Dynamic Proactive Caching in Relay Networks

Rana A. Hassan*, Ahmed M. Mohamed*, John Tadrous†, Mohamed Nafie* ‡, Tamer ElBatt * ‡, and Fadel Digham *
* Wireless Intelligent Networks Center (WINC), Nile University, Giza, Egypt
†Department of Electrical and Computer Engineering, Gonzaga University, WA, USA
‡ Department of EECE, Faculty of Engineering, Cairo University, Giza, Egypt
Email: r.fouad@nu.edu.eg, ah.magdy@nu.edu.eg, tadrous@gonzaga.edu, mnafie@nu.edu.eg, telbatt@ieee.org, fdigham@nu.edu.eg

***Abstract*—We investigate the performance of dynamic proactive caching in relay networks where an intermediate relay station caches content for potential future use by end users. A central base station proactively controls the cache allocation such that cached content remains fresh for consumption for a limited number of time slots called *proactive service window*. With uncertain user demand over multiple data items and dynamically changing wireless links, we consider the optimal allocation of relay stations cache to minimize the time average expected service cost. We characterize a fundamental lower bound on the cost achieved by any proactive caching policy. Then we develop an asymptotically optimal caching policy that attains the lower bound as the proactive caching window size grows. Our analytical findings are supported with numerical simulations to demonstrate the efficiency of the proposed relay-caching.**

I. Introduction

The tremendous increase in demand for spectrum-based services has led end users to experience a major demand and supply mismatch during the whole day. During peak periods, the demand level is high and may reach the network capacity causing congestion. However, during off-peak periods the network resources are underutilized. Thus the concerns of spectrum under-utilization have been raised due to the spatial and temporal variations in the activity of wireless users [1]. This study is also strengthened by the recent data traces collected by major European operators in [2]. Thus, during peak periods service providers incur excessive costs to provide reliable delivery of the requested data, on the contrary at off-peak periods.

It is predicted that mobile data traffic will see a nine-fold increase by the end of 2020 compared to 2014 [3]. That is why networks should consider employing advanced resource allocation techniques in order to balance the rapid increase in the user demand. There has been extensive research to tackle such a problem, some of which has employed reactive resource allocation techniques where the user is served when the request is initiated by the user. Under heavy traffic conditions, reactive techniques suffer from huge penalty as degrade network performance.

Predictability of a wireless user demand is supported by a growing body of evidence that ranges from the launch of Google Instant to the interesting findings on the predictable mobility patterns [4]. Also the user's experienced channel quality metrics (CQM) including received signal strength and interference levels are predictable as well [5], [6].

However, one of the technologies for 5G wireless networks is proactive resource allocation [7]. The main idea of proactive resource allocation is to leverage the predictable characteristics of the users demand to smooth out network traffic variations across the peak and off-peak periods of the day so that spectrum utilization is improved. When a predictive network serves a request before its actual request time, the data is stored in the cache memory of either the wireless device or any intermediate node. Then at the actual time of demand, the requesting application pulls the data directly from the memory instead of accessing the wireless network. By this way the user's quality of experience (QoE) is enhanced.

In [8], the authors introduced a novel proactive resource allocation paradigm by exploiting the predictability of the user behavior. They provided a solid theoretical background and explained significant spectral efficiency gains in different scenarios. In [9], proactive resource allocation schemes under time-invariant and time-varying demand statistics are studied. The authors proposed fundamental lower bounds on the achievable costs, and developed asymptotically optimal policies that approach these bounds when the the window of time slots over which predictable future demand can be proactively served while being fresh for consumption called *proactive service window* is increased. However, the utilization of predicted CQM for proactive resource allocation is not captured in proactive scheduling thus far. Authors in [10] studied proactive resource allocation strategies that exploit both the predictable data and channel demand characteristics, with uncertainties. In [9] and [10] the authors assumed the presence of a single file to be served to the end user implying full certainty about the exact content to be requested, given a request is sent. The authors in [9]- [13] assumed the proactive caching to take place at the end user.

In this work, we study a proactive allocation scheme when the cache is at intermediate node between the base station (BS) and the end user. This caching technique improves our performance metric which is the service cost. In addition it can also improve other metrics, for instance battery and memory consumption at the end users, compared to caching at end user. Also 5G networks are going to enable relay stations (RS) at the network edges, and these smart relays are equipped

This publication was made possible by NPRP grant number NPRP 7-923-2-344 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

978-3-901882-90-6/17 ©2017 IFIP

with storage capabilities that can proactively maintain relevant content for network resource optimization [14]- [16].

We consider a network which consists of a single BS connected to a single RS equipped with storage in order to serve end users. We consider a time-invariant demand statistics model in which all incoming requests from data networks are statistically indistinguishable over time. We introduce in our system model the uncertainty about the requested files which is not addressed in previous work. We further consider time-invariant channel statistics model in which user experiences statistically indistinguishable channels over time.
Our main contributions in this work are as follows.

- We extend the work in [10] by considering the presence of a set of multiple files or data items with different popularity where the user can request any of which every time slot. Thus incorporating an additional element of uncertainty due to the randomness of the requested content.
- We consider the effect of caching at an intermediate node between BS and end user on the service cost.
- We characterize, analytically, a fundamental lower bound on the average expected cost incurred by any proactive caching policy.
- We develop an asymptotically optimal policy that achieves the fundamental lower bound as the proactive window size grows to infinity.
- We show that our system considerably outperforms both non-proactive schemes in terms of cost and the counter-part proactive system with caching at end-user because it exploits channel diversity.

The rest of the paper is organized as follows. In Section II, we present our system model. The proposed lower bound and policy are introduced in Section III. Numerical results are presented in Section IV. The paper is concluded in Section V.

Notation: $\mathbb{P}[.]$ denotes the probability of a random variable and $\mathbb{E}[.]$ denotes the expected value of a random variable.

II. System Model

We consider a network model that comprises a BS serving data demand from end users. A RS equipped with data storage intermediates the connection between end users and BS, as shown in Fig. 1. While in reality data demand is generated by multiple end users, we focus on the service of single user demand for the ease of exposition as a proof of concept. Nevertheless, our analysis and technique directly generalize to the multi-user scenario, which will be addressed in future work [17].

Dynamic Data Content: The network is assumed to serve data content from a finite set $\mathcal{F} = \{1, \cdots, F\}$ of F disjoint files or data items. Each file $f \in \mathcal{F}$ represents a different type of dynamically changing content whose size is S data units. We assume a time slotted operation where in every time slot the content of each file is consistently updated such that the content of any data item f at time t is no longer *fresh* for consumption at time slot $t+T$, for some $T \geq 1$. Examples on such dynamically changing content include news, On-Demand Video services, traffic information, and social network updates.

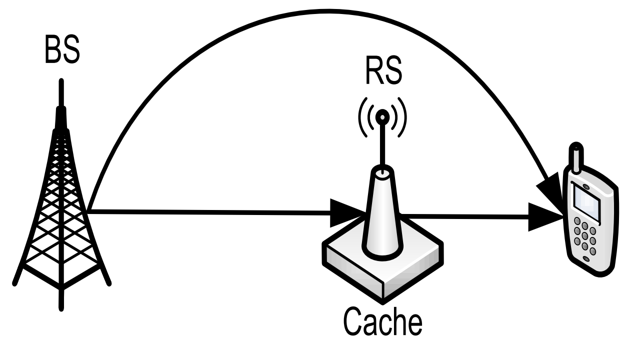

Fig. 1: Caching system at the RS for a single end user.

