Mobile Data Offloading: How Much Can WiFi Deliver?

Kyunghan Lee, Injong Rhee
Department of Computer Science
North Carolina State University
Raleigh, NC 27695, USA
{klee8, rhee}@ncsu.edu

Joohyun Lee, Song Chong, Yung Yi
Department of Electrical Engineering
KAIST
Daejeon, Korea
jhlee@netsys.kaist.ac.kr,
{songchong,yiyung}@kaist.edu

ABSTRACT

This paper presents a quantitative study on the performance of 3G mobile data offloading through WiFi networks. We recruited about 100 iPhone users from metropolitan areas and collected statistics on their WiFi connectivity during about a two and half week period in February 2010. Our trace-driven simulation using the acquired traces indicates that WiFi already offloads about 65% of the total mobile data traffic and saves 55% of battery power without using any delayed transmission. If data transfers can be delayed with some deadline until users enter a WiFi zone, substantial gains can be achieved only when the deadline is fairly larger than tens of minutes. With 100 second delays, the achievable gain is less than only 2-3%. But with 1 hour or longer deadline, traffic and energy saving gains increase beyond 29% and 20%, respectively. These results are in stark contrast to the substantial gain (20 to 33%) reported by the existing work even for 100 second delayed transmission using traces taken from transit buses or war-driving. The major performance difference comes from traces: while bus and war-driving traces contain much shorter connection and inter-connection times, our traces reflects the daily mobility patterns of average users more accurately.

1. INTRODUCTION

Mobile data traffic is growing at an unprecedented rate. Many researchers from networking and financial sectors [2, 3, 13, 17] forecast that by 2014, an average broadband mobile user will consume 7GB of traffic per month which is 5.4 times more than today's average user consumes per month, and the total mobile data traffic throughout the world will reach about 3.6 exabytes per month, 39 times increase from 2009 at a compound annual rate of 108%. It is also predicted that about 66% of this traffic is mobile video data. The main drive behind this explosive growth is the increase in smart mobile devices that offer ubiquitous Internet access and diverse multimedia authoring and playback capabilities.

There are several solutions to this explosive traffic growth problem. The first is to scale the network capacity by building out more cell towers and base stations of smaller cell sizes (e.g., picocell, femtocell) or upgrading the network to the next generation networks such as LTE (Long Term Evolution) and WiMax. However, this is not a winning strategy especially under a flat price structure where revenue is independent of data usage. It is interesting to note that most of these data consumptions come from a small percentage of mobile users: while smartphone users constitute about 3% of the total users in AT&T, they consume about 40% of the network traffic as of the end of 2009 [13]. Besides, expanding the network capacity may even exacerbate the problem by encouraging more data usages since the first deployment of the 4G networks is likely targeting the densely populated metropolitan areas like Manhattan or San Francisco. The second is to adopt a usage based price plan which limits heavy data usages. While price restructuring is rather inevitable, pure usage based plans are likely to backfire by singling out a particular sector of user groups, e.g., smartphone users, which have the highest potential for future revenue growth.

WiFi offloading seems the most viable solution at the moment. Building more WiFi hot spots is significantly cheaper than network upgrades and build-out. Many users are also installing their own WiFi APs at homes and work. If a majority of data traffic is redirected through WiFi networks, carriers can accommodate the traffic growth only at a far lower cost. Given that there are already a wide-spread deployment of WiFi networks, WiFi offloading addresses the "time-to-capacity" issue for the currently pressing need of additional network capacity.

There are two types of offloading: *on-the-spot* and *delayed*. On-the-spot offloading is to use spontaneous connectivity to WiFi and transfer data on the spot; when users move out of the WiFi coverage, they discontinue the offloading and all the unfinished transfers are transmitted through cellular networks. Most of the smart-phones which give priority to WiFi over the cellular interface in data transmissions

ACM CoNEXT 2010, November 30 – December 3 2010, Philadelphia, USA.

can be expected to currently achieve *on-the-spot* offloading. In delayed offloading, each data transfer is associated with a deadline and as users come in and out of WiFi coverage areas, it repeatedly resumes data transfer until the transfer is complete. If the data transfer does not finish within its deadline, cellular networks finally complete the transfer.

