Opportunistic Alignment of Advertisement Delivery with Cellular Basestation Overloads

Ravi Kokku†, Rajesh Mahindra†, Sampath Rangarajan†, Honghai Zhang†
†NEC Laboratories America, Inc., Princeton, USA
‡Current Affiliation: IBM Research, Bangalore, India
ravi.kokku@acm.org, {rajesh, sampath, honghai}@nec-labs.com

Abstract
This paper is motivated by two observations: (1) cellular network operators are actively exploring advertisement delivery as a new means of revenue generation, and (2) cellular basestations perceive intermittent overloads at different times of day. Bringing the two observations together, we design and implement Opal, a novel system for opportunistically aligning advertisement delivery with basestation overload. Such alignment improves the overall perception of network availability for users. To achieve the alignment systematically, Opal builds on an analytical framework for tunable unavailability of network service to users during overload. At the same time, if the network is not overloaded enough during a certain period, Opal schedules enough advertisements to satisfy the advertisement delivery contracts. Opal minimizes the amount of state to be maintained to play advertisements to users uniformly and also maximizes the number of viewers for each advertisement. We implement a prototype of Opal on a Picochip based WiMAX testbed, and demonstrate its efficacy using simulations, analysis and prototype evaluation.

Categories and Subject Descriptors
C.2.1 [Computer-Communication Networks]: Network Architecture and Design—Wireless Communication

General Terms
Design, Performance

Keywords
Cellular Networks, Overload Management, Advertisement Scheduling

1. INTRODUCTION
Recent years have witnessed an unprecedented increase in data traffic on cellular and broadband wireless networks. While forecasts indicate a 130% increase in mobile data traffic in 2011, and an estimated average compound annual growth rate (CAGR) of 92% between 2010-2015 [5], technology improvements contribute to only 55% increase in throughput per year [4]. Consequently, Mobile Network Operators (MNOs) have to increase network provisioning to keep up with the traffic requirements, which significantly increases the costs (see Figure 1 for a qualitative representation of the trends). Further, the requirements fluctuate with time, thereby making worst case network provisioning significantly more expensive than average case. At the same time, as we progress from voice-dominated to data-dominated mobile Internet, the revenue generated per byte of data (and hence per unit wireless resource) reduces significantly. These trends indicate a diverging gap between an operator’s revenues and costs in the opposite direction than desired.

Several techniques attempt to reverse this divergence without compromising on the user-perceived quality of experience, and make cellular network operations profitable in the data-dominated mobile Internet. Examples of such techniques include overload management methodologies (e.g., congestion pricing [34] and admission control [22]), novel architectures such as network virtualization [39] and femtocells [28], and new revenue generators such as enhanced video delivery [44], advertisement delivery [16], etc. Along these lines, this paper explores a novel way of combining overload management (a cost reducing approach) with advertisement delivery (a revenue generating approach). Firstly, as networks get dominated by non-elastic traffic such as video, overload management using admission control will be imminent, thereby exposing network unavailability to users. A forecast by Cisco indicates that 66% of mobile traffic...
will be video by the year 2015 [5]. Secondly, as mobile advertising is rapidly growing, mobile network operators are increasingly interested in leveraging its huge revenue potential, and are also better equipped than any other entity in reaching the users. Advertising, however, would lead to occasional unavailability of network service to users.

In this paper, we design and implement Opal—a system for opportunistically aligning advertisement delivery with basestation overload periods, which reduces the total number of times network service is unavailable to users. The design of Opal is influenced by two observations: (1) Advertisement contracts involve delivery requirements of a much longer timescale than basestation load fluctuations; i.e. it may be sufficient to show an advertisement to a user at some point of time (any time) during a day; whereas, basestation overloads happen at multiple and potentially predictable times during a day, and (2) Advertisement delivery can take much lower bandwidth per user than the user’s demand, and hence can be used during basestation overloads for load shedding; users who cannot be admitted will at least view advertisements instead of being shut-off.

Leveraging the two observations, Opal explicitly creates opportunities to align advertisement delivery to certain users during basestation overloads using (1) a soft admission control approach with long-term fairness, and (2) a tunable unavailability framework that lets an operator strike a trade-off between the amount of time and the number of times a user is interrupted for advertisements. Long-term fairness attempts to ensure that all users see similar number of advertisements. Consequently, by effectively serving advertisements to selected users, Opal ameliorates the effect of network unavailability during overload, and improves their overall quality of experience. If the basestations do not get overloaded enough within a certain duration, Opal still schedules advertisements sufficient number of times to meet advertisement contracts. Opal employs situation-based randomization for advertisement delivery to minimize per-user per-advertisement state maintenance, which works well even for dynamic working sets of users and advertisements.

We evaluate Opal with simulations and prototype implementation on a mobile WiMAX (802.16e) network testbed containing the Picochip WiMAX ASN (Access Service Network) gateway that runs Opal, the Picochip WiMAX base station [13], and Beceem-chipset-based clients [2]. Our prototype evaluation with YouTube [1] workloads shows that Opal selects a subset of users to receive advertisements (i.e. for load shedding) during overloads, such that the rest of the users observe 80-100% reduction in number of YouTube video stalls. Similar experiment with video over UDP shows an improvement in PSNR of up to 20dB for user videos.

Our analysis shows that for masking off a persistent overload of 20% for a duration of 10 minutes at a basestation, it is sufficient to show 3 advertisements to each user of about 35 second duration. Opal is designed as a gateway level solution with minimum dependence on the specific broadband wireless technology; hence, it is easily adaptable on multiple wireless access networks such as WiMAX2, LTE and LTE-Advanced.

To the best of our knowledge, this is the first work that demonstrates that advertisement delivery can be aligned with basestation overloads to minimize the effect of network service unavailability on users. This paper makes two novel technical contributions:

1. We present an analytical framework for tunable unavailability of network service to users. The framework captures an interesting tradeoff between the number of times a user is interrupted for seeing advertisements and the amount of time at a stretch advertisements are shown.

2. We design and implement a simple yet effective situation-based approach to schedule advertisements to users such that (a) each advertisement is viewed by maximum number of users, and (b) each user sees similar number of advertisements. The solution does not require per-user-per-advertisement state maintenance.