User Demand Profile: We assume that the time slot size is sufficient to serve only one data item request. We assume that future requests are not perfectly anticipated. This means that the user demand for files in the next T time slots is known only statistically. We assume that file requests are independent and identically distributed (i.i.d.) random variables across time. We assume the presence of a set of requested files denoted by $\mathcal{R}$. We define a random variable $r(t) \in \mathcal{R}$ where $\mathcal{R} = \{0, \cdots, F\}$ to represent the index of the requested file at time slot t, where $r(t) = 0$ indicates that no file is requested at time slot t. The popularity of file $f \in \mathcal{F}$ is $p_f = \mathbb{P}(r(t) = f)\ \forall t$, which is considered to be given and constant across time. The probability of demand is denoted by p where $p = \sum_{f=1}^{F} p_f$. Also, the user remains silent with probability q where, $q = 1 - p$.

Let $\rho(r(t)) \in \{0, 1\}$ be a random variable that represents the presence of a request or not in time slot t where,

$$\rho(r(t)) = \begin{cases} 1, & \text{for } r(t) \geq 1 \\ 0, & \text{for } r(t) = 0 \end{cases}.$$

We assume that the files' popularity is characterized by a Zipf distribution with parameter ψ, which is commonly used to model content popularity in data networks. Hence, the popularity of file f is expressed as

$$p_f = \gamma \frac{1}{f^{\psi}}.$$

where $\gamma = \frac{1}{\sum_{i=1}^{F} \frac{1}{i^{\psi}}}$ and $f = 1, \cdots, F$.

The Zipf parameter ψ characterizes the distribution by controlling the relative popularity of files. Larger values of ψ imply steeper distribution and hence more certainty on the exact file to be requested, whereas smaller values of ψ indicate a more uniform distribution and less certainty about the user preference. We assume that the BS knows the user demand profile, which captures the statistical characteristics of the future demand.

Channel Model: We assume that the user experiences in each time slot a channel gain g_{B}, which represents the wireless channel between the BS and the user. The channel gain is one of N_{B} discrete channel realizations from a set $G_{\mathrm{B}} = \{g_{\mathrm{B}}^{(1)}, \cdots, g_{\mathrm{B}}^{(N_{\mathrm{B}})}\}$ with corresponding probabilities $\alpha = \{\alpha_1, \cdots, \alpha_{N_{\mathrm{B}}}\}$, where $\sum_{n=1}^{N_{\mathrm{B}}} \alpha_n = 1$. Similarly, for the link between the RS and the user, the user experiences a channel gain g_{R} which is one of N_{R} possible discrete channel states from a set $G_{\mathrm{R}} = \{g_{\mathrm{R}}^{(1)}, \cdots, g_{\mathrm{R}}^{(N_{\mathrm{R}})}\}$ with corresponding probabilities $\beta = \{\beta_1, \cdots, \beta_{N_{\mathrm{R}}}\}$, where $\sum_{n=1}^{N_{\mathrm{R}}} \beta_n = 1$. We assume that the channels on both links are independent of each other and each channel realization is i.i.d. across different time slots, yet remains constant over the time slot duration. We assume that the channel between the BS and the RS $g_{\mathrm{B,R}}$ is fixed and independent across time slots, where it could be a wired backhaul link or a line of sight microwave link. All channel realizations are assumed to be non negative and finite. We consider uncorrelated channels, especially we consider a large timescale operation where a time slot lasts for a few minutes, long enough to have independent channel realizations.

III. Problem Formulation

Operational Cost: We consider the total cost to serve the user from both the BS and RS. We denote $C_d(x)$, $d \in \{\mathrm{B}, \mathrm{R}\}$ as the cost function for serving $x \geq 0$ data units in a time slot. The cost function is assumed to be monotonically increasing and strictly convex. Furthermore, the cost due to communication over a channel with gain $g \geq 0$ is $\eta(g)$. We can view that for example the cost to use the channel can be related to the amount of power, where the power needed to serve a certain content of size B bits over a channel with gain g under noise variance N_0 is $\frac{(2^B-1)N_0}{g}$. Thus the cost decreases with $\frac{1}{g}$ and vice-versa.

Reactive Service Model: The reactive network is considered to be our baseline system. Under reactive paradigm, user requests are served only after they have been actually initiated and they receive service in the same slot of initiation. Here, there is no proactive caching and all requested files are transmitted from the BS directly to the user. The amount of load generated by the BS at time slot t for a reactive network is described as follows,

$$L_{\mathrm{B}}^{\mathrm{re}}(t) = S\rho(r(t))\eta(g_{\mathrm{B}}(t)).^1 \tag{1}$$

The superscript re indicates reactive operation. The time average expected cost under the reactive model is as follows,

$$C^{\mathrm{re}} = \limsup_{t\to\infty} \frac{1}{t} \sum_{l=0}^{t-1} \mathbb{E}\left[C_{\mathrm{B}}\left[L_{\mathrm{B}}^{\mathrm{re}}(l)\right]\right] = C_{\mathrm{B}}\left[L_{\mathrm{B}}^{\mathrm{re}}(1)\right], \tag{2}$$

since both $r(t)$ and $g_{\mathrm{B}}(t)$ are i.i.d. random variables across time.

Proactive Service Model: We assume that both the BS and the RS are aware of the statistics of both the user demand profile and the channel realization across T time slot proactive service window. We denote $x_r^t(k)$ as the amount of proactive download from file $r \in \mathcal{R}$ in time slot k for the future request to be made in time slot t where $t-T \leq k \leq t-1$. The proactive download term cannot be negative and cannot exceed the total file size S. The amount of load generated by the BS at time slot t for the proactive network is given by

$$\begin{aligned} L_{\mathrm{B}}^{\mathrm{pro}}(t) = & \left(S - \sum_{k=t-T}^{t-1} x_r^t(k)\right)\rho(r(t))\eta(g_{\mathrm{B}}(t)) \\ & + \sum_{f\subset\mathcal{F}} \sum_{k=t+1}^{t+T} x_f^k(t)\eta(g_{\mathrm{B,R}}). \end{aligned} \tag{3}$$

The superscript pro indicates proactive operation. The first term represents the on-time service component resulting from the non-proactively served part of the requested file r, where $\sum_{k=t-T}^{t-1} x_r^t(k)$ is the past applied proactive caching. The second term is the proactive service of future requests in the upcoming T-slot interval, while the term $\sum_{k=t+1}^{t+T} x_f^k(t)$ captures the proactive service to be applied over the next T slots.