Most smartphones with WiFi are already performing on-the-spot offloading by default. But delayed offloading is relatively new. Its notion is very close to that of delay-tolerant networks where applications can tolerate some amount of delays. Our philosophy is that many data transfers can tolerate delays. It is true that users want to have data immediately. But if network carriers provide more incentives in price for users to use transfers with longer deadlines, it will create demands for them because users will select more judiciously their transfer deadlines based on their own needs. Several usage scenarios are possible. (1) Alice records video of a family outing at a park using her cell phone and wants to archive it in her data storage in the Internet. She does not need the video immediately until she comes home after a few hours. (2) Bob wants to email Alice a roll of pictures that he took last week, but it does not have to be available immediately and besides the carrier charges less if he opts to have it delivered within thirty minutes. (3) Bob is traveling this afternoon from New York to Los Angeles and he just realizes that he can use some entertainment during the long flight. As he has several hours before the trip, he schedules to download a couple of movies on his cell phones. (4) Alice wanted to download an e-book on her iPad, but as the carrier charges less for a two-hour download than immediate download, she opted for that service. But she found out later the book was actually delivered in thirty minutes as she stopped by a coffee shop that provided a free WiFi connectivity. The first scenario is in fact currently implemented in the Urban Tomography project [1].

There is no doubt that both on-the-spot and delayed offloading reduce the load on 3G networks. But an important, yet under-addressed question is how much benefits offloading can bring to network providers and users. Network carriers are interested in knowing how much traffic load WiFi offloading takes away from cellular networks under a given or future WiFi network deployment. On-the-spot offloading is currently being offered through smartphones. Since carriers do not have control over WiFi networks that users connect to, they have no idea how much on-the-spot offloading helps them even now letting along the future. How much does the new notion of delayed offloading help reduce their traffic given the projected amount of data growth in the future? The answers to these questions can provide clues on their price and cost restructuring strategies. Users are also interested in offloading because of economic reasons, e.g., a potential decrease of subscription fees or better service with the same fees. The average delays of offloaded data are also important to users. If they can predict in advance how long the actual data transfers will take on average based on their own mobility patterns, they can use that information in choosing the right price and deadlines for their transfer services. Users are also interested in actual energy saving that delayed offloading can achieve. All the above questions are fundamentally tied to the mobility patterns of users as users may come in and out of WiFi coverage. In this paper, we offer rough and rule-of-thumb answers to these questions.

There have been several recent studies [6,7,15,16] on the related topic. Some [7,15,16] have studied in the context of energy saving with assumption that data can tolerate a delay of one minute to a few hours and the other [6,15] in the context of on-the-spot or short delayed (up to 100 seconds) offloading. None of them have looked at the benefits of the full scale delayed offloading. Most important, the data sets used in these studies are highly limited. In [6], the authors use several traces of a war driving around a city using their own vehicles and also 20 city transit buses. These data sets are very limited for answering our questions as they do not account for the *temporal coverage* of actual users in their daily lives (i.e., they do not necessarily ride buses or cars all the time) and their characteristics, e.g., how often and long users enter and leave a WiFi zone and what data rate they experience when they stay in a zone. Their results are meaningful only if mobile data are generated in a city transit bus or in their war driving scenarios. The authors report about 10 to 30% of the total traffic can be offloaded using on-the-spot offloading and with up to 100 second delays, delayed offloading can achieve about 20 to 33% additional gains over on-the-spot offloading. In [16], the authors study energy saving efficiency using a set of walk traces, each walk taking a few hours with an instrumented mobile device. For our study, this data is of limited use because each trace is too short to account for the daily life patterns of users. More details on related work can be found in Section 4.

We offer, to the best of our knowledge, the first quantitative answers to some of these questions by conducting an extensive measurement study in South Korea. For our measurement study, we first designed and implemented an iPhone application that tracks WiFi connectivity. We recruited about 100 iPhone users from the Internt who downloaded our application to their phones and used it for about a two and half week period in February 2010. About 55% of the users live in Seoul and the others in the other major cities in Korea. None of the users, to our knowledge, are related to the authors. We briefed the users about the types of the measured data and their objectives. The phone is configured to connect to various WiFi networks as the users travel including its carrier's WiFi network. The application runs in the background to record the locations of WiFi stations to which each user connects, the connection times and durations, and the data transfer rates between WiFi stations and mobile phones, and then periodically upload the recorded data to our server. These data are used to carry out trace-driven simulation of offloading with diverse data traffic and WiFi deployment scenarios.