The rest of the paper is organized as follows. Section 2 formulates the problem by discussing the design considerations. Section 3 presents the design of Opal. Section 4 describes our prototype, and Sections 5 and 6 present the evaluation. Section 7 identifies limitations and discusses future work. Section 8 discusses related work to place Opal in context. Section 9 concludes.

2. PROBLEM FORMULATION

In this section, we expand on the two aspects that form the basis for Opal—cellular basestation overloads, and operator-enabled advertising.

2.1 Basestation Overload in Perspective

Overload is a condition in which the users’ aggregate guaranteed bandwidth requirement exceeds the basestation capacity, thereby leading to reduced effective bandwidth per user. Depending on the network provisioning (i.e., base-station density in a unit geographical area), this condition can happen several times in wireless networks, especially during certain times of day. One way to handle overload is provisioning network resources to handle worst-case requirements to avoid all overloads; however, worst-case provisioning can be prohibitively expensive. Alternately, overload management can be done by controlled load-shedding (i.e., either reducing per-user bandwidth or reducing the number of active users) based on the type of traffic.

Studies suggest that users are increasingly accessing video content over mobile networks [5]. For non-elastic traffic such as video, several past proposals suggest the use of admission control [51] during overload. Admission control becomes necessary for non-elastic traffic since throughput below a certain threshold for a user can cause unacceptable degradation of user experience [51]. Note that this is true irrespective of whether the payload is streaming video (over UDP) or stored video (carried over UDP or TCP). With streaming video over UDP, the effect is perceived as blocky video frames. Whereas with stored video, the effect is perceived as intermittent video stalls to enable buffering; during the stall, the user sees a frozen video frame.

To understand the effect of overload on video stalls, we model the problem analytically (described in detail in the appendix) and discuss the results here using Equation 10 and Equation 12. Assuming that all traffic carries stored video (of bandwidth \( \lambda \)) over UDP or TCP. Figure 2 shows the impact of an hour of basestation overload on the number of stalls \( b_o \) and the amount of stall time \( b_t \) in each stall instance. The overload factor \( O \) is defined as the fraction of additional capacity required to satisfy users’ bandwidth.
requirements over the actual basestation capacity. The different lines correspond to different amount of buffering done by a client player after each buffer underrun. As expected, the number of stalls and the stall time increases with increasing overload.

What is more interesting and not immediately obvious is that the number of stalls increases more dramatically with overload than the time of stall. This is more inconvenient to users since, for instance, two separated short stalls typically give a more adverse experience to a user than one long stall. The different lines indicate that larger buffers at the client side lead to the more desirable effect of increased stall time, but fewer number of stalls. However, (1) configuring too large a buffer size would unnecessarily delay the video playback in the beginning, and (2) adapting the buffer size dynamically based on overload would require a new protocol between the network infrastructure and the client players. Admission control, however, creates a bad perception to users of network unavailability, which if not controlled, can make users migrate to other operators. Hence, the network operator has the challenge of minimizing the number of users observing network unavailability, while also maximizing the quality of experience for the admitted users.

2.2 Operator-enabled Advertising

Mobile advertising, although still in its nascent stage, is advancing rapidly with both content providers and mobile network operators trying to leverage the huge revenue potential [8]. Recent forecasts indicate that the U.S. local advertising market will reach $144.9 billion in 2014 [25]. To leverage this growth effectively using the mobile media, 56% of marketers plan to increase their mobile marketing budget [26]. Also, predictions indicate that location-based mobile spending by users will reach $4 billion by 2015 [12].

While mobile advertising can also be enabled through content providers, the cellular network domain appears to be more favorable to operator-enabled advertising for several reasons [32, 3, 30]: (1) it is much easier for the operator to acquire subscriber personal and location information, mobility patterns, etc. that help targeted and relevant advertising, (2) the operators already have a trust relationship with the customers established, and customers may be more tolerant of the operators using their personal information, (3) operators typically have much larger customer bases than most individual content providers, and (4) an operator is better aware of the best channel (e.g. SMS, MMS, banners, etc.) and best time (e.g. based on channel and overload conditions) to reach the user for a particular type of advertisement. Hence, from an advertisement agency’s perspective, an operator has a better reach to users than any specific content provider. Commercial solutions like Optism [16] are already providing mechanisms for operator-enabled advertising.

In fact, in the future, the SLAs between users and network operators may be mainly in terms of advertising and not calling plans; the operators can offer incentives to users for agreeing to receive advertisements [30]. In support of this hypothesis, a recent survey reveals that users are willing to receive occasional advertisements for reduced price of service [10]. These advertisements may be delivered to the user at any point of time during active network access. For instance, while it is common to deliver advertisements either at the beginning or at the end of user requested videos, some content providers today (such as YouTube [1] and Hulul [9]) deliver advertisements several times in the middle of long videos. In other words, users are aware that the operator will occasionally deliver advertisements and interrupt their activity.

2.3 Opal’s Idea

Opal aligns advertisement delivery with overloads as much as possible to minimize the disruption users perceive. Opal explicitly creates long enough opportunities for advertisement delivery to selected users during overloads. The design, however, should meet several requirements. Firstly, Opal is required to minimize disruption on any one particular user, in an attempt to equalize the quality of experience across all users. Secondly, Opal should ensure that an advertisement is displayed to as many unique users as possible to satisfy an advertisement owner and maximize the revenue generated. Finally, Opal has to account for dynamically changing working sets of users: a given user typically accesses the network actively only at certain periods of time.

In the next section, we describe the design of Opal that addresses each of these challenges. Since videos are more amenable to interruptions, and videos also contribute a large percentage of bytes to the total data traffic, our goal is to target only video users during overload for advertisement delivery. Identification and separation of video users from others, however, requires additional mechanisms, such as deep packet inspection and bandwidth reservations. Fortunately, this can be achieved by previously known techniques [14, 19, 44], and hence we do not expand on them. In this paper, we focus on the design and evaluation of Opal assuming that all users are accessing video. We later discuss how Opal can easily incorporate bandwidth reservation mechanisms to separate users accessing videos from those accessing other kinds of traffic such as Web and file transfers.