Similarly, the amount of load generated from the RS at time slot t for proactive network is described as follows,

$$L_{\mathrm{R}}^{\mathrm{pro}}(t) = \sum_{k=t-T}^{t-1} x_r^t(k)\rho(r(t))\eta(g_{\mathrm{R}}(t)). \tag{4}$$

Problem statement: Our goal is to determine the optimal proactive download to minimize the time average expected cost for the system to deliver the user demand. Our optimization problem is written as $C_T^{\mathrm{pro}} =$

$$\begin{aligned} \min_{x_r^t(k)\forall r,t,k} \limsup_{t\to\infty} \quad & \frac{1}{t}\sum_{l=0}^{t-1} \mathbb{E}\left[C_{\mathrm{B}}\left[L_{\mathrm{B}}^{\mathrm{pro}}(l)\right] + C_{\mathrm{R}}\left[L_{\mathrm{R}}^{\mathrm{pro}}(l)\right]\right] \\ \text{subject to} \quad & x_f^k(l) \geq 0, \ \forall f, k, l, \\ & \sum_{k=t-T}^{t-1} x_r^t(k) \leq S, \ \forall r, t. \end{aligned} \tag{5}$$

The exact solution of (5) is intractably complex due to the infinite dimensionality of the problem. Instead, we aim to develop efficient proactive caching policy that can efficiently utilize statistical predictions and operate arbitrarily close to optimal as prediction window size grows. Fortunately, our analysis will also show that the performance of this policy, of finite complexity, converges considerably fast to the optimal as shown in simulations, thereby possesses close-to-optimal performance even for moderate values of the prediction window.

[1] This is an application-layer optimization where service provider needs to conduct measurements and define what the load is and what the cost is. For instance, the load could be the power consumption in Watt in a given time slot, and the cost is the money paid to supply this power.

IV. Lower Bound and Proposed Policy

We begin our design by investigating the limiting behavior of any proactive caching policy, which will inspire the dynamics of our developed policy. We characterize a fundamental lower bound on the minimum possible operational cost. Then we develop our asymptotically optimal policy which operates arbitrarily close to the bound as the proactive window size grows.

Theorem 1. *Under time-invariant demand and channel statistics model, the optimal proactive scheduling cost of* (5), *satisfies*

$$C_T^{\text{pro}} \geq C_b. \tag{6}$$

where, C_b is represented as follows

$$\begin{aligned} C_b = & \min_{\widetilde{x}_f(g_{\mathrm{B}},g_{\mathrm{R}},r)\forall r,g_{\mathrm{B}},g_{\mathrm{R}}} \sum_{g_{\mathrm{B}}\in G_{\mathrm{B}}} \sum_{g_{\mathrm{R}}\in G_{\mathrm{R}}} \sum_{r\in\mathcal{R}} \mathbb{P}(g_{\mathrm{B}})\mathbb{P}(g_{\mathrm{R}}) \\ & \mathbb{P}(r) \times C_B\Big[\big(S - \sum_{h_{\mathrm{B}}\in G_{\mathrm{B}}} \sum_{h_{\mathrm{R}}\in G_{\mathrm{R}}} \sum_{d\in\mathcal{R}} \mathbb{P}(h_{\mathrm{B}})\mathbb{P}(h_{\mathrm{R}})\mathbb{P}(d) \\ & \widetilde{x}_r(h_{\mathrm{B}},h_{\mathrm{R}},d)\big)\eta(g_{\mathrm{B}})\rho(r) + \sum_{f\subset\mathcal{F}} \widetilde{x}_f(g_{\mathrm{B}},g_{\mathrm{R}},r)\eta(g_{\mathrm{B,R}})\Big] \\ & + C_R\Big[\sum_{h_{\mathrm{B}}\in G_{\mathrm{B}}} \sum_{h_{\mathrm{R}}\in G_{\mathrm{R}}} \sum_{d\in\mathcal{R}} \mathbb{P}(h_{\mathrm{B}})\mathbb{P}(h_{\mathrm{R}})\mathbb{P}(d) \\ & \widetilde{x}_r(h_{\mathrm{B}},h_{\mathrm{R}},d)\eta(g_{\mathrm{R}})\Big] \\ & \textit{subject to } \; 0 \leq \widetilde{x}_f(g_{\mathrm{B}},g_{\mathrm{R}},r) \leq S \;\forall\; f,r,g_{\mathrm{B}},g_{\mathrm{R}}. \end{aligned} \tag{7}$$

Proof. Refer to Appendix A. □

The minimum of the above optimization problem exists and is unique. Existence follows since the objective function is convex; the composition in $\widetilde{x}_f(g_{\mathrm{B}},g_{\mathrm{R}},r)$ is linear $\forall f,r,g_{\mathrm{B}},g_{\mathrm{R}}$ and the sum of strictly convex functions is strictly convex, thus the constraint set is compact. In the objective of (7), the term $\sum_{h_{\mathrm{B}}\in G_{\mathrm{B}}} \sum_{h_{\mathrm{R}}\in G_{\mathrm{R}}} \sum_{d\in\mathcal{R}} \mathbb{P}(h_{\mathrm{B}})\ \mathbb{P}(h_{\mathrm{R}})\mathbb{P}(d)\widetilde{x}_r(h_{\mathrm{B}},h_{\mathrm{R}},d)$ corresponds to the average proactive service assigned to a request from the user before it is actually realized. The term $\widetilde{x}_f(g_{\mathrm{B}},g_{\mathrm{R}},r)$ is the total expected proactive service assigned to all possible requests from the user when f is the current requested file and the channel realizations g_{B} and g_{R}.

The theorem establishes that no proactive caching policy can achieve a lower cost than the non-trivial bound C_b. We note that the optimization of C_b is convex and yields a unique solution by the strict convexity of $C_d(x)$. Such optimization is numerically tractable and its solution can be numerically computed, e.g. through dual-based or interiorpoint methods.

Next, we develop a stationary policy π that asymptotically achieves the lower bound C_b.

Definition 1. *We consider a proactive policy π that observes the requested file f and channel realizations g_{B} for the link between the BS and the user and g_{R} for the channel between the RS and the user each time slot t. The policy assigns proactive controls* $x_r^k(t) = \dfrac{\widetilde{x}_f(g_{\mathrm{B}},g_{\mathrm{R}},r)}{T}, \forall f,k, 1\leq t\leq T.$

Policy π is a stationary policy that observes the current demand and channel realizations, solves the optimization problem in (7) and accordingly assigns proactive control value $x_r^k(t) = \frac{\widetilde{x}_f(g_{\mathrm{B}},g_{\mathrm{R}},r)}{T}$ for all potential requests that may be requested in the upcoming T slots. Then, we can develop the asymptotic optimal policy π as follows.

Theorem 2. *Denote the time average expected cost under the policy π by C_T^{π}. Then the policy π is asymptotically optimal in the sense that*

$$\limsup_{T\to\infty} \mid C_T^{\pi} - C_T^{\text{pro}} \mid = 0. \tag{8}$$

Proof. Refer to Appendix B. □

By using the strong law of large numbers, equal allocation of proactive service throughout the prediction window of size T, policy π achieves the global lower bound as $T\to\infty$. Having established the key characteristics of proactive scheduling under demand and channel uncertainties, we next move on to deeper insights on the system performance through numerical simulations.

V. Simulation Results

Throughout this section, we assume that the BS and the RS are aware of the statistics of both the user demand and the channel realizations, where the BS spends $S = 1$ data units for each request.

We assume that the end user experiences one of two possible channel realizations $\{g_{\mathrm{B}}^{(1)}, g_{\mathrm{B}}^{(2)}\}$ on the link between the BS and the user and also experiences one of two possible channel realizations $\{g_{\mathrm{R}}^{(1)}, g_{\mathrm{R}}^{(2)}\}$ on the link between the RS and the user with probabilities $\{\alpha_1,\alpha_2\}$ and $\{\beta_1,\beta_2\}$ respectively , where $\alpha_2 = 1-\alpha_1$ and $\beta_2 = 1-\beta_1$. We consider $g_{\mathrm{B}}^{(1)}$ and $g_{\mathrm{R}}^{(1)}$ to be the bad channel realizations while $g_{\mathrm{B}}^{(2)}$ and $g_{\mathrm{R}}^{(2)}$ to be the good channel realizations.