From our data, we find that users are in a WiFi coverage zone for 70% of their time on average (63% during the day time). They stay in a coverage area for about 2 hours on average, and after leaving the area, they return to an WiFi area within 40 minutes (this time interval is called *inter-connection times*). The distributions of these statistics have a strong heavy-tail tendency. Data rates from the phone to our measurement server in the Internet are about 1.26 Mbps on average during the daytime and 2.76 Mbps the nighttime. The full analysis is presented in Section 2.2.

Using the data traces we obtained from the experiments, we run a trace-driven simulation to measure the efficiency of on-the-spot and delayed offloading. Our simulation uses the measured data rates from our traces and each data transfer by a user in a WiFi zone is assumed to run at the actual transfer rate experienced by the user in our trace. This ignores the effect of changed load (e.g., contention) on the network bandwidth in the future. The same simulation strategy is used in [6]. The results below must be interpreted as upper-bounds if the carriers can sustain the measured data rates through additional WiFi resource provisioning in the future.

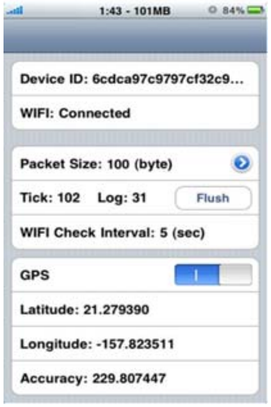
The followings are the key findings from our simulation.

1. On-the-spot offloading can offload about 65% of the total traffic load. This is achieved without using any delayed transfer. When delayed offloading is used with 100 second delay deadlines, the achievable gain over on-the-spot is very insignificant: 2-5%. This result is in stark contrast to the result from [6] which reports 10 times bigger gains with the same deadline. Our analysis indicates that in order for delayed offloading to get significant gains, the deadline must be much longer than 100 seconds because of long inter-connection times. When data transfers are opted by users for delayed transfers with a deadline of one hour and longer, the gain over on-the-spot becomes larger than about 29%.

2. On-the-spot offloading alone (without any delayed transfer) can achieve about 55% energy saving for mobile devices because WiFi offloading can reduce the transmission time of mobile devices substantially. However, for delayed transfers with very short deadlines like 100 seconds, the achievable energy saving gain over on-the-spot offloading is highly limited to about 3%. But with one hour delay, the achievable energy saving gain increase to around 20%.

3. For a prediction-based offloading strategy like Breadcrumbs [6,15] to be useful, it has to predict over several tens of minutes since the inter-connection time has an average of 40 minutes. Because of the heavy-tail tendency of the interconnection times, this prediction will be even harder.

4. The average completion time of data transfers is much shorter than their delay deadlines. While on-the-spot offloading obviously achieves faster transfer than using 3G networks only, it is surprising that video file transfers of size larger than 30 MB with one-hour deadline are consistently faster than no offloading. Furthermore, the 3G network usage reduction gain of these transfers is more than 50% over on-the-spot offloading and more than 80% over no offloading, which implies 50% or more cost reduction for the carriers to deliver such transfers and translates directly into price reductions for users.


1:43 - 101MB 84%

Device ID: 6cdca97c9797cf32c9...

WIFI: Connected

Packet Size: 100 (byte)

Tick: 102 Log: 31

WIFI Check Interval: 5 (sec)

GPS

Latitude: 21.279390

Longitude: -157.823511

Accuracy: 229.807447

Figure 1: An iPhone App, DTap for measuring WiFi availability.

More detailed analysis of our simulation can be found in Sections 3. We present some discussions on the limitations of our work and future work in Section 5.