3. Opal DESIGN

Opal is instantiated as a gateway-level solution in the cellular operator’s access network external to the basestations. Since the gateway will typically handle traffic for multiple basestations, it hosts multiple Opal instances, each handling traffic for one basestation. In what follows, we begin with a discussion of service level agreements (SLAs) between the user and the network operator, and the network operator and the advertisement agency that wishes to advertise over the operator’s network. We then describe an architecture for overload management, and an advertisement scheduling approach to meet the different service level agreements.
3.1 Service Level Agreements

In designing Opal, we consider that each user $j$ is provided a minimum reserved downlink bandwidth $\lambda_j$, which is set based on contracts (data plans) between users and the network operator\(^1\). Observe that the wireless resource usage (such as MAC resource slots) for achieving a given bandwidth varies with the MCS (modulation and coding scheme) for a user. Hence, to avoid getting penalized by users with bad channel quality, the mobile network operator may actually sign a contract with a user $j$ to provide a contingent reserved bandwidth as follows: The MNO defines $R_{j, \text{eff}}$ as the effective bitrate above which the network provides at least the minimum reserved bandwidth of $\lambda_j$. Then, if $R_{j, \text{eff}}$ is the instantaneous effective bitrate during system operation, the adjusted reserved bandwidth $\lambda_j$ is

$$\lambda_j = \lambda_j \cdot \min(1, \frac{R_{j, \text{eff}}}{R_j}) \quad (1)$$

Opal strives to provide at least a bandwidth of $\lambda_j$ to each user $j$ when the basestation is overloaded.

The contract between an advertisement agency and a network operator includes the minimum number of times an advertisement is displayed to users over a given time period (e.g. a day). Opal attempts to meet this requirement both when the basestation is overloaded and when it is underloaded.

We assume that the base stations provide the gateway on which Opal is deployed, a feedback of two quantities every $\delta$ units of time—(i) the base station utilization, and (ii) the current modulation and coding scheme (MCS) of each active user. The second quantity helps keep $\lambda_j$ updated based on the user’s changing channel conditions. Both quantities help Opal synchronize with the basestation capacity, and also identify overload conditions. This information is either already available on commercial Macroe cell base stations (e.g. the NEC WiMAX basestation [13] provides this information via SNMP), or can be added easily. The basestation utilization depicts how fully the wireless channel is utilized, and is defined as the ratio of the resource slots used for transmission and the total available slots in the last $\delta$ units of time.

3.2 Overload Management

The basic idea of Opal is depicted in Figure 3. Opal considers two groups of users: regular users that receive their traffic share of $\lambda_j$, and targeted users that receive advertisements during basestation overloads. Opal initially maps all users to the regular group, and employs proportional fair resource allocation (using the periodic MCS feedback from the basestation) across the users. Opal employs per-user queues, and monitors the service rate (averaged over the last 10$\delta$ units of time) and the queue length for each user. If a user’s queue builds up beyond a threshold and the service rate for the user is below $\lambda_j$, some active users with minimum advertisement views (represented by $z_j$ for flow $j$) are migrated to the target group (See Algorithm 1). Maintaining two groups of users as above essentially achieves a notion of soft admission control.

The goal of the migration approach in line 3 in the algorithm is to discover the appropriate number of users to be maintained in the target group such that the remaining users receive adequate service rate. This may happen in multiple steps due to the condition in line 3. Every time a user is moved to the target group, $z_j$ is incremented by 1. The users moved to the target group remain there for $V_1$ units of time. $V_1$ is configurable by network operators, and represents the amount of time advertisements are shown to a user at a stretch. Each user that is moved into the target group is marked (see lines 4-6 in Algorithm 1) to ensure that he will not get repeatedly selected for advertisement delivery even if he has very low $z_j$; such a user will get selected again only after all other active users have also been shown an advertisement once. Additionally, a user is unmarked once a certain amount of time has passed (e.g. 15 minutes) after being marked, assuming that repeating an advertisement beyond this time is tolerant to the user during overload. This marking achieves a notion of a round within which an active user is selected only once.

---

\(^1\)While such contracts do not exist in today’s cellular networks, we believe that the increasing data traffic often containing time- and bandwidth-sensitive traffic such as video, and the scarcity of wireless spectrum will make such contracts imminent in the near future.
The algorithm has an interesting property that if there is positive discrepancy in \( z_j \) (i.e. difference between maximum and minimum \( z_j \)) among the current active users, the algorithm \textit{never increases the discrepancy} in \( z_j \); the discrepancy either reduces or remains the same. In other words, \texttt{Opal} greedily attempts to equalize the number of advertisements seen by users.

The overload factor \( \mathcal{O} \) is defined as the fraction of additional capacity required to satisfy users’ bandwidth SLAs over the actual basestation capacity \( C \). Then,

\[
\mathcal{O} = \frac{\sum \lambda_j - C}{C} \tag{2}
\]

The above equation and the condition in line 3 of Algorithm 1 ensures that \texttt{Opal} automatically discovers the appropriate number of users that should be in the regular group, even as the capacity \( C \) fluctuates.

Notice that we define \( \mathcal{O} \) in terms of the SLA agreed between the users and the network operators, and not in terms of the traffic demand of the users. We assume that each user has a finite queue in \texttt{Opal}, on which we employ active queue management. Specifically, if the queue size exceeds a threshold \( R_j \), one packet every 10 packets is dropped. This is helpful in indicating congestion if the user has TCP flows, and is shown to work well for broadband wireless networks [48]. If the queue size exceeds \( 2R_j \), all subsequent packets are dropped. In the prototype, we set

\[
R_j = 0.2 \cdot \lambda_j \tag{3}
\]

Finally, a synchronizer ensures that only enough number of packets are sent to the basestation periodically to match with the capacity of the basestation, and hence avoid queue build up at the base station. Note that every time the basestation utilization is close to 100%, the aggregate service rate of all the users represents the basestation capacity \( C \). The design and evaluation of a robust synchronizer that adapts to changing available capacity of the basestation is presented in our previous work [44], and hence we do not discuss it here. Without the synchronizer, \texttt{Opal} cannot detect overload effectively based on just the utilization feedback from the basestation.

### 3.3 Tunable Unavailability Framework

We now derive a simple analytical framework to highlight \texttt{Opal}’s functionality better, and mainly to help a network operator configure the value of \( V_t \) in an informed manner.