We assume that the cost function of demand from BS or RS are a polynomial function $C_d(x) = x^4$ for $d\in\{\mathrm{B},\mathrm{R}\}$. For the channel cost function, it is inversely proportional to the channel gain value, $\eta(x) = \dfrac{1}{x}$.

We consider the *cost reduction gain* as the main QoS performance metric. It provides a measure of the percentage change of the cost under the reactive model compared to the proactive model. Cost reduction gain is described as follows,

$$\gamma = \frac{\mid C^{\text{re}} - C_T^{\text{pro}} \mid}{C^{\text{re}}} \times 100\%. \tag{9}$$

A. Impact of number of files on the cost reduction gain

The impact of increasing the number of files in the system on the cost reduction gain γ is shown in Fig. 2 and Fig. 3. For this scenario, the channel realizations between the BS and the user are set to $g_{\mathrm{B}}^{(1)} = 0.1$ and $g_{\mathrm{B}}^{(2)} = 0.2$ with probability $\alpha_1 = \alpha_2 = 0.5$ and the channel realizations between the RS and the user are set to $g_{\mathrm{R}}^{(1)} = 0.3$, $g_{\mathrm{R}}^{(2)} = 0.4$ with probability $\beta_1 = \beta_2 = 0.5$ and for the channel between BS and RS is set to $g_{\mathrm{B,R}} = 1$.

In Fig. 2 and Fig. 3, we plot the cost reduction gain γ versus the number of files. The cost reduction gain decreases

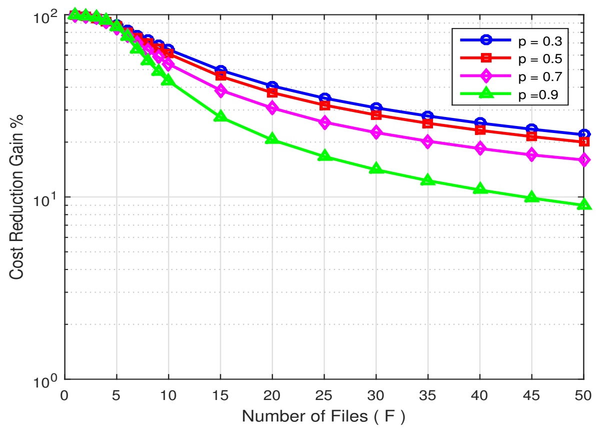

Fig. 2: The impact of number of files on the cost reduction gain under different probability of demand.

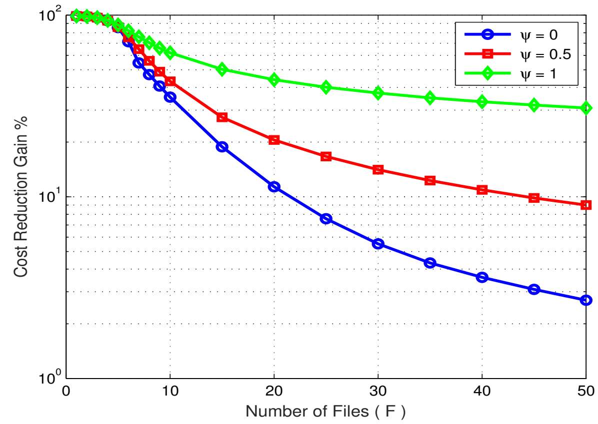

Fig. 3: The impact of the number of files on the cost reduction gain under different Zipf parameter.

with increasing the number of files under different demand probabilities. This is attributed to the observation that when the number of files increases, the uncertainty in the demand of the files increases which leads to a decrease in γ.

In Fig. 2, we fix the Zipf parameter to $\psi = 0.5$. By increasing the demand probability for the files, p, the cost reduction gain decreases. This is a result of increasing the load in the network which means less opportunities for proactive service, thus cost increases which means that cost reduction gain decreases.

In Fig. 3, we assume that the demand probability $p = 0.9$. When the Zipf parameter increases the cost reduction gain increases as well. This is due to the fact that when the Zipf parameter is equal to zero, the files popularity is uniform which will lead to more cost as uniform popularity means maximum uncertainty which leads to highest chances of inaccurate proactive service thus the BS in turn reduces its proactive operation. On the other hand, when the Zipf parameter tends to increase, the cost reduction gain increases. This is a result of higher certainty and more accurate proactive service, that leads to increasing the proactive operation of the BS. Thus, there will be more losses in the scenario where $\psi = 0$ and these losses decrease by increasing ψ. So the cost reduction gain ,γ, significantly increases when the Zipf parameter increases.

B. Impact of Zipf parameter on cost reduction gain

In this scenario, the number of files is fixed to $F = 10$. We set the channel realizations and their probabilities as follows, $g_{\mathrm{B}}^{(1)} = 0.1$ and $g_{\mathrm{B}}^{(2)} = 0.2$ with probability $\alpha_1 = \alpha_2 = 0.5$ and $g_{\mathrm{R}}^{(1)} = 0.3$, $g_{\mathrm{R}}^{(2)} = 0.4$ with probability $\beta_1 = \beta_2 = 0.5$. The channel between BS and RS is set to $g_{\mathrm{B,R}} = 1$.

Fig. 4 shows the effect of Zipf parameter on the cost reduction gain as explained in Fig. 3 where the cost reduction gain increases when the Zipf parameter increases. In addition, it shows how the probability of demand affects γ, when the probability of demand increases the cost reduction gain decreases as illustrated in Fig. 2. The effect of the probability

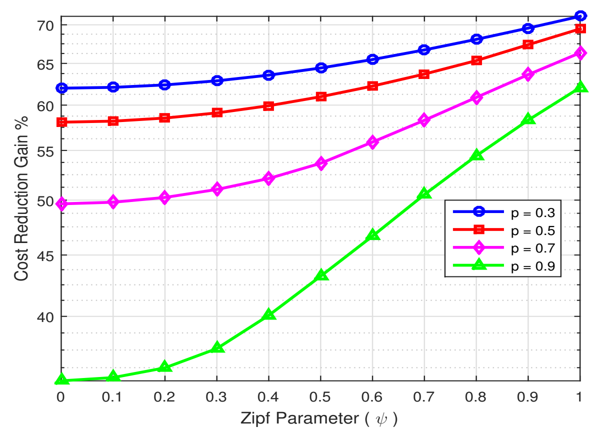

Fig. 4: Cost reduction gain versus the Zipf parameters under different probability of demand.

of demand on γ decreases by increasing the Zipf parameter as the file popularity goes from uniform to biased popularity.

C. Comparison with caching at end user

We compare our system performance with the caching at end user scheme which is primarily considered in [10]. In this simulation setup, the channel realizations between the BS and the user are set to $g_{\mathrm{B}}^{(1)} = 0.1$ and $g_{\mathrm{B}}^{(2)} = 0.2$ with probability $\alpha_1 = \alpha_2 = 0.5$ and the channel realizations between the RS and the user are set to $g_{\mathrm{R}}^{(1)} = 0.1$, $g_{\mathrm{R}}^{(2)} = 0.2$ with probability $\beta_1 = \beta_2 = 0.5$ and for the channel between BS and RS is set to $g_{\mathrm{B,R}} = 1$. The Zipf parameter is set to $\psi = 0.5$. We assume that the number of files in our system to be $F = 1$ in order to be able to compare it with the system model in [10], which assumes a single data item service.

