frac{\log Pr(\mathcal{U}_d)}{C \log C}$. Since $P_P(\text{outage}) \leq Pr(\mathcal{U}_d)$, then, the prediction diversity gain of the predictive system, $d_P(\gamma)$, satisfies $d_P(\gamma) \geq d_{\mathcal{U}_d}(\gamma)$. The number of arrivals in a $T+1$-slots period is Poisson with rate $C^\gamma(T+1)$.

Hence,

$$Pr(\mathcal{U}_d) = \sum_{k=C(T+1)+1}^{\infty} \frac{(C^\gamma(T+1))^k}{k!} e^{C^\gamma(T+1)} \doteq \frac{1}{\sqrt{2\pi(C(T+1)+1)}} \left(\frac{C^\gamma(T+1)e}{C(T+1)+1}\right)^{C(T+1)+1} \times e^{-C^\gamma(T+1)}. \quad (4)$$

From the above result, $d_{\mathcal{U}_d}(\gamma)$ is obtained as

$$d_{\mathcal{U}_d}(\gamma) = (1+T)(1-\gamma). \quad (5)$$

Therefore,

$$d_P(\gamma) \geq (1+T)(1-\gamma). \quad (6)$$

2) Lower Bound: We define the event $\mathcal{L}_d(n)$ as the event that the number of arrivals at the beginning of slot $n-T$ is strictly larger than $C(T+1)$ in the steady state, that is,

$$\mathcal{L}_d(n) \triangleq \{Q(n-T) > C(T+1)\}.$$

Lemma 3: If $Q(n-T) > C(T+1)$, then an outage occurs at slot n, and the converse is not necessarily true, i.e., $Pr(\mathcal{L}_d(n)) \leq P(\text{outage})$.

Proof: Since the arriving requests at the beginning of slot $n-T$ have a deadline of n, then they may exist in the system only in the interval $n-T, \cdots, n$. Moreover, for the causal system, the arrivals of time slot $n-T$ can not start receiving service at any slot prior to $n-T$. Hence, if the system starts serving those requests along the interval $n-T, \cdots, n$, the maximum number of requests that it can serve is $C(T+1)$. But, since $Q(n-T) > C(T+1)$, an outage will occur at slot n.

Since the arrival process $Q(i)$, $i > 0$ is stationary, then as $n \to \infty$, $Pr(\mathcal{L}_d(n))$ is independent of n, hence, we use $Pr(\mathcal{L}_d(n)) = Pr(\mathcal{L}_d)$.

To show that $Q(n-T) > C(T+1)$ is not a necessary condition for an outage at time n, consider the following scenario, $Q(n-T-1) > C(T+1)$ but $C < Q(n-T) \leq C(T+1)$. From the definition of SP1, the requests arriving at the beginning of slot $n-T$ will not start receiving service before the system completes the service of requests arriving at the beginning of slot $n-T-1$. Since, $Q(n-T-1) > C(T+1)$, the system will incur an outage at slot $n-1$. Then at slot n (the deadline of $Q(n-T)$ requests) the system cannot serve all of the $Q(n-T)$ as $Q(n-T) > C$, and will go in outage although $Q(n-T) \leq C(T+1)$. Consequently, $Pr(\mathcal{L}_d) \leq P_P(\text{outage})$. ■

Let $d_{\mathcal{L}_d}(\gamma) \triangleq \lim_{C\to\infty} -\frac{\log Pr(\mathcal{L}_d)}{C \log C}$. Since, $Pr(\mathcal{L}_d) \leq P_P(\text{outage})$, then $d_P(\gamma) \leq d_{\mathcal{L}_d}(\gamma)$. To characterize $d_{\mathcal{L}_d}(\gamma)$,

$$Pr(\mathcal{L}_d) = \sum_{k=C(T+1)+1}^{\infty} \frac{(C^\gamma)^k}{k!} e^{-C^\gamma} \doteq \frac{1}{\sqrt{2\pi(C(T+1)+1)}} \left(\frac{C^\gamma e}{C(T+1)+1}\right)^{C(T+1)+1} \times e^{-C^\gamma}. \quad (7)$$

Hence,

$$d_{\mathcal{L}_d}(\gamma) = (1+T)(1-\gamma). \quad (8)$$

Consequently,

$$d_P(\gamma) \leq (1+T)(1-\gamma). \quad (9)$$

From (6), (9), the diversity of the predictive system is given by

$$d_P(\gamma) = (1+T)(1-\gamma). \quad (10)$$

Comparing $d_N(\gamma)$ and $d_P(\gamma)$, it is obvious that the predictive system with fixed T operating according to SP1 **enhances the diversity by a factor of $1+T$, i.e., a prediction diversity gain of $1+T$.**

C. Random T

Now, we consider a more general case where $T_{q(n)}, 0 \leq q \leq Q(n), n > 0$ is a sequence of i.i.d. nonnegative integer-valued random variables defined over a finite support $T_{min}, T_{min}+1, \cdots, T_{max}$. First, we start with the scenario where probability mass function (PMF) of $T_{q(n)}$ does not scale with C and establish the critical dependence of the achievable diversity gain on T_{min}. Then, we show that even when $T_{min} = 0$, one can achieve significant prediction diversity gain under certain assumption on the scalability of the PMF with C. More specifically, let the PMF of $T_{q(n)}$ be given by

$$Pr(T_{q(n)} = k) \triangleq \begin{cases} p_k, & T_{min} \leq k \leq T_{max}, \\ 0, & \text{otherwise}, \end{cases} \quad (11)$$

where $T_{min} \geq 0$ and $T_{max} \geq T_{min}$. Note that, the scenario $T_{min} = T_{max} = 0$ corresponds to the non-predictive system, and $T_{min} = T_{max} = T > 0$ corresponds to the predictive system with T deterministic. First, we assume that the p_k's are constants that do not depend on C. We refer to the cumulative distribution function (CDF) of the random variable $T_{q(n)}$ as

$$F(k) = F_k = \begin{cases} 1, & k > T_{max}, \\ \sum_{i=T_{min}}^{k} p_i, & T_{min} \leq k \leq T_{max}, \\ 0, & k < T_{min}. \end{cases} \quad (12)$$

Similar to the deterministic T scenario, we use an argument based on tight upper and lower bounds on the outage probability to characterize the diversity of the random T case.