Consider a time interval \( T \) during which the basestation is continuously overloaded. For simplicity, let all users have the same reserved bandwidth of \( \lambda \). Let \( M \) be the number of users that can be supported in the regular group at an average service rate of \( \lambda \), and let \( N \) be the number of users in the target group receiving advertisements at a rate of \( \lambda_a \). Let \( G = M + N \) represent the total number of users. Then Equation 2 can be rewritten as

\[
\mathcal{O} = \frac{GA - C}{C} \Rightarrow G = \frac{(1 + \mathcal{O})C}{\lambda} \tag{4}
\]

If \( C \) is the capacity of the basestation, then \( C = M\lambda + N\lambda_a = (G - N)\lambda + N\lambda_a \), which on simplifying gives

\[
N = \frac{G\lambda - C}{\lambda - \lambda_a} \tag{5}
\]

Substituting the value of \( G \) from Equation 4,

\[
N = \frac{\mathcal{O}C}{\lambda - \lambda_a} \tag{6}
\]

Now, with \texttt{Opal}, let \( V_t \) be the time for which a user gets moved into the target group. \texttt{Opal} ensures that \( N \) users exist in the target group at any point of time, so that \( M \) users receive a service rate of \( \lambda \). Hence, referring to Figure 4, consider the duration \( T \) to be intervals of \( V_t \), each accounting for \( N \) user migrations. Then, the total number of user migrations into the target group in time \( T \) is given by \( N \cdot \left( \frac{T}{V_t} \right) \). With \( G \) active users in the system, the number of migrations per user is

\[
U_n = \frac{1}{V_t} \cdot \frac{NT}{G} \tag{7}
\]

Using Equations 4 and 6 and simplifying,

\[
U_n = \frac{T\mathcal{O}\lambda}{V_t(1 + \mathcal{O})(\lambda - \lambda_a)} \tag{8}
\]

Hence, given a choice of \( V_t \), and the basestation overload, and known values \( \lambda \) and \( \lambda_a \), a network operator can estimate the number of advertisements a user will be shown (See Figure 5). Alternately, to control the number of interruptions \( U_n \) per user, an operator may choose to use a longer \( V_t \) during higher overloads and show advertisements for a
Algorithm 2 \texttt{ad\_sched\_overload (User U)}

1: \textbf{while} (time in target group < $V_t$) do
2: \hspace{1em} $Ad \leftarrow$ random entry from adlist
3: \hspace{1em} sched (U, $Ad$)
4: \textbf{end while}
5: Move U to regular group

Algorithm 3 \texttt{ad\_sched\_underload}

1: \textbf{while} (exists $Ad$ with shows below $S_{ad}$) do
2: \hspace{1em} U $\leftarrow$ random entry from active users
3: \hspace{1em} $Ad \leftarrow$ $Ad$ with minimum shows
4: \hspace{1em} sched (U, $Ad$)
5: \textbf{end while}

longer period of time. The graph shows that for masking off a persistent overload of 20% for a duration of 10 minutes at a basestation, it is sufficient to show 3 advertisements to each user of about 35 second duration. More importantly, the effect of increasing $V_t$ is more dramatic initially and then flattens relatively afterwards. This behavior is advantageous because, the operator can easily strike the tradeoff by choosing an operating point $V_t$ in the region where the curve begins to flatten; this operating point, however, is different for different overload factors.

3.4 Advertisement Scheduling

\texttt{Opal} schedules as many advertisements as possible during basestation overloads to the targeted users. However, if the basestation does not get overloaded enough during a day, each advertisement is shown at least at a certain number of times in a given period to satisfy advertisement contracts. We next address the question “which advertisement should be scheduled at any instant of time?” This question has to be addressed with the objective that each advertisement is seen by as many unique users as possible for maximal coverage.

\textbf{During overload}, specific users are first selected by Algorithm 1 based on $z_i$ to be moved to the target group. Hence, to maximize the number of unique advertisements served to users without explicitly maintaining state, \texttt{Opal} selects an advertisement randomly. Algorithm 2 depicts the basic idea of advertisement scheduling during basestation overload. Random selection of the advertisements avoids any synchronization between the number of advertisements and number of active users that can cause the same advertisement being repeatedly shown to a user. Further, random selection ensures that the advertisement list can be dynamic, i.e., advertisements can be removed and added.

\textbf{During underload}, specific advertisements are first selected to meet the contracts; i.e., the advertisements that have been shown below $S_{ad}$ number of times are scheduled for delivery. The threshold $S_{ad}$ increases linearly with time with a slope equal to the frequency with which advertisements have to be displayed. To maximize the number of unique users seeing a given advertisement without maintaining state, \texttt{Opal} selects users randomly. Algorithm 3 represents the overall idea, and Figure 6 summarizes the advertisement scheduling behavior. During underload situation, if the number of shows for an advertisement is below $S_{ad}$ the advertisement is scheduled for delivery. Each overload period may cause greater than $S_{ad}$ number of advertisements to be shown, and hence the overload period may be followed by a no-show period when no advertisements are shown to users. The number of shows for an advertisement and $S_{ad}$ are reset periodically. Note that the only state \texttt{Opal} maintains is the number of times an advertisement has been shown, and not to which specific users it was shown. Hence, advertisement delivery is independent of the set of users connected to the basestation at any point in time.

4. PROTOTYPE

In this section, we describe the prototype implementation on a WiMAX network testbed. We first provide a brief background of the relevant features of WiMAX. Note that other broadband technologies such as LTE have similar features, thereby making \texttt{Opal}’s design equally applicable. The mobile WiMAX network comprises of two components (1) the access service network (ASN) including the ASN gateways and the base stations, and (2) the connectivity services network (CSN) that provides IP connectivity. The ASN Gateway interfaces the base stations with the CSN, and handles many tasks such as admission control, mobility management, QoS and policy enforcement, and caching of subscriber profiles. Each gateway typically handles hundreds to thousands of base stations. The WiMAX base station uses an orthogonal frequency division multiple access (OFDMA) MAC frame structure for scheduling downlink and uplink transmissions to/from the subscriber stations (clients). The base station schedules a MAC frame periodically, typically every 5 ms.

Our prototype WiMAX platform consists of a Picochip [18] WiMAX basestation (IEEE 802.16e compliant), a WiMAX Profile-C Access Service Network (ASN) gateway, an Internet (CSN) gateway and several Beceem [2] PCMCIA and USB clients (See Figure 7). The ASN gateway functionality is deployed on a standard dual-core Linux machine connected to the base station with a 100 Mbps link.