As shown in Fig. 5, our system performance outperforms that in [10] when the probability of demand increases. This is the effect of the RS. In [10] when the demand increases the link between the BS and the end user needs to support more data. While in our model, when the demand increases,

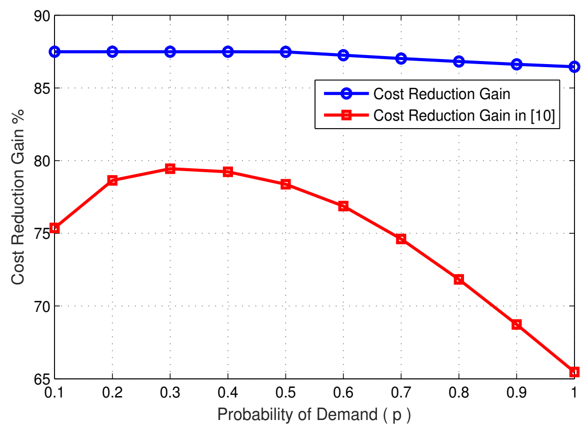

Fig. 5: Cost reduction gain versus the probability of demand for the user.

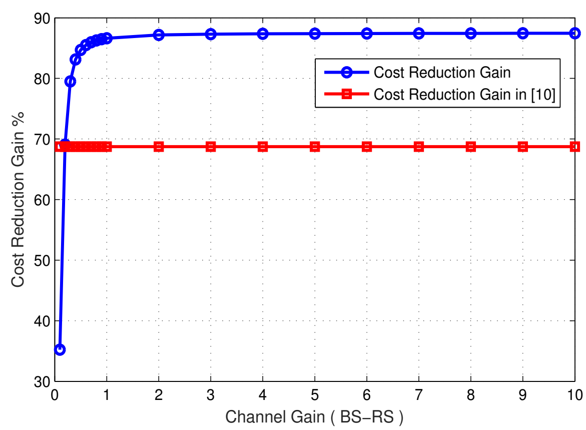

Fig. 6: Cost reduction gain versus the BS- RS channel gain.

the content can reach the user via two links, BS user link and RS user link. The two links in our model are better than the single link in [10] where it increase the opportunity of serving the user because this exploits channel diversity. As shown in Fig. 5 at low demand probability, the cost reduction gain, γ, increases. While at high demand probability, γ decreases for the reasons mentioned above.

In Fig. 6, we use the Zipf parameter $\psi = 0.5$ and we fix the probability of demand to be $p = 0.9$. We vary the quality of the link between the BS and the RS. We assume that the statistics for the channel between the RS and the end user are the same as those for the link between the BS and the user, $g_{\mathrm{B}}^{(1)} = g_{\mathrm{R}}^{(1)} = 0.1$ and $g_{\mathrm{B}}^{(2)} = g_{\mathrm{R}}^{(2)} = 0.2$ with probabilities $\alpha_1 = \beta_1 = 0.5$.

The cost reduction gain in [10] is constant for all BS- RS channel gains, as in this model there is no RS thus the change in this channel gain will not affect the cost reduction gain. However, for our system when the channel gain between the BS and the RS increases the cost reduction gain γ increases as the cost for transmission will decrease as the channel gain is better. The cost reduction gain begins to stay constant at $g_{\mathrm{B,R}} = 5$, which means that at this channel γ reaches its maximum, i.e. the cost decreases until it reaches its minimum level to serve the demand of the user.

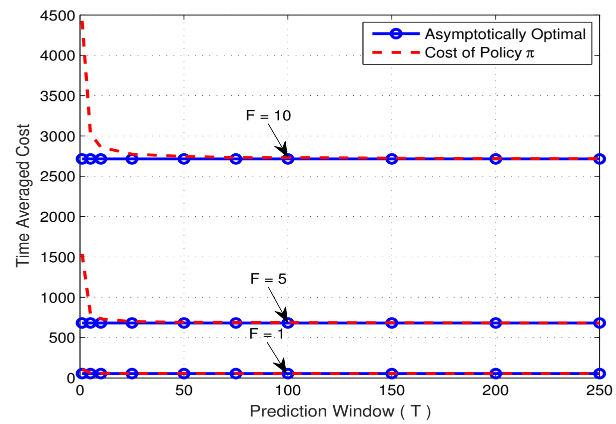

Fig. 7: Impact of proactive window size on the average cost.

D. Impact of prediction window size on the expected cost

In Fig. 7, we plot the time average cost under policy π against the prediction window size T. In this scenario, the channel realizations between the BS and the user are set to $g_{\mathrm{B}}^{(1)} = 0.1$ and $g_{\mathrm{B}}^{(2)} = 0.2$ with probability $\alpha_1 = \alpha_2 = 0.5$ and the channel realizations between the RS and the user are set to $g_{\mathrm{R}}^{(1)} = 0.3$, $g_{\mathrm{R}}^{(2)} = 0.4$ with probability $\beta_1 = \beta_2 = 0.5$ and for the channel between BS and RS is set to $g_{\mathrm{B,R}} = 1$. We assume the Zipf parameter to be $\psi = 0.5$ where the probability for the user demand is $p = 0.9$. As shown in Fig. 7, policy π converges to achieve the lower bound C_b rapidly when the prediction window size grows. We see the impact of the prediction window size for different number of files in the system $F = 1, 5, 10$.

For the scenario with $F = 1$, $F = 5$ and $F = 10$ the policy converges to the lower bound at $T = 25$, $T = 60$ and $T = 150$, respectively. As suggested by intuition, when the number of files increases the uncertainty and randomness increases thus the system needs larger service window to realize sufficient numbers of randomness elements and achieve the law of large numbers and converge to the lower bound.

VI. Conclusion

We proposed and investigated the performance of dynamic proactive caching in relay networks. We studied the impact of demand and channel uncertainties on the design of a proactive scheduler, where an intermediate RS caches content for potential future use by end users under time-invariant demand over multiple data items and channel statistics models. We have established fundamental lower bound on the achievable cost through proactive scheduling, we developed asymptotically optimal policy that attains the lower bound rapidly as the proactive scheduling window size increases. Our numerical results demonstrate the performance gain of proactive caching at the RS, compared to a reactive content retrieval baseline scheme as well as a proactive caching at the end user introduced earlier in [10].