1) Upper Bound: To find an upper bound on the outage probability, it is required to have an event whose probability of occurrence is at least equal to the probability of an outage occurrence. Hence, any necessary but not sufficient condition on the outage event has probability of occurrence that acts as an upper bound on the actual outage probability. However, for the tightness on of the upper bound, we use the event

$$\mathcal{U}_r(n) \triangleq \left\{ \bigcup_{T_{min} \leq k \leq T_{max}} \left(\sum_{j=0}^{k} \sum_{i=T_{min}}^{k} Q_i(n-j-i) > C(k+1) \right) \right\},$$

where $Q_k(n)$ is the number of requests with $T = k$ arriving at the beginning of time slot n, as a necessary but not sufficient

condition for the outage event occurring at time n in the steady state.

Lemma 4: If an outage event occurs at time slot n, then the event $\mathcal{U}_r(n)$ is also occurring at the same slot, and the converse is not necessarily true, i.e., $P_P(\text{outage}) \leq Pr(\mathcal{U}_r(n))$.

Proof: Since there is an outage at slot n, this outage is on the requests whose deadline is slot n. Those requests start existing in the system from the beginning of time slot $n-T_{max}$ till the beginning of time slot $n-T_{min}$. Given an outage at slot n, either of the following events must have occurred.

First, the number of arrivals with $T = T_{min}$ at the beginning of time slot $n-T_{min}$, i.e., $Q_{T_{min}}(n-T_{min})$, plus all of the arrivals with $T = T_{min}$ in the previous T_{min} slots are strictly larger than $C(T_{min}+1)$. That is, $\sum_{j=0}^{T_{min}} Q_{T_{min}}(n-T_{min}-j) > C(T_{min}+1)$. This event is following directly from the proof of Lemma 2. Or, second, the number of the requests arriving at the beginning of time slot $n-T_{min}-1$ with $T = T_{min}+1$, i.e., $Q_{T_{min}+1}(n-T_{min}-1)$ plus all of the requests whose $T = T_{min}+1$ arriving in the previous $T_{min}+1$ slots plus all of the requests whose $T = T_{min}$ arriving in slots $n-2T_{min}-1, n-2T_{min}, \cdots, n-T_{min}$ are strictly larger than $C(T_{min}+2)$. That is, $\sum_{j=0}^{T_{min}+1}\sum_{i=T_{min}}^{T_{min}+1} Q_i(n-j-k) > C(T_{min}+2)$, and so on till the event that $\sum_{j=0}^{T_{max}}\sum_{i=T_{min}}^{T_{max}} Q_i(n-j-i) > C(T_{max}+1)$. Consequently, the union of all of those events is a consequence of having an outage at slot n, meaning that $\mathcal{U}_r(n)$ is a necessary condition for an outage at slot n. Since the arrival process is stationary, then at steady state as $n \to \infty$, $Pr(\mathcal{U}_r(n))$ is independent of n, hence we use $Pr(\mathcal{U}_r) = Pr(\mathcal{U}_r(n))$.

To show that the event $\mathcal{U}_r(n)$ is not sufficient for an outage at slot n, consider the following counter example, $Q_{T_{max}}(n-2T_{max}) > C(T_{max}+1)$ and $Q_i(n-i-j) \leq C/(T_{max}-T_{min}+1), T_{min} \leq j, i \leq T_{max}$, $j \geq 0$ and $j+i \neq 2T_{max}$. In this case, one of the sufficient events for $\mathcal{U}_r$ is satisfied, that is $\sum_{j=0}^{T_{max}}\sum_{i=T_{min}}^{T_{max}} Q_i(n-j-i) > C(T_{max}+1)$. However, in this case, there is no outage at slot n as the last outage before slot n must have occurred at slot $n-T_{max}$, and then the system will proceed with no outages till slot n at least. Consequently, $P_P(\text{outage}) \leq \Pr(\mathcal{U}_r)$. ∎

Let $d_{\mathcal{U}_r}(\gamma) \triangleq \lim_{C\to\infty} -\frac{\log(Pr(\mathcal{U}_r))}{C\log(C)}$. Since $P_P(\text{outage}) \leq \Pr(\mathcal{U}_r)$, hence, $d_P(\gamma) \geq d_{\mathcal{U}_r}(\gamma)$.

$$Pr(\mathcal{U}_r) = Pr\left(\bigcup_{T_{min}\leq k\leq T_{max}}\left(\sum_{j=0}^{k}\sum_{i=T_{min}}^{k} Q_i(n-j-i) > C(k+1)\right)\right) \tag{13}$$

$$\leq \sum_{k=T_{min}}^{T_{max}} Pr\left(\sum_{j=0}^{k}\sum_{i=T_{min}}^{k} Q_i(n-j-i) > C(k+1)\right) \tag{14}$$

$$= \sum_{k=T_{min}}^{T_{max}}\sum_{l=C(k+1)+1}^{\infty} \frac{((k+1)F_kC^\gamma)^l}{l!}e^{-(k+1)F_kC^\gamma} \tag{15}$$

$$\doteq \sum_{k=T_{min}}^{T_{max}}\sum_{l=C(k+1)+1}^{\infty} \frac{(e/l)^l\left((k+1)F_kC^\gamma\right)^l}{\sqrt{2\pi l}} \times e^{-(k+1)F_kC^\gamma}. \tag{16}$$

Since F_k is monotonically nondecreasing of k, and for large values of C, the dominant exponent in the above summation will be the one for $k = T_{min}$. Thus,

$$Pr(\mathcal{U}_r) \dot{\leq} \sum_{l=C(T_{min}+1)+1}^{\infty} \frac{(e/l)^l\left((T_{min}+1)p_{T_{min}}C^\gamma\right)^l}{\sqrt{2\pi l}} \times e^{-(T_{min}+1)p_{T_{min}}C^\gamma} \leq Pr(\mathcal{U}_r).$$

Therefore,

$$Pr(\mathcal{U}_r) \doteq \sum_{l=C(T_{min}+1)+1}^{\infty} \frac{(e/l)^l\left((T_{min}+1)p_{T_{min}}C^\gamma\right)^l}{\sqrt{2\pi l}} \times e^{-(T_{min}+1)p_{T_{min}}C^\gamma}. \tag{17}$$