We implement \texttt{Opal} on the ASN Gateway using the Click Modular Router Framework [45] as a user-level module that intercepts all data packets to the basestation. We modify the Picochip basestation MAC code to provide feedback on the average basestation utilization and the average MCS per client to the \texttt{Opal} instance every $\delta = 100$ milliseconds, since this feedback was missing from the Picochip basestation. We implement the Advertisement Server on Node 1 (a Linux machine) in the figure connected directly to the
ASN Gateway. The Advertisement Server listens on a control socket for messages from Opal. When it receives a message containing the index of the client that has been moved to the target group and the advertisement index, the Advertisement server streams advertising to the target group via the Ad server. Assuming the Ad server is implemented within the gateway or co-located in the same data center, the overhead of message delivery is minimal. On the arrival or departure of each packet, only the arrival rate and the service rate of the appropriate flow is updated. Hence, Opal’s overhead does not hinder its deployability, given the novel operator-enabled-advertising feature it enables for revenue generation.

5. PROTOTYPE EVALUATION

In this section, we evaluate the efficacy of Opal on the WiMAX network testbed shown in Figure 7. We describe the setup, workloads and metrics before discussing the results.

**Setup:** We perform all the measurements in static settings and with walking mobility in an indoor office testbed, since the Picochip basestation is a low-end system designed for providing Femto-cell coverage (6 to 8 users). Nevertheless, the testbed is sufficient to highlight several aspects of Opal, which will remain equally valid on Macrocell basestations. We also study Opal in larger scale scenarios with a simulator in the next section. All prototype experiments are done with over-the-air transmissions between the basestation and the clients on the 2.585-2.595 GHz channel (10MHz). To emulate clients at different channel qualities (i.e. different MCS), we place client laptops in different cubicle locations in the office.

**Workloads:** We consider a mix of workloads including video over TCP, video over UDP, and Iperf generated traffic. For video over TCP, we stream videos from YouTube [1] over the Internet. We use Adobe Flash Player plugin in Internet Explorer to view the videos. For video over UDP, we use VLC for Linux on both the server (node 1) and the clients. VLC is used for both user requested videos and for advertisement delivery. Even for experiments with Iperf, the server resides on node 1. We repeat each experiment several times to ensure statistical confidence in the experimental results.

**Metrics:** For video over UDP, we use PSNR—a standard metric of video quality that is a function of the mean square error between the transmitted video and that displayed at the receiver. We use the EvalVid tool [38] to compute the PSNR. Klaue et al. [38] show a mapping between PSNR and qualitative user rating; PSNR>37 is considered as excellent quality and PSNR between 31 and 37 is considered good.

---

**Figure 7:** Opal prototype on WiMAX testbed.

**Figure 8:** Snapshot of user-requested video and advertisement delivery in our prototype.
For video over TCP, we cannot use PSNR for two reasons: (1) we do not have access to the video content from YouTube, and (2) more importantly, PSNR does not incorporate the effect of video stalls due to buffering. Observe that since TCP ensures reliable delivery of video frames, videos only incur stalls and not dropped frames; whereas, PSNR measures the effect of dropped or late frames, and is better suited for video streaming over UDP. Hence, for video over TCP, we use the number of stalls in the received video to indirectly measure user satisfaction in viewing the received video.

5.1 Results

The main goal of prototype experiments is to show that Opal effectively migrates users to the target group such that users in the regular group receive their reserved bandwidth, thereby improving the quality of experience for users; unacceptable quality of experience is perceived as network unavailability by users. We also demonstrate Opal’s adaptation to changing basestation capacity due to user mobility.

5.1.1 Iperf Flows

We begin with a simple experiment containing Iperf TCP flows. We consider six clients placed statically in different office locations. They are placed at similar locations such that they receive data at a MCS of QPSK\(^1/2\). Each client is configured with a reserved bandwidth \(\lambda_j\) of 750 Kbps, and \(R_j^{\text{req}}\approx 3\) Mbps (MCS of QPSK\(^1/2\)). The sum of the reserved bandwidths exceeds the basestation capacity (=3 Mbps at QPSK\(^1/2\)), and hence emulates an overload condition when all clients are active. With Opal, the users migrated to the target group receive low-rate flows as advertisements, with \(\lambda_j\approx 60\) seconds. Figure 9 shows the throughput obtained by three of the clients with and without Opal. With Opal, some users get selected to receive advertisements such that the other users achieve their reserved bandwidth. Figure 9(b) shows that without Opal, the bandwidth is shared across users, and is much lower than anyone’s reserved bandwidth.

5.1.2 TCP Video Flows

We again consider six clients placed statically in different office locations as above, requiring MCS of QPSK. The clients access three different YouTube videos (two users each video), namely Shrek Trailer(580 Kbps average rate), Matrix Trailer(650 Kbps average rate) and UP Trailer(700 Kbps average rate). Each client is configured with \(\lambda_j\approx 750\) Kbps. We then measure the service rates obtained by the different flows with and without Opal, and count the number of stalls that the six clients observe while viewing the videos. As shown in Figure 10, all clients see significant number of stalls during video playout without Opal. However, with Opal, all users view good quality video with 0 to 2 stalls for receiving advertisements. Figure 11 shows the service rate received by two users with time for the YouTube videos and the advertisement videos. The graph shows that even though YouTube video rate fluctuates significantly around the average, Opal effectively identifies the number of supportable users in the regular group, and moves the rest to the target group.

5.1.3 UDP Video Flows

In this experiment, we show the efficacy of Opal with UDP-based video streams. We consider six clients placed at different locations such that two each are at MCS levels of 64QAM, 16QAM and QPSK, respectively. Different videos are streamed out from the VLC server in a loop to each client. We measure the PSNR observed at the client based on the frames received, dropped and late (i.e. arrive after the playout deadline). In Figure 12(a), we show the average PSNR in different one minute intervals for three out of the six users. The graph shows that in different intervals, different clients are chosen to be migrated to target group.

Figure 9: Service rates Received by the 6 clients with and without Opal for Iperf TCP-Traffic.

Figure 10: Stalls in YouTube videos seen by the 6 clients with and without Opal.

Figure 11: Service Rate received by 2 out of the 6 users steaming YouTube video with Opal.
and shown advertisements. We represent the users moved to the target group with black bars. Note that the black bars do not represent a PSNR value. The other clients receive good quality video and hence, high PSNR values. In the case without \textit{Opal}, as shown in Figure 12(b), all six clients receive bad quality video all the time, and hence show poor PSNR values.