Appendix A
Proof of Theorem 1

The objective of our optimization problem is

$$C_T^{\text{pro}} = \limsup_{t\to\infty} \frac{1}{t}\sum_{l=0}^{t-1} \mathbb{E}\left[C_{\text{B}}\left[L_{\text{B}}^{\text{pro}}(l)\right] + C_{\text{R}}\left[L_{\text{R}}^{\text{pro}}(l)\right]\right]. \quad (10)$$

By conditioning on all random variables: requested files and all possible channel realizations, we can write C_T^{pro} as follows,

$$C_T^{\text{pro}} = \limsup_{t\to\infty} \frac{1}{t}\sum_{l=0}^{t-1} \sum_{g_{\text{B}}\in G_{\text{B}}} \sum_{g_{\text{B}}\in G_{\text{R}}} \sum_{r\in\mathcal{R}} \mathbb{P}(g_{\text{B}}(l) = g_{\text{B}}, g_{\text{R}}(l) = g_{\text{R}}, r(l) = r) \times \mathbb{E}\left[C_{\text{B}}\left[L_{\text{B}}^{\text{pro}}(l)\right] + C_{\text{R}}\left[L_{\text{R}}^{\text{pro}}(l)\right] \mid g_{\text{B}}(l) = g_{\text{B}}, g_{\text{R}}(l) = g_{\text{R}}, r(l) = r\right]. \quad (11)$$

Due to statistical independence, $\mathbb{P}(g_{\text{B}}(l) = g_{\text{B}}, g_{\text{R}}(l) = g_{\text{R}}, r(l) = r) = \mathbb{P}(g_{\text{B}}(l) = g_{\text{B}})$ $\mathbb{P}(g_{\text{R}}(l) = g_{\text{R}})\mathbb{P}(r(l) = r)$ and (11) can be written as

$$C_T^{\text{pro}} = \limsup_{t\to\infty} \frac{1}{t}\sum_{l=0}^{t-1} \sum_{g_{\text{B}}\in G_{\text{B}}} \sum_{g_{\text{R}}\in G_{\text{R}}} \sum_{r\in\mathcal{R}} \mathbb{P}(g_{\text{B}}(l) = g_{\text{B}}) \mathbb{P}(g_{\text{R}}(l) = g_{\text{R}})\mathbb{P}(r(l) = r) \times \mathbb{E}\left[C_{\text{B}}\left[L_{\text{B}}^{\text{pro}}(l)\right] + C_{\text{R}}\left[L_{\text{R}}^{\text{pro}}(l)\right] \mid g_{\text{B}}(l) = g_{\text{B}}, g_{\text{R}}(l) = g_{\text{R}}, r(l) = r\right]. \quad (12)$$

By using (3) and (4) and substituting in (12), we can write the cost as follows,

$$\begin{aligned} C_T^{\text{pro}} = \; & \limsup_{t\to\infty} \frac{1}{t}\sum_{l=0}^{t-1} \sum_{g_{\text{B}}\in G_{\text{B}}} \sum_{g_{\text{R}}\in G_{\text{R}}} \sum_{r\in\mathcal{R}} \mathbb{P}(g_{\text{B}}(l) = g_{\text{B}}) \\ & \mathbb{P}(g_{\text{R}}(l) = g_{\text{R}})\mathbb{P}(r(l) = r) \times \Big(\mathbb{E}\Big[C_B\big[\\ & \big(S - \sum_{k=l-T}^{l-1} x_r^l(k)\big)\eta(g_{\text{B}}(l))\rho(r(l)) + \sum_{f\subset\mathcal{F}} \sum_{k=l+1}^{l+T} \\ & x_f^k(l)\eta(g_{\text{B,R}}) \,|\, g_{\text{B}}(l) = g_{\text{B}}, g_{\text{R}}(l) = g_{\text{R}}, r(l) = r\big]\Big] + \\ & \mathbb{E}\Big[C_R\big[\sum_{k=l-T}^{l-1} x_r^l(k)\eta(g_{\text{R}}(l))\rho(r(l)) \,|\, g_{\text{B}}(l) = g_{\text{B}}, \\ & g_{\text{R}}(l) = g_{\text{R}}, r(l) = r\big]\Big]\Big). \end{aligned} \quad (13)$$

Since channel realizations for both the BS and the RS as well as the requested files are i.i.d., we have

$$\begin{aligned} C_T^{\text{pro}} = \; & \limsup_{t\to\infty} \frac{1}{t}\sum_{l=0}^{t-1} \sum_{g_{\text{B}}\in G_{\text{B}}} \sum_{g_{\text{R}}\in G_{\text{R}}} \sum_{r\in\mathcal{R}} \mathbb{P}(g_{\text{B}})\mathbb{P}(g_{\text{R}})\mathbb{P}(r) \\ & \times \mathbb{E}\Big[C_B\big[\big(S - \sum_{k=l-T}^{l-1} x_r^l(k)\big)\eta(g_{\text{B}})\rho(r) \\ & + \sum_{f\subset\mathcal{F}} \sum_{k=l+1}^{l+T} x_f^k(l)\eta(g_{\text{B,R}}) \mid g_{\text{B}}, g_{\text{R}}, r\big]\Big] \\ & + \mathbb{E}\Big[C_R\big[\sum_{k=l-T}^{l-1} x_r^l(k)\eta(g_{\text{R}})\rho(r) \mid g_{\text{B}}, g_{\text{R}}, r\big]\Big]. \end{aligned} \quad (14)$$

We apply Jensen's inequality in (14), as $C_d(x)$ is assumed to be strictly convex. Both the requested file $r(l)$ and the channel realizations $g_{\text{B}}(l)$ and $g_{\text{R}}(l)$ are independent of $\sum_{k=l-T}^{l-1} x_r^l(k)$, as the current requested file and channel realizations will not affect the past services. However, they will affect the future service as the future services are dependent on the requested file and the channel realizations in the current time slot. We can then write the bound as follows,

$$\begin{aligned} C_T^{\text{pro}} \geq \; & \limsup_{t\to\infty} \frac{1}{t}\sum_{l=0}^{t-1} \sum_{g_{\text{B}}\in G_{\text{B}}} \sum_{g_{\text{R}}\in G_{\text{R}}} \sum_{r\in\mathcal{R}} \mathbb{P}(g_{\text{B}})\mathbb{P}(g_{\text{R}})\mathbb{P}(r) \\ & \times C_B\Big[\big(S - \mathbb{E}\big[\sum_{k=l-T}^{l-1} x_r^l(k) \mid g_{\text{B}}, g_{\text{R}}, r\big]\eta(g_{\text{B}})\rho(r) \\ & + \mathbb{E}\big[\sum_{f\subset\mathcal{F}} \sum_{k=l+1}^{l+T} x_f^k(l) \mid g_{\text{B}}, g_{\text{R}}, r\big]\eta(g_{\text{B,R}})\Big] \\ & + C_R\Big[\mathbb{E}\big[\sum_{k=l-T}^{l-1} x_r^l(k)\eta(g_{\text{R}})\rho(r)\Big]. \end{aligned} \quad (15)$$

Since $\frac{1}{t}\sum_{l=0}^{l-1} 1 = 1$, $\sum_{g_{\text{B}}\in G_{\text{B}}} \mathbb{P}(g_{\text{B}}) = 1$ and $\sum_{g_{\text{B}}\in G_{\text{B}}} \mathbb{P}(g_{\text{B}}) = 1$ we can apply Jensen's inequality again and by using $\limsup_{t\to\infty}(-f(t)) = -\liminf_{t\to\infty}(f(t))$, (15) can be written as,

$$\begin{aligned} C_T^{\text{pro}} \geq \; & \sum_{g_{\text{B}}\in G_{\text{B}}} \sum_{g_{\text{R}}\in G_{\text{R}}} \sum_{r\in\mathcal{R}} \mathbb{P}(g_{\text{B}})\mathbb{P}(g_{\text{R}})\mathbb{P}(r) \times C_B\Big[\\ & \big(S - \liminf_{t\to\infty} \frac{1}{t}\sum_{l=0}^{t-1} \mathbb{E}\big[\sum_{k=l-T}^{l-1} x_r^l(k) \mid g_{\text{B}}, g_{\text{R}}, r\big]\big) \\ & \eta(g_{\text{B}})\rho(r) + \limsup_{t\to\infty} \frac{1}{t}\sum_{l=0}^{t-1} \mathbb{E}\big[\sum_{f\subset\mathcal{F}} \sum_{k=l+1}^{l+T} x_f^k(l) \\ & \mid g_{\text{B}}, g_{\text{R}}, r\big]\eta(g_{\text{B,R}})\Big] + C_R\Big[\limsup_{t\to\infty} \frac{1}{t}\sum_{l=0}^{t-1} \\ & \mathbb{E}\big[\textstyle\sum_{k=l-T}^{l-1} x_r^l(k)\big]\eta(g_{\text{R}})\rho(r)\Big]. \end{aligned} \quad (16)$$