Furthermore, all of the terms in the sum of (17) decay faster than l^{-l}. Since l takes on values larger than $C(T_{min}+1)$,

$$Pr(\mathcal{U}_r) \doteq \frac{1}{\sqrt{2\pi(C(T_{min}+1)+1)}} \times \left(\frac{(T_{min}+1)p_{T_{min}}C^\gamma e}{C(T_{min}+1)+1}\right)^{C(T_{min}+1)+1} \times e^{-(T_{min}+1)p_{T_{min}}C^\gamma}. \tag{18}$$

Then, $d_{\mathcal{U}_r}(\gamma)$ will be characterized as follows.

$$d_{\mathcal{U}_r}(\gamma) = (1+T_{min})(1-\gamma). \tag{19}$$

Therefore,

$$d_P(\gamma) \geq (1+T_{min})(1-\gamma). \tag{20}$$

2) Lower Bound: To characterize a lower bound on the outage probability we seek a sufficient but not necessary event for an outage to occur. Hence, we introduce the following event,

$$\mathcal{L}_r(n) = \left\{\bigcup_{T_{min}\leq k\leq T_{max}}\left(\sum_{i=T_{min}}^{k} Q_i(n-i) > C(k+1)\right)\right\}, n \to \infty$$

as a sufficient but not necessary condition for an outage at slot n at steady state.

Lemma 5: If $\mathcal{L}_r(n)$ has occurred, then slot n is an outage, and the converse is not true, i.e., $Pr(\mathcal{L}_r(n)) \leq P_P(\text{outage})$.

Proof: To show that $\mathcal{L}_r(n)$ is sufficient for an outage at slot n, consider the following events. First, $Q_{T_{min}}(n-$

$T_{min}) > C(T_{min}+1)$. Since the system can at most serve $C(T_{min}+1)$ requests in an interval of $T_{min}+1$ slots, then, if $Q_{T_{min}}(n-T_{min}) > C(T_{min}+1)$ the system will suffer an outage at slot n. Second, $Q_{T_{min}+1}(n-T_{min}-1)+Q_{T_{min}}(n-T_{min}) > C(T_{min}+2)$. Note that, $Q_k(n-k)$ is the number of requests whose $T = k$ and deadline is slot n. Thus, the second event means that, if the number of the requests whose deadline is slot n, and arrive at the system starting from slot $n-T_{min}-1$, is strictly larger than $C(T_{min}+2)$, then there must be an outage at slot n. In general, the event that $\sum_{i=T_{min}}^{k} Q_i(n-i) > C(k+1), T_{min} \leq k \leq T_{max}$ is sufficient for an outage at slot n. Consequently, $\mathcal{L}_r(n)$ is sufficient for an outage at slot n.

To prove that $\mathcal{L}_r(n)$ is not necessary for an outage at slot n, consider the following counter example. $\sum_{i=T_{min}}^{k} Q_i(n-i) \leq C(k+1), \forall T_{min} \leq k \leq T_{max}$, but $\sum_{i=T_{min}}^{T_{max}} Q_i(n-i) > C$, and $Q_{T_{max}}(n-T_{max}-1) > C(T_{max}+1)$. In this case, the system will not be able to serve the requests whose deadline is slot n except at slot n itself where there may be a possibility to serve part of them. However, the number of those requests is larger than C, then the system will go in outage in slot n. Consequently, $P_P(\text{outage}) \geq Pr(\mathcal{L}_r(n))$. Since the arrival process is stationary, then at the steady state, as $n \to \infty$, $Pr(\mathcal{L}_r(n))$ is independent of n, hence, we use $Pr(\mathcal{L}_r) = Pr(\mathcal{L}_r(n))$. ∎

Let $d_{\mathcal{L}_r}(\gamma) = \lim_{C\to\infty} -\frac{\log(Pr(\mathcal{L}_r))}{C\log C}$. Since $P_P(\text{outage}) \geq Pr(\mathcal{L}_r)$, then $d_{\mathcal{L}_r}(\gamma) \geq d_P(\gamma)$.

$$Pr(\mathcal{L}_r) = Pr\left(\bigcup_{T_{min}\leq k\leq T_{max}} \left(\sum_{j=T_{min}}^{k} Q_j(n-j) > C(k+1)\right)\right) \tag{21}$$

$$\leq \sum_{k=T_{min}}^{T_{max}} Pr\left(\sum_{j=T_{min}}^{k} Q_j(n-j) > C(k+1)\right) \tag{22}$$

$$= \sum_{k=T_{min}}^{T_{max}} \sum_{i=C(k+1)+1}^{\infty} \frac{(F_k C^\gamma)^i}{i!} e^{-F_k C^\gamma} \tag{23}$$

$$\doteq \sum_{i=C(T_{min}+1)+1}^{\infty} \frac{(p_{T_{min}} C^\gamma)^i}{i!} e^{-p_{T_{min}} C^\gamma} \tag{24}$$

$$\leq Pr(\mathcal{L}_r). \tag{25}$$

Therefore,

$$Pr(\mathcal{L}_r) \doteq \sum_{i=C(T_{min}+1)+1}^{\infty} \frac{(p_{T_{min}} C^\gamma)^i}{i!} e^{-p_{T_{min}} C^\gamma}. \tag{26}$$

Using Stirling's approximation,

$$Pr(\mathcal{L}_r) \doteq \sum_{i=C(T_{min}+1)+1}^{\infty} \frac{(e/i)^i \left(p_{T_{min}} C^\gamma\right)^i}{\sqrt{2\pi i}} e^{-p_{T_{min}} C^\gamma}. \tag{27}$$

Since all terms in the above sum decay faster than i^{-i}, then, as $C \to \infty$,

$$Pr(\mathcal{L}_r) \doteq \frac{1}{\sqrt{2\pi(C(T_{min}+1)+1)}} \times \left(\frac{p_{T_{min}} C^\gamma e}{C(T_{min}+1)+1}\right)^{C(T_{min}+1)+1} \times e^{-p_{T_{min}} C^\gamma}. \tag{28}$$

Based on (28), $d_{\mathcal{L}_r}(\gamma)$ will be characterized as follows.