5.1.4 Efficacy with User Mobility

In this experiment, we demonstrate the efficacy of \textit{Opal}'s dynamic overload management in response to wireless capacity fluctuations due to user mobility. We setup the basestation with four clients running Iperf [11] TCP flows. Each client has a reserved rate $\lambda_j$ of 1.5 Mbps. Initially, all the clients are placed at a location close to the basestation such that they receive data at the highest rate (64QAM). At about 50 seconds into the experiment, two clients are moved away from the basestation such that they receive data at the lowest rate (QPSK). We repeat the experiments with and without \textit{Opal}. The case without \textit{Opal} emulates traditional admission control approaches that only make decisions at the arrival of a new flow. Moreover, majority of the traditional schedulers employ some variant of proportional fairness across users, that proportionally penalize the clients with lower MCS.

Figure 13 shows that all client receive their reserved 1.5 Mbps with and without \textit{Opal} initially, since all of them are at the same channel quality, and the sum of their reserved bandwidths is below the basestation capacity. At 50 seconds, Figure 13(a) shows that, without \textit{Opal}, both users 3 and 4 receive reduced service rates. However, as shown in Figure 13(b), users 1, 2 and 3 receive the required service rates with \textit{Opal}, while user 4 receives an advertisement.

6. SIMULATION STUDY

We now study the performance of \textit{Opal} using simulations to evaluate with large number of users; these experiments could not be done on our prototype testbed due to its small scale. We use an in-house system-level OFDMA simulator that can simulate multiple basestations and multiple clients for each basestation, fast channel fading, user arrival and departure, and different types of traffic (Video, FTP and Web traffic).

We implement the algorithms for overload management and advertisement delivery in the simulator for evaluating \textit{Opal}. In the simulation runs, we setup different number of users each configured with a reserved bandwidth $\lambda_j=300$ Kbps, and running FTP or video traffic. We set $R_j^{\text{req}}=9$ Mbps for all users. When \textit{Opal} marks a user as targeted, the user receives 10Kbps of bandwidth for advertisement delivery for $t_{\text{ad}}=30$ seconds; we later experiment with increasing bandwidth requirement for advertisement delivery. When a user returns to the regular group, its reserved rate is reconfigured to the original value. The setup has 10 unique advertisements, and the advertisement contract is to display each advertisement once every 10 minutes. The experiments simulate a duration of one hour.

With simulations, we specifically study the efficacy of \textit{Opal} to detect overload with variable traffic and number of users, the efficacy of advertisement distribution across users, and the effect of variable bandwidth requirement for advertisements itself.

6.1 User Migration

In this section, we show that \textit{Opal} is effective in selecting the appropriate number of users to be migrated to the target group. We first run the simulation with 25 users in the system—a case in which the basestation is always underloaded. Each user continuously receives FTP flows. To meet the advertisement contracts, \textit{Opal} migrates certain number of users into the target group at appropriate times. As shown by the CDF in Figure 14(b), 5% of the total time users are targeted for advertisement delivery. Otherwise, most of the time the 25 users are in the regular group (Figure 14(a)).

We repeat the simulation experiments with 30, 32 and 34 users in the system. Since the reserved rate per user is kept the same during the three runs, the overload of the system increases with number of users. As shown in Figure 14(a), \textit{Opal} ensures that about 26-28 users always meet their reserved bandwidth irrespective of the load in the system. Note that since channel conditions fluctuate due to fast fading, the basestation capacity keeps fluctuating; this also indicates that traditional admission control techniques that admit or reject flows on their arrival are not effective because capacity can reduce after admitting users. \textit{Opal} adapts to these fluctuations effectively and maintains the appropriate number of users in the regular group. Figure 14(b) shows that \textit{Opal} increases the number of targeted users with system overload.

We now vary the bandwidth required for the advertisements with 30 users in the system. We plot CDF of the number of regular users and targeted users during the 1 hour. Intuitively, if the advertisement rates are increased, then the number of users that the system can support at a time reduces. To study this effect with \textit{Opal}, we repeat
the simulation runs for advertisement rates of 10 Kbps, 100 Kbps and 200 Kbps per advertisement. As seen in Figures 15(a) and (b), the number of users in the regular group decreases with increase in the advertisement bandwidth, as expected. What is more interesting is that the gap between the lines in Figure 15(b) increases at the top (e.g. the 200 Kbps line relatively flattens compared to the 100 Kbps line). This is explained by Equation 6: a linear increase in $\lambda_a$ causes more than a linear increase in $N$. Hence, an operator should choose as low a bandwidth as possible for advertisements.

6.2 Advertisement Scheduling

We next study the behavior of advertisement scheduling during underload and overload. We configure the basestation with 25 active users for underload and 30 active users to cause overload. In Figure 16, we plot the number of users viewing advertisements as a function of time. In Figure 16(a), we see that Opal schedules each advertisement every 10 minutes during persistent underload to satisfy the advertising contract. There are 10 unique advertisements, so 10 users are chosen each time. In a scenario in which the basestation is overloaded, as shown in Figure 16(b), Opal mostly shows advertisements to 1-3 users. The exact number of users that are shown advertisements varies with time due to the fluctuating channel capacity (0-5 in this experiment).

From the same simulation run, we plot a 2-D array of the advertisement index versus the user index for both underload and overload. This plot shows the number of views of a particular advertisement per user, and can be used to count the total number of shows for each advertisement. As seen in Figure 17(a), the aggregate views of each advertisement is equalized across all the advertisements. This results shows the effectiveness of selecting an advertisement with the least shows during underload. The advertisements are also uniformly distributed among the users. Since Opal selects users randomly for advertisement delivery, it maximizes the number of unique users that view a particular advertisement, without any explicit state maintenance.

Figure 17(b) shows the 2-D array when the basestation is overloaded. Since at any time instance, Opal selects the user with the least advertisement views, the aggregate views of each user are effectively equalized across all active users. Most users view a uniform combination of the advertisements and see the same advertisement at most 2 or 3 times. The views are not exactly equal, however, because of the random number generator and not because of Opal itself.