2. MEASUREMENT STUDY

2.1 Experimental Setup

The performance of offloading highly depends on the patterns of WiFi coverage and user mobility. Accurate modeling of offloading performance calls for a measurement study. We first develop an Apple's iPhone application, called *DTap* (Delay Tolerant APplication) that records the statistics of WiFi connectivity in the background and periodically sends the recorded statistics to a server (see Figure 1 for a screenshot). Running in background, DTap scans for WiFi connectivity at every three minute interval. As scanning for WiFi, iPhone connects to the AP, if any, with the strongest signal strength among those to which it has a past history of connections. Note that the captured WiFi APs include the private APs at home and work, commercial APs installed by the carrier of the mobile phone and the third party companies (e.g., Boingo). As our participants are actively using their iPhones, these APs are mostly included in their past histories. After connecting to a WiFi network, it measures data throughput and round trip times by pinging the server with a 100 byte packet ten times. This measures the end-to-end data rate between the client phone and our server. This is obviously not the most accurate measurement method but it is reasonable under the constraint that DTap should minimally consume the bandwidth as well as the battery of a

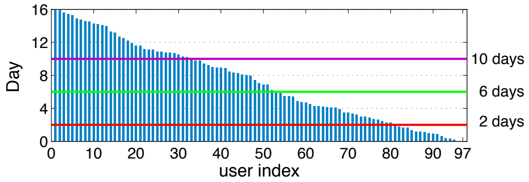


Figure 2: Number of valid experimental days for each user.

user. DTap records in a log file the GPS location where the connection occurs, and the duration, data rate and time of the connection.

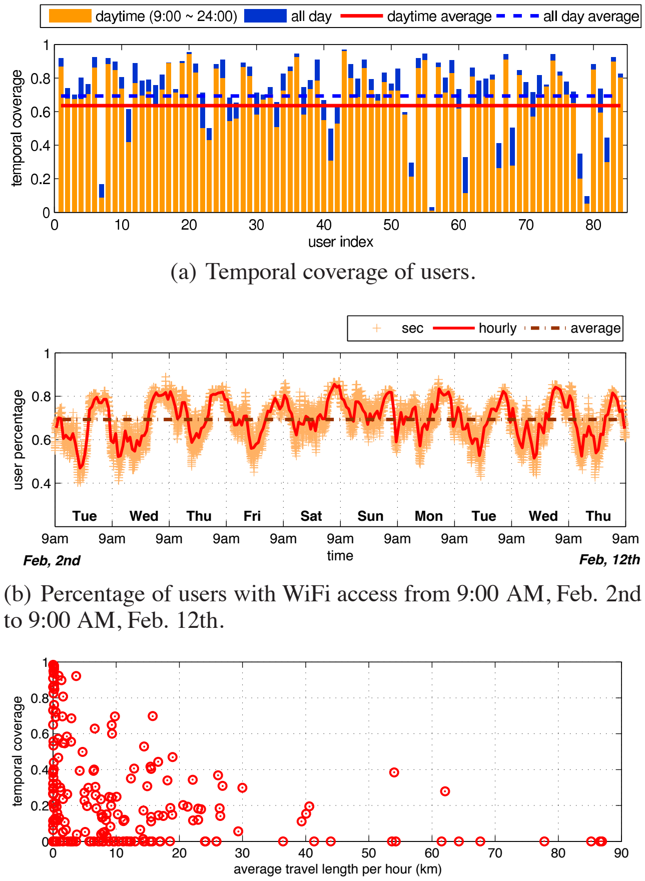
DTap does not perform offloading. This is because performing offloading directly in client's phones for arbitrarily generated data drains too much battery power, which faces strong resistance from volunteers. Even with just the WiFi scanning and pinging tests, the phone drains the battery power very quickly. Instead, we take an approach of collecting detailed traces of WiFi connectivity for a fixed period of time and later using the traces to simulate offloading under diverse traffic patterns.

The log files are uploaded to our server using ftp connections daily between 4:00 AM to 4:30 AM. The daily log file size is typically less than 1 MB. DTap runs with customized parameters which are contained in an XML configuration file automatically updated to client phones daily. Each row of the log file contains the following tuple: *(device id, time stamp, event name, field 1, ... field n)*. The device id is the unique id of a phone. The time stamp is the time when the corresponding tuple is recorded. Multiple event names are used in our experiment, depending on which, the number and the values of associated fields are decided. They are summarized in Table 1. The AP list represents all the APs that the phone can currently detect. GPS location is associated with the location accuracy information provided by the phone.