$$d_{\mathcal{L}_r}(\gamma) = (1+T_{min})(1-\gamma). \tag{29}$$

Therefore,

$$d_P(\gamma) \leq (1+T_{min})(1-\gamma). \tag{30}$$

From (20) and (30), it follows directly that

$$d_P(\gamma) = (1+T_{min})(1-\gamma). \tag{31}$$

As can be seen from (31), the diversity gain of random T scenario is dominated by the requests with $T = T_{min}$. However, if $T_{min} \geq 1$, it is still guaranteed that $d_P(\gamma) > d_N(\gamma)$, but in case of $T_{min} = 0$, the system will not see any improvement in the diversity gain **under the previous definition of PMF** in (11). Despite this, our numerical results, reported in Section III-E, still show remarkable gains in **the outage probability** for a wide range of system parameters. Moreover, when the fraction of requests corresponding to $T = 0$ decays as C grows, which is reasonable to expect as most of the new demand corresponds to data traffic that may not be time critical, then the proactive resource allocation framework is able to harness improved prediction diversity gains. This can be viewed as follows. Assume $T_{min} = 0$ and $p_{T_{min}} = p_0 = C^{-\alpha}$, $\alpha > 0$, i.e., $p_0 \to 0$ as $C \to \infty$. By substituting in (18), (28) with $p_0 = C^{-\alpha}$, the diversity gain of the predictive network will be given by,

$$d_P(\gamma) = 1 + \alpha - \gamma \tag{32}$$

as long as $1+\alpha-\gamma$ is smaller than $2(1-\gamma)$ or equivalently, $\alpha \leq 1-\gamma$. Otherwise, be referring to (16), (23), the diversity gain will be determined by the requests with $T = 1$ and will be given by

$$d_P(\gamma) = 2(1-\gamma). \tag{33}$$

Consequently, when the urgent arrivals do not scale with C, the predictive resource allocation technique can achieve strictly larger diversity gain than the non-predictive scenario.

D. The Impact of Prediction Errors

Thus far, we have shown that the proposed proactive resource allocation paradigm will significantly enhance the prediction diversity gain under the assumption of perfect, i.e., error free, prediction. Now, we introduce a model that takes into account the effect of prediction error on the traffic behavior and, consequently, the prediction diversity gain. In our analysis we consider the deterministic T scenario, and assume that the traffic of the non-predictive system is characterized by the process $Q(n), n > 0$ which represents the number of

arriving requests at the beginning of time slot n with $T = 0$. This process is Poisson with rate C^{γ}. Moreover, the system is operating according to SP1. The following two events are as the causes of prediction errors in our model:

1) The predictive network mistakenly predicts a request and serves it causing a waste of resources.
2) The predictive network fails to predict a request and, as a consequence, it encounters an urgent arrival (unpredicted request that should be served in the same slot of arrival) later.

Based on these two sources of errors, the traffic of the predictive network is no longer the same as the traffic of the non-predictive network. Instead, the predictive network with errors has a new traffic model $Q^E(n)$, $n > 0$ which can be regarded as the superposition of two arrival process: 1) $Q'(n)$ is the arrival process of the predicted requests. It represents the number of arriving requests at the beginning of time slot n with deadline $n + T$. 2) $Q''(n)$ is the arrival process of the unexpected requests. It represents the number of arriving requests at the beginning of time slot n and must be served in the same slot because the network has failed to predict them.

In general, the prediction mechanism employed by the predictive network can be thought of as a system with input traffic $Q(n)$ and output traffic $Q^E(n) = Q'(n)+Q''(n)$ where a deadline of $n + T$ is associated with $Q'(n)$. The predictive network, hence, can achieve an improved diversity gain performance as long as the diversity gain obtained through the new arrival process $Q^E(n)$ is larger than the diversity gain of the non-predictive network, i.e., $1 - \gamma$.

For instance, we consider a special case of prediction mechanisms in which: $Q'(n)$ is Poisson with rate $C^{\gamma'}$, where $\gamma' \in \Re$, and $Q''(n)$ is Poisson with rate $C^{\gamma''}$, $\gamma'' \leq \gamma$ such that

$$C^{\gamma'} + C^{\gamma''} \geq C^{\gamma}. \tag{34}$$

Note that, $\gamma'' \leq \gamma$ is following directly from a natural constraint on the arrival rate of the unexpected requests that cannot exceed the arrival rate of requests in the non-predictive scenario. The prediction mechanism is supposed to reduce the rate of those urgent requests. Moreover, the constraint (34) is an implication of the imperfect prediction technique. Precisely, the inaccurate estimates of the potential requests causes error event 1 which yields additional unnecessary traffic. Hence, a necessary and sufficient condition for perfect prediction is $\gamma' = \gamma$ and $\gamma'' = -\infty$, where in this case $Q^E(n) = Q'(n) = Q(n + T)$.

Moreover, we assume that, given γ' and γ'', both processes $Q'(n)$ and $Q''(n)$ are independent. Hence, by setting $\gamma' = \alpha'\gamma$ and $\gamma'' = \alpha''\gamma$ where $\alpha' \in \Re$ and $\infty < \alpha'' \leq 1$, the diversity gain of the predictive network will be given by[1]

$$d_P(\gamma) = \min\{(1 + T)(1 - \alpha'\gamma), 1 - \alpha''\gamma\}. \tag{35}$$

[1]Following similar analysis to that of Section III, the outage probability of the predictive network with errors is asymptotically dominated by either $C + 1$ urgent arrivals or $C(T + 1) + 1$ predictable arrivals.

Consequently, the optimum operating point for the system is such that the two quantities inside the $\min\{,\}$ are equal, i.e.,

$$(1 + T)(1 - \alpha'\gamma) = (1 - \alpha''\gamma) \tag{36}$$

or

$$T = \frac{(\alpha' - \alpha'')\gamma}{1 - \alpha'\gamma}. \tag{37}$$

Furthermore, the predictive system can achieve a *strictly* improved diversity gain over the non-predictive system if and only if,

$$\min\{(1 + T)(1 - \alpha'\gamma), 1 - \alpha''\gamma\} > 1 - \gamma. \tag{38}$$

Hence, if the system is operating according to (36), it is sufficient to have

$$(1 - \alpha''\gamma) > (1 - \gamma) \tag{39}$$

or equivalently

$$\alpha'' < 1 \tag{40}$$

in order to see an improved diversity gain. Consequently, for any value of $T > 0$,

$$\frac{(\alpha' - \alpha'')\gamma}{1 - \alpha'\gamma} > 0$$

yielding

$$\alpha'' < \alpha' \leq \frac{1}{\gamma}. \tag{41}$$

Finally, from (40), and in the infinite C asymptotic, condition (34) will reduce to,

$$\alpha' \geq 1. \tag{42}$$

Note that, (42) implicitly satisfies the lower bound of (41).