6.3 Improvement in Quality of Experience

Finally, we simulate a scenario (with and without Opal) containing different number of users that continuously access videos. The system without Opal schedules advertisements periodically to satisfy advertisement contracts. We simulate client buffering (and re-buffering every time the buffer goes empty) to study the number of times the user observes stalls while viewing video. We set the client buffer size $B=200$ KB. Figure 18(a) shows the number of interruptions (due to stalls and advertisements) observed by users with and without Opal. The graph shows that Opal reduces the total number of interruptions by 30-50% by appropriately scheduling advertisements, compared to a system without Opal. Note that users observe stalls with Opal also, although they are few in number. The benefit of Opal increases with increasing load. However, the increase in benefit does not continue beyond a
certain number of users since the number of advertisements that need to be shown increases significantly with overload for a given choice of \( V_t \) (Recall from Figure 4). For higher overload, Opal has to show advertisements for longer duration. Figure 18(b) shows the CDF of stall time without Opal. The graph shows that the stall time increases with increasing load, but not as dramatically as the number of stalls, as shown in Figure 18(a). Recall that this behavior is exactly as predicted in Figure 2.

7. LIMITATIONS AND DISCUSSION

Opal makes an assumption that network unavailability is more annoying to a user than watching advertisements intermittently. More specifically, we assume that frequent pauses during buffer underflows annoy users more than viewing advertisements occasionally. While the assumption is intuitive, admittedly, only a real deployment and active long-term feedback from users can conclusively validate this assumption. This user study is an interesting topic for our future work.

The metric \( z_j \) is simplistic in the paper as it only captures the total advertisement viewed by the user. However, it can be extended to more sophisticated formulations for incorporating skew in user-requested video bandwidths, variable advertisement plans between the operators and users (e.g. receiving maximum of 5 Ads/day vs. 20 Ads/day), control of frequency of interruptions, etc. Some of this sophistication is also explored in our previous work [44]. We believe that this is a policy question, while our focus in this paper is mainly on the mechanisms for aligning advertisements with basestation overloads. Nevertheless, the appropriate set of factors to incorporate into \( z_j \) is an interesting topic, and requires further research, including a user perception study.

As an optimization, an MNO can use deep packet inspection (DPI) or other detection techniques to estimate the average rate such that the quality of user experience does not deteriorate. In this case, \( \lambda_j \) may be set as the minimum of the value given by Equation 1 and the rate estimated by DPI. For example, most YouTube traffic headers contain the average rate of the video, even if the arrival rate fluctuates significantly.

The design of Opal needs to be extended appropriately to incorporate targeted advertising, although the system will incur more overhead than the current design. When a user is chosen to be moved to the target group, the advertisement can be chosen from a pool of advertisements that are relevant to the interests of that particular user.

Opal can incorporate location-based advertising in the future. Advertisements based on users’ course-grained location such as a city or a county can be easily delivered with the current architecture of Opal. The advertisement database can be customized for each specific basestation, since each basestation covers only a small geographical area. However, advertisements based on user’s current fine-grained location (e.g. Ad of a Pizza Hut when a user is within a mall) has to be delivered as and when appropriate and not just when basestation overload occurs. We believe that this is an independent problem, and Opal itself should not be used to deliver such Ads. The system for delivering fine-grained location-based advertisements can work independently of Opal.

While end-host adaptation techniques can be used to lower the impact during overloads, such techniques have been shown to take long time to adapt [23]. Majority of video services like Netflix, Hulu and YouTube employ tcp-based transmission which makes it hard to do fast rate-control. Moreover, a video below a certain quality may not be acceptable to users, thereby hitting an overload condition where servers cannot adapt any longer, beyond which mechanisms such as Opal are required. Nevertheless, combining server adaptation with Opal in mobile settings with fluctuating capacities on the wireless channel is an interesting topic for future work.

Since Opal is designed as a Gateway-level solution, it moves scheduling functionality of the basestation into the gateways. For any effective resource management at the gateway-level (in the mobile core), scheduling of wireless resources has to be moved into the gateways. Otherwise, any decision taken at the gateways will be ineffective. Such advanced features can only be deployed in mobile networks either by modifying the basestations or by moving the scheduling into gateways. Note that solutions such as Sandvine [19] shape traffic and influence scheduling at basestations, and are already deployed widely in cellular networks. Moreover, the packet scheduling decisions taken at the gateways are at timescales of 100s of milliseconds. Finer timescale wireless frame scheduling to counter fading, multipath, rate-adaptation and to take advantage of MIMO, multi-user diversity etc. are still taken at the basestation.

An alternate approach to unicast transmission of advertisements on a per user basis is to use multicast. Using multicast reduces the amount of bandwidth spent for advertisements, which is especially important during overload. However, this would require migrating users in groups such that the cost of multicast setup (including dynamic admission of users to groups and the operation at robust transmission rates to reach all users of the group) is justified. Nevertheless, understanding the tradeoffs between unicast and multicast for advertisements is an interesting topic for future work.

Yet another approach of displaying advertisements is to deliver them during underload and store locally on a user device, and play them during overload through control signalling. While this approach reduces the bandwidth required for Ad delivery during overload, it needs synchronization between the network infrastructure and the clients to ensure that users see a smooth transition between videos and advertisements.
8. RELATED WORK

We discuss related work in three categories: content adaptation based on available bandwidth, admission control and mobile advertising. While the former two contain resource management solutions that eventually reduce network costs by making effective utilization of resources, the last category increases revenue for network operators, as we discussed in Figure 1.

Content adaptation: Several techniques have been proposed that involve receiving feedback at the content source on the link characteristics (e.g. bandwidth, delay or loss) or perceived user QoE (e.g. video quality) at the receiver, and adapting transmission at the source [35, 36, 37, 54]. For instance, Kim et al. [37] explore a probe-based channel-adaptive video streaming solution over 3G links using RTCP feedback, and Juan et al. [35] use coarse-grained feedback from the cellular base stations to adapt the frames generated by the video source. Alternatively, some approaches, including several commercial solutions, modify video streams at intermediate nodes for adapting to link capacity variations [17, 15, 50]. Finally, several related efforts explore differential treatment of low priority frames during base station scheduling to make enough room for high priority (i.e., important) frames during overloads [42, 43, 49]. These content adaptation solutions are complementary to Opal, and are required in conjunction to combat the rapid growth of data traffic.