Table 1: Event names and associated fields in the log file

Event Name	Associated Fields
WiFi connectivity	0 or 1
AP Lists	SSID1, . . . , SSIDn
GPS	latitude, longitude, accuracy
Data rate	rate

We recruited 97 volunteers who own iPhone 3G/3GS from an iPhone user community in Korea and asked them to install DTap in their phones for a period of 18 days in January and February 2010. The volunteers come from diverse occupational backgrounds and various major cities in Korea (60% from Seoul). For data integrity, we have excluded a very small number of daily traces which show no movement (as users might have forgotten to carry their phones). Figure 2 shows the number of experimental days for each participant. The total number of valid daily traces we collect is 705.



(a) Temporal coverage of users.

(b) Percentage of users with WiFi access from 9:00 AM, Feb. 2nd to 9:00 AM, Feb. 12th.

(c) Temporal coverage for hourly mobility

Figure 3: Temporal coverages per user, time and hourly mobility

2.2 Key Observations

We measure the following statistics relevant to offloading: the total time duration of WiFi connectivity, the data rate during connections, the distributions of connection times and inter-connection times and the correlations of the total travel lengths with the data rate and time of WiFi connectivity time.

Temporal coverage

The performance of offloading highly depends on the time duration that a user stays in a WiFi coverage area which is defined as *temporal coverage*. Figure 3(a) shows the daily average temporal coverage recorded by each participant. It also plots the coverage recorded during the daytime. The averages across all the users are 70% for all day and 63% for the daytime only. Difference between all day and day time averages arises because most participants are likely to have WiFi connectivity at home. Figure 3(b) shows the percentage of users that have WiFi connectivity at any given time averaged over one second and hour periods respectively. It indicates that at any time, about 70% of users stay in a WiFi

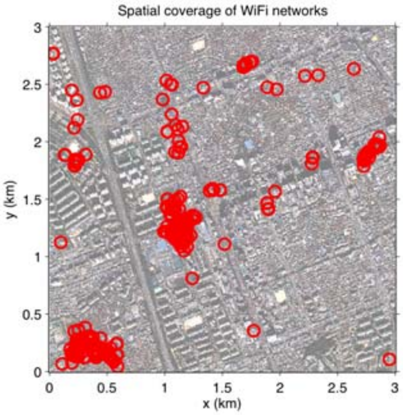


Figure 4: Locations of WiFi APs detected by the participants in a 3km by 3km area with the most dense WiFi deployment inside Seoul (more visible in color).

coverage area.

There is a substantial difference between the data from [6] that reports only 11% temporal coverage. This difference comes from the fact that their measurements are done using war-driving or in transit buses. They do not account for the natural mobility of users and their natural sojourn times in a WiFi zone. Typically users spend most of time in office and home. This type of information is missing as average users are not likely to spend most of their time only inside a car or bus. To verify our conjecture, we also record the traveling distances of each user for each hour. This can be calculated as the log contains GPS data. We map the temporal coverage during each hour to the travel distance that the user makes during that time period. Figure 3(c) shows the results. The results indicate that users with high mobility (i.e., including those moving in a car) have very low temporal coverage.

We measure *spatial coverage* which is defined to be the fraction of an area that is under any WiFi coverage. Our traces give only a rough estimation of spatial coverage since they do not capture all possible WiFi APs located in the city because the walkabouts of participants do not cover the whole area. But it certainly gives a lower bound. Figure 4 shows the locations of WiFi APs that the users visit in a 3 km by 3 km area of the city where the users visit most. We measure the spatial coverage by drawing 50 m radius circles, a typical WiFi range, around each WiFi detected AP and totaling the areas of the drawn circles. Our analysis shows that the spatial coverage is about 8.3% (20.6% for 100m radius circles).