The design of the prediction mechanism, therefore, should ensure that the values of T, α' and α'' satisfy

$$T = \frac{(\alpha' - \alpha'')\gamma}{1 - \alpha'\gamma}, \tag{43}$$

$$1 \leq \alpha' \leq \frac{1}{\gamma}, \tag{44}$$

$$\alpha'' < 1 \tag{45}$$

in order to achieve the *largest* diversity gain that is *strictly* better than that of the non-predictive network.

For example, consider the design of a prediction mechanism that is supposed to predict the requests by $T = 4$ slots in advance. The designed mechanism is required to result in a diversity gain of 0.92 which is strictly larger than that of the non-predictive system with $\gamma = 0.8$. Then, the prediction diversity gain of 0.92 is the best that the prediction mechanism can achieve if $(1 + T)(1 - \alpha'\gamma) = (1 - \alpha''\gamma) = 0.92$, meaning that $\alpha'' = 0.1$, i.e., the mechanism results in urgent arrivals with rate $C^{0.1\gamma}$. Consequently, the mechanism should guarantee that $\alpha' = 1.02$. That is, the predicted requests should arrive at a rate of $C^{1.02\gamma}$. Note that, when the prediction mechanism is perfect, a prediction diversity gain of 1 is attained.

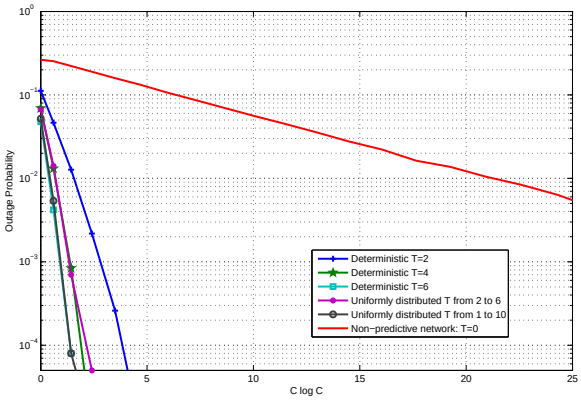

Fig. 2. Outage probability vs. $C \log C$ with $\gamma = 0.8$.

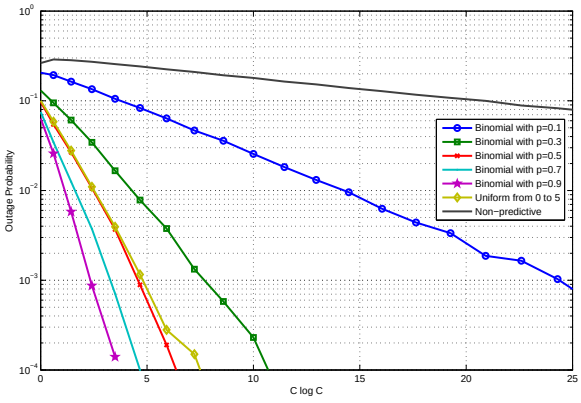

Fig. 3. Effect of different distributions T on the outage performance ($\gamma = 0.9$).

E. Numerical Results

In this section we present numerical results that illustrate the performance gain offered by the proactive resource allocation framework. In Fig. 2 we plot the outage probability of predictive and non-predictive networks versus $C \log C$. The simulation is based on SP1 with $\gamma = 0.8$. At each value of C, the system is simulated for 10^3 time slots and the performance is averaged over 10^2 simulation runs. It is clear, from the results, that there is a remarkable reduction in the resources required to attain a certain level of outage probability when the network employs a predictive mechanism for resource allocation. Moreover, for the two simulated random T scenarios, although T_{min} is chosen to be 2 and 1, the corresponding outage probability curves are upper bounded by the outage probability of the predictive case with deterministic $T = 2$. This actually may be a consequence of the small values of C in this figure. Here, the averaging effect over the range between T_{min} and T_{max} appears to have a more favorable impact on the performance than increasing T_{min}.

Fig. 3 investigates the effect of the distribution of T on the outage (blocking) probability. Here, we consider a class of

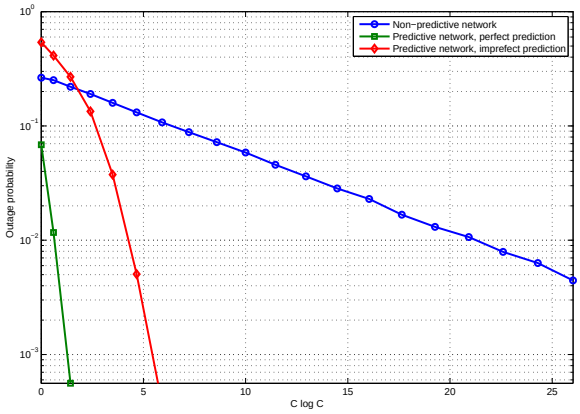

Fig. 4. Effect of imperfect prediction on the performance of the outage probability.

binomial distributions with finite support from $T_{min} = 0$ to $T_{max} = 5$ and parameter p. That is,

$$Pr(T = t) = p_t = \binom{T_{max}}{t} p^t (1-p)^{T_{max}-t}$$

where $T_{min} \leq t \leq T_{max}$. The predictive system is then simulated for different values of p and the outage probability results are depicted. Moreover, the uniform distribution of T over the interval $T_{min} = 0$ to $T_{max} = 5$ is plotted on the same figure. From the results, one can argue that the outage performance is sensitive to the value of $p_{T_{min}}$ over the simulated range of C. Since the binomial distributions of T, $p_{T_{min}}$ is monotonically decreasing with p and thus, as the weight of the arrivals with $T = 0$ increases the outage behavior becomes worse although all of the outage curves have the same diversity gain in infinite C asymptotic. Also, in case of a uniform distribution, the outage probability curve is quite close to that of the binomial distribution with $p = 0.5$ although $p_{T_{min}}$ of the uniform is larger that its peer of the binomial with $p = 0.5$. The reason behind this behavior is that, the weights of the higher values of T in case of the uniform distribution are larger than their peers in case of the binomial distribution with $p = 0.5$. This advantage enables the scheduler to efficiently reduce the outage probability despite the relatively large probability corresponding to $T = 0$ in the uniformly distributed T.