By replacing $\limsup_{t\to\infty}$ by $\liminf_{t\to\infty}$ as $C_{\text{B}}(x)$ and $C_{\text{R}}(x)$ are monotonically increasing in x this yields,

$$\begin{aligned} C_T^{\text{pro}} \geq \; & \sum_{g_{\text{B}}\in G_{\text{B}}} \sum_{g_{\text{R}}\in G_{\text{R}}} \sum_{r\in\mathcal{R}} \mathbb{P}(g_{\text{B}})\mathbb{P}(g_{\text{R}})\mathbb{P}(r) \times C_B\Big[\\ & \big(S - \liminf_{t\to\infty} \frac{1}{t}\sum_{l=0}^{t-1} \sum_{k=l-T}^{l-1} \mathbb{E}\big[x_r^l(k) \mid g_{\text{B}}, g_{\text{R}}, r\big]\big) \\ & \eta(g_{\text{B}})\rho(r) + \liminf_{t\to\infty} \frac{1}{t}\sum_{l=0}^{t-1} \sum_{f\subset\mathcal{F}} \sum_{k=l+1}^{l+T} \mathbb{E}\big[x_f^k(l) \\ & \mid g_{\text{B}}, g_{\text{R}}, r\big]\eta(g_{\text{B,R}})\Big] + C_R\Big[\liminf_{t\to\infty} \frac{1}{t}\sum_{l=0}^{t-1} \sum_{k=l-T}^{l-1} \\ & \mathbb{E}\big[x_r^l(k)\big]\eta(g_{\text{R}})\rho(r)\Big]. \end{aligned} \quad (17)$$

By letting $\widetilde{x}_f(g_{\text{B}}, g_{\text{R}}, r) = \liminf_{t\to\infty} \frac{1}{t}\sum_{l=0}^{t-1} \sum_{k=l+1}^{l+T} \mathbb{E}\big[x_f^k(l) \mid g_{\text{B}}, g_{\text{R}}, r\big]$, we can write $\liminf_{t\to\infty} \frac{1}{t}\sum_{l=0}^{t-1} \sum_{k=l-T}^{l-1} \mathbb{E}\big[x_r^l(k)\big]$ as follows,

$$\liminf_{t\to\infty} \; \frac{1}{t}\sum_{l=0}^{t-1} \sum_{k=l-T}^{l-1} \mathbb{E}\big[x_r^l(k)\big]$$

$$= \liminf_{t\to\infty} \frac{1}{t} \sum_{l=0}^{t-1} \sum_{h_{\mathrm{B}}\in G_{\mathrm{B}}} \sum_{h_{\mathrm{R}}\in G_{\mathrm{R}}} \sum_{d\in\mathcal{R}} \mathbb{P}(h_{\mathrm{B}})\mathbb{P}(h_{\mathrm{R}})\mathbb{P}(d) \sum_{k=l-T}^{l-1} \mathbb{E}\left[x_r^l(k) \mid h_{\mathrm{B}}, h_{\mathrm{R}}, d\right]$$
$$= \sum_{h_{\mathrm{B}}\in G_{\mathrm{B}}} \sum_{h_{\mathrm{R}}\in G_{\mathrm{R}}} \sum_{d\in\mathcal{R}} \mathbb{P}(h_{\mathrm{B}})\mathbb{P}(h_{\mathrm{R}})\mathbb{P}(d)\widetilde{x}_r(h_{\mathrm{B}}, h_{\mathrm{R}}, d). \tag{18}$$

Thus, we can write our bound as follows

$$C_T^{\mathrm{pro}} \geq \sum_{g_{\mathrm{B}}\in G_{\mathrm{B}}} \sum_{g_{\mathrm{R}}\in G_{\mathrm{R}}} \sum_{r\in\mathcal{R}} \mathbb{P}(g_{\mathrm{B}})\mathbb{P}(g_{\mathrm{R}})\mathbb{P}(r) \times C_B\Big[\big(S - \sum_{h_{\mathrm{B}}\in G_{\mathrm{B}}} \sum_{h_{\mathrm{R}}\in G_{\mathrm{R}}} \sum_{d\in\mathcal{R}} \mathbb{P}(h_{\mathrm{B}})\mathbb{P}(h_{\mathrm{R}})\mathbb{P}(d) \widetilde{x}_r(h_{\mathrm{B}}, h_{\mathrm{R}}, d)\big)\eta(g_{\mathrm{B}})\rho(r) + \sum_{f\subset\mathcal{F}} \widetilde{x}_f(g_{\mathrm{B}}, g_{\mathrm{R}}, r) \eta(g_{\mathrm{B,R}})\Big] + C_R\Big[\sum_{h_{\mathrm{B}}\in G_{\mathrm{B}}} \sum_{h_{\mathrm{R}}\in G_{\mathrm{R}}} \sum_{d\in\mathcal{R}} \mathbb{P}(h_{\mathrm{B}})\mathbb{P}(h_{\mathrm{R}})\mathbb{P}(d) \widetilde{x}_r(h_{\mathrm{B}}, h_{\mathrm{R}}, d)\eta(g_{\mathrm{R}})\rho(r)\Big]. \tag{19}$$

The constraints of the optimization problem in (5) imply that $0 \leq \widetilde{x}_f(g_{\mathrm{B}}, g_{\mathrm{R}}, r) \leq S \,\forall\, r, d, g_{\mathrm{B}}, g_{\mathrm{R}}$. By minimizing the right-hand-side of the previous equation over all feasible choices of $\widetilde{x}_r(g_{\mathrm{B}}, g_{\mathrm{R}}, d)$, our theorem is proved.

Appendix B
Proof of theorem 2

It suffices to prove that $\limsup_{T\to\infty} C_T^\pi = \liminf_{T\to\infty} C_T^\pi$. We start by $\limsup_{T\to\infty} C_T^{\mathrm{pro}}$. Since the policy C_π^T is a stationary policy that depends only on the current demand realization, the cost under this policy can be written as follows,

$$C_\pi^T = \sum_{g_{\mathrm{B}}\in G_{\mathrm{B}}} \sum_{g_{\mathrm{R}}\in G_{\mathrm{R}}} \sum_{r\in\mathcal{R}} \mathbb{P}(g_{\mathrm{B}}(t) = g_{\mathrm{B}})\mathbb{P}(g_{\mathrm{R}}(t) = g_{\mathrm{R}}) \mathbb{P}(r(t) = r) \times \mathbb{E}\Big[C_B\big[\big(S - \sum_{k=t-T}^{t-1} x_r^t(k)\big)\eta(g_{\mathrm{B}}(t)) \rho(r(t)) + \sum_{f\subset\mathcal{F}} \sum_{k=t+1}^{t+T} x_f^k(t)\eta(g_{\mathrm{B,R}})\big] \mid g_{\mathrm{B}}, g_{\mathrm{R}}, r\Big] + \mathbb{E}\Big[C_R\big[\sum_{k=t-T}^{t-1} x_r^t(k)\eta(g_{\mathrm{R}}(t))\rho(r(t))\big] \mid g_{\mathrm{B}}, g_{\mathrm{R}}, r\Big]. \tag{20}$$