Admission Control: Admission control for both voice traffic and multimedia traffic has received significant attention in the past [22]. Several works focus on reducing call handoff and blocking probability through bandwidth reservations [24, 31, 41], and giving isolated resources to handoff calls [29]. For example, Wu et al. [53] design a dynamic admission control that is adaptive to a wide range of system parameters and traffic conditions. Kwon et al. [40] explore admission control for multimedia traffic; they dynamically adjust bandwidth for existing flows to admit each newly arriving flow. However, they do not consider fluctuating channel conditions or video rates. In our previous work [44], we explore a soft admission control approach for UDP flows carrying streaming video, which incorporates long-term fairness into the system design itself to provide similar perception of the network service to all users.

Mobile Advertising: Although mobile advertising is still in its nascent stage, several commercial solutions have begun to appear already [16]. Examples of such solutions include permission based advertising [16], location oriented search and advertising [20], targeted advertising [32, 46], etc. Bulander et al. [27] provide a nice summary of the different design considerations for a mobile advertising system including the end-users, advertisers and the delivery system. None of these solutions, however, consider any networking related issues such as scheduling and bandwidth management. Several research efforts are underway on complementary issues such as user location tracking and privacy [33], improving relevance of advertisements to users [26], etc. On the scheduling aspects, Tripathi et al. [52] model and solve the advertisement scheduling problem with the assumption that the advertisement firm is a separate entity, and buys a certain amount of resources from the cellular network operator for advertisement delivery. Reyck et al. [47] solve a broadcast scheduling problem to decide which advertisements to send out to which customers and at what time, given a limited capacity of broadcast timeslots, while maximizing the utility for both users and advertisers.

9. CONCLUSION

Opal is a novel system to align advertisement delivery with cellular base station overloads. Opal builds on two observations: (1) Cellular network operators are exploring advertisement delivery as a new revenue generator, and (2) Cellular base stations perceive intermittent overloads at different timescales such as different times of day. Opal is designed over a simple framework for tunable unavailability of network service to users. As operators incorporate advertisement delivery into their infrastructure to leverage the large revenue potential, research on novel architectures to enable efficient delivery of advertisements becomes increasingly important. While this paper touches on one such topic, several interesting avenues remain for future work.

Acknowledgment

This paper benefited significantly from MobiSys 2011 reviewer comments, for which we are very thankful. We also thank our shepherd, Gaetano Boriello, for overseeing the editing process.

10. REFERENCES

Appendix: Analysis of stall time and number of stalls in a system with no soft admission control

In this section, we analyze the effect of admitting user flows beyond the capacity of the basestation in a system without Opal. Specifically, we analyze the effect on video flows. The players on the clients typically buffer a part of the video before playing out. If the buffer becomes empty, they stall playing and buffer enough bytes before playing again. Let us say, if the video player on the client does not have any video frames to play, it buffers packets until it receives $B$ bytes of data. Let $b_t$ be the buffering time, and hence the stall time that a user observes. Let each user be serviced with an equal rate $\mu$. If $M$ is the number of users that can be supported a bandwidth of $C$, and $N$ is the additional number of users in the system, then the additional number of users in the system, $N$, can be calculated as:

$$N = M \times \frac{C}{\mu}$$

where $C$ is the capacity of the basestation and $\mu$ is the rate at which the video is transmitted. If $M$ is the number of users that can be serviced with a bandwidth of $C$ and $N$ is the additional number of users in the system, then the additional number of users in the system, $N$, can be calculated as:

$$N = M \times \frac{C}{\mu}$$

where $C$ is the capacity of the basestation and $\mu$ is the rate at which the video is transmitted. If $M$ is the number of users that can be serviced with a bandwidth of $C$ and $N$ is the additional number of users in the system, then the additional number of users in the system, $N$, can be calculated as:

$$N = M \times \frac{C}{\mu}$$

where $C$ is the capacity of the basestation and $\mu$ is the rate at which the video is transmitted. If $M$ is the number of users that can be serviced with a bandwidth of $C$ and $N$ is the additional number of users in the system, then the additional number of users in the system, $N$, can be calculated as:

$$N = M \times \frac{C}{\mu}$$

where $C$ is the capacity of the basestation and $\mu$ is the rate at which the video is transmitted. If $M$ is the number of users that can be serviced with a bandwidth of $C$ and $N$ is the additional number of users in the system, then the additional number of users in the system, $N$, can be calculated as:

$$N = M \times \frac{C}{\mu}$$

where $C$ is the capacity of the basestation and $\mu$ is the rate at which the video is transmitted. If $M$ is the number of users that can be serviced with a bandwidth of $C$ and $N$ is the additional number of users in the system, then the additional number of users in the system, $N$, can be calculated as:

$$N = M \times \frac{C}{\mu}$$

where $C$ is the capacity of the basestation and $\mu$ is the rate at which the video is transmitted.
Every $p_t$ amount of time, the buffer is empty

Num intervals $= T/b_t + p_t$

Figure 19: No. of stalls in an overloaded system without soft admission control.

$C = M\lambda$ and $O = N/M$. Then,

$$\mu = \frac{C}{M + N} = \frac{M\lambda}{M(1 + O)} = \frac{\lambda}{1 + O} \quad (9)$$

Now, the buffering time $b_t$ for $B$ bytes is given by

$$b_t = \frac{B}{\mu} = \frac{B}{\lambda}(1 + O) \quad (10)$$

The playout rate of the video for a user is $\lambda$. Since, the playout rate is higher than the service rate per user (Eq. 9), each user will receive packets for ($b_t + p_t$) amount of time and play for $p_t$ amount of time. This is depicted in Figure 19. Hence, for an interval of ($b_t + p_t$),

$$\mu. (b_t + p_t) = \lambda p_t$$

Using Equation 9 and simplifying, the above equation leads to

$$\frac{\lambda}{1 + O}. (b_t + p_t) = \lambda p_t$$

$$\Rightarrow p_t = \frac{b_t}{O}$$

$$\Rightarrow p_t = \frac{B(1 + O)}{\lambda O} \quad (11)$$

Now, in the time interval ($b_t + p_t$), every user experiences a stall of $b_t$ time for buffering. Hence the total number of stalls experienced per user in time $T$ is,

$$b_n = \frac{T}{(b_t + p_t)}$$

which simplifies using the expressions for $b_t$ and $p_t$ to

$$b_n = \frac{T O \lambda}{B(1 + O)^2} \quad (12)$$