Finally, Fig. 4 compares the outage probability of the imperfect prediction mechanism described in the example of Section III-D to the outage performance of two networks: a non-predictive network with $\gamma = 0.8$ and an idealistic predictive network with perfect prediction at $T = 4$. It is obvious that errors in prediction cause the outage curve to be shifted to the right of the ideal curve as well as a small loss in the diversity gain (characterized in Section III-D). However, the predictive network with errors still achieves a significantly improved outage performance over the non-predictive scenario.

IV. Different QoS Users: The Good Citizen Phenomenon

The previous section demonstrates the potential gains that can be leveraged from the proactive resource allocation framework when all the requests belong to the same class of QoS. In this section we consider a network with two QoS classes that can be considered as primary and secondary users sharing the same resources. We investigate the effect of prediction by **the primary user only** on the prediction diversity gain of the secondary network. Clearly, our analysis can be extended to allow for prediction by the secondary user as well; but we choose to limit ourselves to this special case for simplicity. We assume that the number of secondary arrivals at the beginning of time slot n is $Q^s(n)$, where $Q^s(n)$ follows a Poisson distribution with rate $\lambda^s = C^{\gamma^s}$, $0 \leq \gamma^s \leq 1$. The number of primary requests arriving at the beginning of time slot n is $Q^p(n)$ that follows a Poisson distribution with rate $\lambda^p = C^{\gamma^p}$, where $0 \leq \gamma^p \leq 1$. We assume that the system is dominated by primary arrivals, that is, $\lambda^p > \lambda^s$ or, equivalently, $\gamma^p > \gamma^s$. The secondary and primary arrival processes are ergodic and independent.

A. Non-Predictive Primary User

We analyze the outage probability of the secondary user and its diversity gain when the primary user is non-predictive. At the beginning of time slot n, the system is supposed to witness $Q^p(n)+Q^s(n)$ arriving requests with deadline is slot n, i.e., must be served in the same slot of arrival. The primary system has a fixed capacity C per slot. In order to enhance the utilization of its resources, the primary user allows secondary requests to be served by the remaining resources from serving the primary requests. Thus, at slot n, the remainder of $C - Q^p(n)$ is assigned to serve the secondary requests. Based on this technique, the outage probability of the primary system is identical to $P_N(\text{outage})$ calculated in (1). We denote the primary outage probability in this case by $P_N^p(\text{outage})$. As a result, the primary diversity gain is given by

$$d_N^p(\gamma^p, \gamma^s) = 1 - \gamma^p. \tag{46}$$

The secondary system under non-predictive primary network encounters an outage at a given slot when the remaining resources from serving the primary requests at this slot are less than the number of arriving secondary requests at the beginning of the same slot. Thus, if the primary network suffers an outage in a certain slot with at least one arriving secondary request, the secondary system goes in outage as well. The secondary system, consequently, encounters an outage at slot n if and only if

$$Q^p(n) + Q^s(n) > C \quad \text{and} \quad Q^s(n) > 0.$$

Let the outage probability of the secondary network when the primary network is non-predictive be denoted by $P_N^s(\text{outage})$, hence

$$P_N^s(\text{outage}) = Pr\left(Q^p(n) + Q^s(n) > C, Q^s(n) > 0\right). \tag{47}$$

The two random variables $Q^p(n) + Q^s(n)$ and $Q^s(n)$ are dependent but their joint distribution can simply be obtained by transformation of variables. By setting $Y = Q^p(n)+Q^s(n)$ and $U = Q^s(n)$, the exact expression of $P_N^s(\text{outage})$ will be given by

$$\begin{aligned} P_N^s(\text{outage}) &= Pr(Y > C, U > 0) \\ &= \sum_{y=C+1}^{\infty} \sum_{u=1}^{y} \frac{C^{\gamma^p(y-u)+\gamma^s u}}{(y-u)!u!} e^{-(C^{\gamma^p}+C^{\gamma^s})}. \end{aligned} \tag{48}$$

The diversity gain of the secondary system coexisting with a non-predictive primary network is defined by

$$d_N^s(\gamma^p, \gamma^s) \triangleq \lim_{C\to\infty} \frac{-\log P_N^s(\text{outage})}{C \log C}.$$

For large values of C, the outer sum of the right hand side of (48) is dominated by $y = C + 1$. However, the inner sum is not dominated by a single value of u because of $(y-u)!u!$ in the denominator. Consequently, as $C \to \infty$, $P_N^s(\text{outage})$ can be written as

$$P_N^s(\text{outage}) \doteq \sum_{u=1}^{C+1} \frac{C^{\gamma^p(C+1-u)+\gamma^s u}}{(C+1-u)!u!} e^{-(C^{\gamma^p}+C^{\gamma^s})}. \tag{49}$$

Characterizing $d_N^s(\gamma^p, \gamma^s)$ from (49) is, however, difficult, so we consider another approach based on the asymptotic behavior of upper and lower bounds on $P_N^s(\text{outage})$.