The sum $\sum_{k=t-T}^{t-1} x_r^t(k)$ is independent of both demand and channel realizations. Define a random variable $Q(h_{\mathrm{B}}, h_{\mathrm{R}}, d)$ which counts the number of occurrences of the joint realization of r,g_{B} and g_{R}. Thus,

$$\sum_{k=t-T}^{t-1} x_r^t(k) = \sum_{h_{\mathrm{B}}\in G_{\mathrm{B}}} \sum_{h_{\mathrm{R}}\in G_{\mathrm{R}}} \sum_{d\in\mathcal{R}} \frac{\widetilde{x}_r(h_{\mathrm{B}}, h_{\mathrm{R}}, d)Q(h_{\mathrm{B}}, h_{\mathrm{R}}, d)}{T}. \tag{21}$$

The strong law of large numbers implies that (21) can be written as follows,

$$\limsup_{T\to\infty} \sum_{h_{\mathrm{B}}\in G_{\mathrm{B}}} \sum_{h_{\mathrm{R}}\in G_{\mathrm{R}}} \sum_{d\in\mathcal{R}} \frac{\widetilde{x}_r(h_{\mathrm{B}}, h_{\mathrm{R}}, d)Q(g_{\mathrm{B}}, g_{\mathrm{R}}, d)}{T} = \mathbb{P}(h_{\mathrm{B}})\mathbb{P}(h_{\mathrm{R}})\mathbb{P}(d)\widetilde{x}_r(h_{\mathrm{B}}, h_{\mathrm{R}}, d), \qquad w.p.1. \tag{22}$$

From here, we can say that the cost of demand under the policy achieve our bound when prediction window increases.

$$\limsup_{T\to\infty} C_T^\pi = \sum_{g_{\mathrm{B}}\in G_{\mathrm{B}}} \sum_{g_{\mathrm{R}}\in G_{\mathrm{R}}} \sum_{r\in\mathcal{R}} \mathbb{P}(g_{\mathrm{B}})\mathbb{P}(g_{\mathrm{R}})\mathbb{P}(r)C_B\Big[\big(S - \sum_{h_{\mathrm{B}}\in G_{\mathrm{B}}} \sum_{h_{\mathrm{R}}\in G_{\mathrm{R}}} \sum_{d\in\mathcal{R}} \mathbb{P}(h_{\mathrm{B}})\mathbb{P}(h_{\mathrm{R}})\mathbb{P}(d) \widetilde{x}_r(h_{\mathrm{B}}, h_{\mathrm{R}}, d)\big)\eta(g_{\mathrm{B}})\rho(r) + C_R\Big[\sum_{h_{\mathrm{B}}\in G_{\mathrm{B}}} \sum_{h_{\mathrm{R}}\in G_{\mathrm{R}}} \sum_{d\in\mathcal{R}} \mathbb{P}(h_{\mathrm{B}})\mathbb{P}(h_{\mathrm{R}})\mathbb{P}(d)\widetilde{x}_r(h_{\mathrm{B}}, h_{\mathrm{R}}, d)\eta(g_{\mathrm{R}})\rho(r)\Big]. \tag{23}$$

By noting that the right-hand-side of (23) is identical to C_b, then $\limsup_{T\to\infty} C_T^\pi \leq \liminf_{T\to\infty} C_T^\pi$. By the definition of C_T^π it follows that $\limsup_{T\to\infty} C_T^\pi = \liminf_{T\to\infty} C_T^\pi$.

References

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

[2] M. Feknous T. Houdoin, B. Le Guyader, J. De Biasio, A. Gravey, and J.A. Torrijos Gijon, "Internet traffic analysis: A case study from two major European operators," in *2014 IEEE Symposium on Computers and Communication (ISCC)* , vol., no., pp.1,7, 23-26 June 2014.

[3] "Ericsson mobility report," http://www.ericsson.com/res/docs/2015/ericsson-mobility-report-june-2015.pdf, June 2015.

[4] C. Song, Z. Qu, N. Blumm, A. Barabas, "Limits of Predictability in Human Mobility", Science, vol. 327, no. 5968, pp. 1018-1021, Feb. 2010.

[5] R. Di Taranto, L. S. Muppirisetty, R. Raulefs, D. Slock, T. Svensson, and H. Wymeersch, "Location-aware communications for 5G networks" *IEEE Signal Processing Magazine* , vol. 31, no. 6, pp. 102112, Nov. 2014.

[6] M. Malmirchegini and Y. Mostofi, "On the spatial predictability of communication channels," *IEEE Transactions on Wireless Communications*, vol. 11, no. 3, pp. 964978, 2012.

[7] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," IEEE Communications Magazine, vol. 52, no. 2, pp. 7480, 2014.

[8] J. Tadrous, A. Eryilmaz, and H. El Gamal, "Proactive resource allocation: Harnessing the diversity and multicast gains" in *IEEE Transactions on Information Theory*, vol. 59, no. 8, pp. 48334854, 2013.

[9] J. Tadrous and A. Eryilmaz, "On optimal proactive caching for mobile networks with demand uncertainties," in *IEEE/ACM Transactions on Networking*, vol. 24, no. 5, pp. 2715 - 2727, Sep. 2015.

[10] L. S.Muppirisetty, J. Tadrous, A. Eryilmaz, and H. Wymeersch, " On proactive caching with demand and channel uncertainties," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)* , vol., no., pp. 1174-1181, Sep. 2015.

[11] E. Batu, M. Bennis, and M. Debbah, " Social and spatial proactive caching for mobile data offloading," in *2014 IEEE International Conference on Communication Workshops (ICC)* , vol., no., pp. 581-586, June 2014.

[12] E. Batu, M. Bennis, and M. Debbah, " Living on the edge: The role of proactive caching in 5G wireless networks," in *IEEE Communications Magazine* , vol. 52, no. 8, pp. 82-89, Aug. 2014.

[13] J. Tadrous, and A. Eryilmaz, " On Optimal Proactive Caching for Mobile Networks with Demand Uncertainties," in *IEEE/ACM Transactions on Networking* , vol. 24, no. 5, pp. 2715-2727, Oct. 2016.

[14] E. Batu, M. Bennis, and M. Debbah, " Think before reacting: Proactive caching in 5G small cell networks," in *1st KuVS workshop on Anticipatory Networks* , vol., no., pp., Sep. 2014.

[15] E. Batu, M. Bennis, and M. Debbah, " A transfer learning approach for cache-enabled wireless networks," *in 2015 13th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, vol., no., pp. 161-166, May 2015.

[16] E. Batu, M. Bennis, and M. Debbah, " Edge Caching for Coverage and Capacity-aided Heterogeneous Networks," in *arXiv preprint arXiv:1605.00319*, vol., no., pp., May 2016.

[17] R. Hassan, A. Mohamed, J. Tadrous, M. Nafie, T. ElBatt and F. Digham, " Dynamic Proactive Caching for Multiusers in Relay Networks," *to be submitted.*