1) Upper Bound on $P_N^s(\text{outage})$: Since $Pr(\mathcal{A}, \mathcal{B}) \leq Pr(A)$ with equality if and only if $\mathcal{A} \subseteq \mathcal{B}$, then

$$P_N^s(\text{outage}) \leq Pr(Q^p(n) + Q^s(n) > C) \tag{50}$$

$$= \sum_{k=C+1}^{\infty} \frac{(C^{\gamma^p} + C^{\gamma^s})^k}{k!} e^{-(C^{\gamma^p}+C^{\gamma^s})} \tag{51}$$

$$\doteq \left(\frac{(C^{\gamma^p} + C^{\gamma^s})e}{C+1}\right)^{C+1} \frac{e^{-(C^{\gamma^p}+C^{\gamma^s})}}{\sqrt{2\pi(C+1)}}. \tag{52}$$

Since $P_N^s(\text{outage}) \leq Pr(Q^p(n) + Q^s(n) > C)$, then

$$d_N^s(\gamma^p, \gamma^s) \geq \lim_{C\to\infty} \frac{-\log(Pr(Q^p(n) + Q^s(n) > C))}{C \log C} \tag{53}$$

$$= 1 - \max\{\gamma^p, \gamma^s\} \tag{54}$$

$$= 1 - \gamma^p. \tag{55}$$

2) Lower Bound on $P_N^s(\text{outage})$: We consider the event that there is at least one secondary arrival with a primary outage at slot n as a sufficient but not necessary condition on a secondary outage at slot n. That is,

$$\mathcal{L}_N^s(n) \triangleq \{Q^p(n) > C, Q^s(n) > 0\}, \quad n \to \infty.$$

Note that, the event $\mathcal{L}_N^s(n)$ is not necessary for a secondary outage at slot n as there may be $Q^p(n) < C$ but $Q^s(n) > C - Q^p(n)$ which results in a secondary outage at slot n too. Furthermore, at steady state, $Pr(\mathcal{L}_N^s(n))$ becomes independent of n as both arrival processes, $Q^p(n)$ and $Q^s(n)$, are stationary, hence we use $Pr(\mathcal{L}_N^s)$ instead. Since $\mathcal{L}_N^s(n)$ is a sufficient

condition for a secondary outage, then $P^s_N(\text{outage}) \geq Pr(\mathcal{L}^s_N)$ and $d^s_N(\gamma^p, \gamma^s) \leq \lim_{C\to\infty} \frac{-\log Pr(\mathcal{L}^s_N)}{C \log C}$.

$$Pr(\mathcal{L}^s_N) = Pr(Q^p(n) > C, Q^s(n) > 0) \quad (56)$$

$$= Pr(Q^p(n) > C).Pr(Q^s(n) > 0) \quad (57)$$

$$= \sum_{k=C+1}^{\infty} \frac{(C^{\gamma^p})^k}{k!} e^{-(C^{\gamma^p})}.\left(1 - e^{-C^{\gamma^s}}\right) \quad (58)$$

$$\doteq \left(\frac{C^{\gamma^p} e}{C+1}\right)^{C+1} \frac{e^{-C^{\gamma^p}}}{\sqrt{2\pi(C+1)}}.\left(1 - e^{-C^{\gamma^s}}\right) \quad (59)$$

Therefore

$$\lim_{C\to\infty} \frac{-\log Pr(\mathcal{L}^s_N)}{C \log C} = 1 - \gamma^p$$

and

$$d^s_N(\gamma^p, \gamma^s) \leq 1 - \gamma^p. \quad (60)$$

From (55), (60), it follows that

$$d^s_N(\gamma^p, \gamma^s) = 1 - \gamma^p. \quad (61)$$

B. Predictive Primary User

In this case, the system can predict the primary arrivals only by T time slots in advance. We assume that T is deterministic and fixed for all primary requests, i.e., the deadline for the primary requests $Q^p(n)$ is $n + T$. The system, however, is assumed to be non-predictive for the secondary requests, i.e., the deadline for the secondary requests $Q^s(n)$ is n. When this system dedicates all the per-slot capacity C to serve the primary requests according to SP1, at least one of the secondary users arriving at the beginning of time slot n will be served if and only if C is strictly larger than the number of primary requests *existing* in the system at the beginning of time slot n. Unfortunately, this service policy does not enhance the outage performance of the secondary system although it minimizes the outage probability of the primary. The main reason behind that is the large variations in the number of served primary requests per slot that takes on values from 0 to C. These variations are quite close to the variations in the number of served primary requests per slot in case of non-predictive primary network. Thereby, a predictive primary network following SP1 does not implicitly enhance the outage probability of the secondary system. Fig. 5 plots the outage probability of the primary and secondary networks versus $C \log C$ under the two types of primary network, predictive and non-predictive. The results are based on simulation of SP1 over $M = 10^3$ slots and averaging over 100 simulation runs. It is clear that the outage probability of the primary system when the primary network is predictive is significantly improved over its peer when the primary network is non-predictive. However, it can be noted that, minimizing the outage probability of the primary network via SP1, when the primary network is predictive, does not enhance the outage probability of the secondary network. Instead, the primary system becomes greedy to serve as much as possible of primary requests in the system without considering the secondary arrivals.

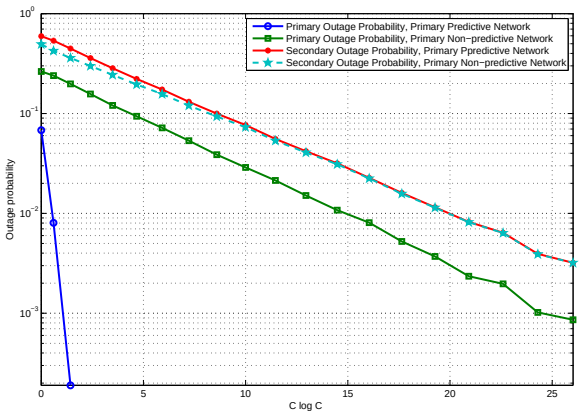

Fig. 5. Outage probability vs. $C \log C$ for primary and secondary networks under the two types of primary network: predictive and non-predictive. All are calculated assuming SP1 ($\gamma^p = 0.75$, $\gamma^s = 0.05$ and $T = 4$).

However, the primary network can behave in a less selfish manner by modifying its service policy so that the prediction at the primary side only can be useful from the point of view of the outage probability of the secondary network and at the same time the diversity gain of the primary network is not affected. The main idea behind this is to devise a new service policy for the primary network that minimizes the probability of the *dominant* outage event instead of minimizing the overall outage probability of the primary network. Thus, the diversity gain of the primary network will not be affected while more opportunities for secondary requests will be created as the primary network will not act greedily as in SP1. Consequently, the outage probability of the secondary network will be enhanced at the same diversity gain of the primary network.

One possible new service policy is the following.

Definition 6: Service Policy 2 (SP2):
The primary network is assigned a fixed capacity per slot of $C - \lfloor C^\beta \rfloor$ where $\beta < 1$. It uses this fixed capacity to serve as much as possible of primary requests in the system at each time slot starting from the requests with the closest deadlines to the furthest ones. This policy is almost the same as SP1, the only difference is that, part of the primary network capacity here (which is $\lfloor C^\beta \rfloor$) is taken out from the primary network and dedicated to serve the secondary requests.

The diversity gain of the primary network when following SP2 is exactly the same if SP1 is followed. That is, $d^p_P(\gamma^p) = (1 + T)(1 - \gamma^p)$, moreover, we show numerically in Section III-E that the outage probability of the secondary network is improved because of the dedicated capacity of $\lfloor C^\beta \rfloor$.

In SP2 the primary network is assumed to have a fixed capacity per slot and use this capacity to serve the existing primary requests by the same way that SP1 does. Because of the variability of the arriving requests with time, the primary network may determine its per slot capacity adaptively taking into account the number of requests in the system at each slot and their deadlines as well. One of these policies can be described as follows.

Definition 7: Service Policy 3 (SP3):
Let $N^p(n)$ be the number of the primary requests in the system at the beginning of time slot n, and $N_d^p(n)$ be the number of these requests whose deadline is slot n. Then, the capacity of the primary network at slot n is calculated as

$$\min\{C, N_d^p(n) + f \times (N^p(n) - N_d^p(n))\}$$

where $0 \leq f \leq 1$. After that, the network serves the primary requests by the same way of SP1 (serves the requests in order from the closest deadline to the furthest).

Note that, the performance of SP3 is highly dependent on the design parameter f. At $f = 0$, the system, at steady state, is serving only the requests whose deadline is the current slot. In this case the system will be similar to the non-predictive network in terms of primary and secondary outage probabilities. At $f = 1$, the system is exactly following SP1 as at each time slot the largest possible number of the primary requests in the system are served with their deadlines ordered from the closest to the furthest. However, we show numerically in Section III-E that for some values of f the outage probability of the secondary system is significantly enhanced at almost no losses in the outage probability of the primary network.

C. Numerical Results

The performance of a network with primary and secondary users has been evaluated numerically with the same parameters of Fig. 5 and the results are shown in Figs. 6, 7. In Fig. 6 the outage probability of a primary network following SP2 with $\beta = 0.3$ is shown. It is clear from the figure that the outage probability of the secondary network is enhanced over the non-predictive case. However, this improvement comes at the expense of shifting the outage probability curve of the primary network to the right, i.e., the diversity gain of the primary network operating according to SP2 is the same as that of SP1, but the outage probability curve itself is shifted to the right of that of SP1. Moreover, although improved, the outage probability of the secondary network appears to have no gain in the decay rate with $C \log C$.

In Fig. 7, SP3 is simulated for $f = 0.5$. Compared with SP2, the behavior of SP3 is shown to remarkably enhance the outage probability of the secondary user at an almost negligible loss in the primary outage performance. The analysis of the diversity gain of the primary and secondary users operating according to SP3, however, are still under investigations. Overall, it can be concluded that, prediction at the primary side only does not **only** enhance the primary spectral efficiency, but it can be efficiently exploited to significantly improve the spectral efficiency of the coexisting non-predictive secondary users (networks) as well.

V. Conclusions

We have introduced a novel paradigm for resource allocation for wireless networks which exploits the predictability of user behavior to minimize the required spectral resources (e.g., bandwidth) to achieve certain QoS metrics. Unlike the

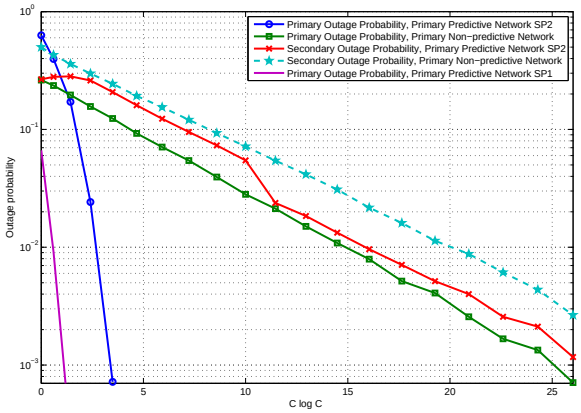

Fig. 6. Outage probability vs. $C \log C$ of the primary and secondary users with $\gamma^p = 0.75$, $\gamma^s = 0.05$, $T = 4$ and $\beta = 0.3$.

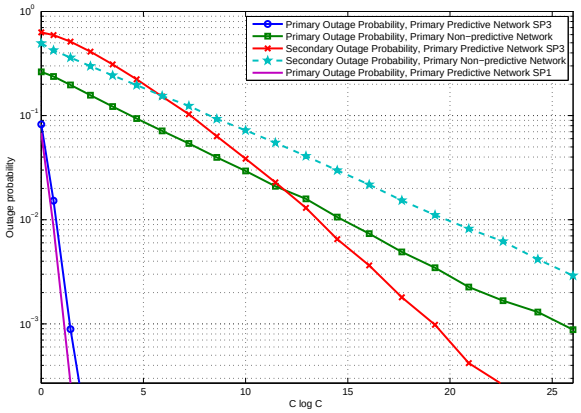

Fig. 7. Outage probability vs. $C \log C$. of the primary and secondary users with $\gamma^p = 0.75$, $\gamma^s = 0.05$, $T = 4$ and $f = 0.5$.

tradition reactive resource allocation approach, in which the network can only start serving a particular user request upon its initiation, our proposed algorithms anticipate future requests, and hence, allows 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 our proactive resource allocation framework to achieve significant spectral efficiency gains in several interesting scenarios. More specifically, we 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 a network with two QoS classes, prediction at one side only does not only enhance its diversity gain, but it also improves the outage probability performance of the other user significantly. Throughout the paper, our theoretical claims were supported by numerical results that illustrate 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, our results should motivate further investigations on the design of efficient prediction algorithms; which will possibly require advanced tools from machine learning in addition to accurate models for user behavior that captures the predictability of traffic requests. Another avenue for future work is the cross layer optimization of content delivery over wireless networks under the proactive resource allocation models (i.e., the potential for multicast, peer-to-peer, and coupling between the time scales of different layers).

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

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