Our data shows that the temporal coverage is about 3.5∼8 times larger than the spatial coverage for a given region, indicating that most users stay inside a WiFi network for a long time once they connect to a WiFi network. Figure 5 shows the CCDF (Complementary Cumulative Density Function) of the stay time (called *connection times*). The average connection times is about 2 hours for all day and 52 minutes for daytime only. Figure 6 shows the CCDF of *inter-connection times*, the time duration after a user leaves a coverage area, until it returns to a coverage area. The average is about 40 minutes for all day and 25 minutes for daytime. An interesting observation from our trace is that both CCDFs show a heavy-tailed tendency and, in particular, fit very well with Weibull distributions[1] using MLE (Maximum Likelihood Estimation). The Weibull distribution has two parameters k and α in its PDF (Probability Density Function) and when k is less than one, the distribution is heavy-tailed and as k gets smaller, it becomes more heavy-tailed. The measured statistics fit very well with $k = 0.31 \sim 0.52$ for connection and inter-connection times. It is interesting to see that the inter-connection time distribution shows similar pat-

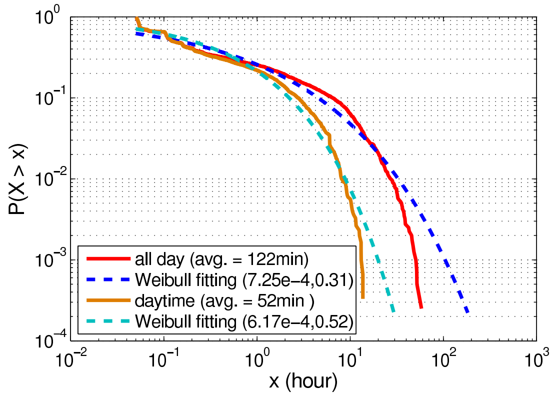


Figure 5: The CCDF of connection duration. The average connection duration is 122 minutes. The distribution fits well with a Weibull distribution with $k = 0.31$ for all day and $k = 0.52$ for daytime. α parameter is also given in the bracket.

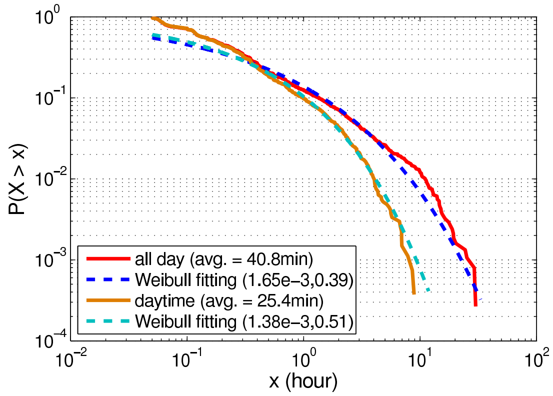


Figure 6: The CCDF of inter-connection times. The average is 41 minutes. The distribution fits well with a Weibull distribution with $k = 0.39$ for all day and $k = 0.51$ for daytime. α parameter is also given in the bracket.

[1] The PDF of the Weibull distribution with the parameters α and k is $\frac{k}{\alpha}(\frac{x}{\lambda})^{k-1}\exp[-(\frac{x}{\alpha})^k]$.

tern to the inter-contact time distribution observed from human mobility which is known to be heavy-tailed [9,14,18].

The heavy-tail tendency of inter-connection times with a large average (25 to 40 minutes) implies that the prediction-based offloading strategies like Breadcrumbs [6,15] may not be so effective. These strategies use past history of user mobility and predict whether users will be entering a WiFi zone with fast transmission rates within a given deadline. Typically these algorithms use 100 seconds for look-ahead times. However, the large average value of inter-connection times requires these algorithms to look ahead farther beyond 100 seconds. Furthermore, because of the heavy-tail tendency of inter-connection times, this prediction may not be so accurate.

End-to-end rates

Figure 7(a) shows the cumulative distributions of end-to-end data rates reported by users. It shows a variety of experienced rates with their average being around 1.97Mbps. This average is highly skewed by the data rate during the night time. Figure 7(b) shows the data rates averaged across users for each time period. It can be seen that the high data rates during the night time are around 2.76Mbps and on average, users are experiencing around 1.26Mbps during the daytime. During the night time, users are likely connecting to their home APs. This data shows that offloading during the night time is going to be very effective if users can tolerate large delays. We also map the measured data rate per hour for each user to their hourly traveling distance (Figure 3). It is shown that user mobility has weaker correlation with data rate than temporal coverage.

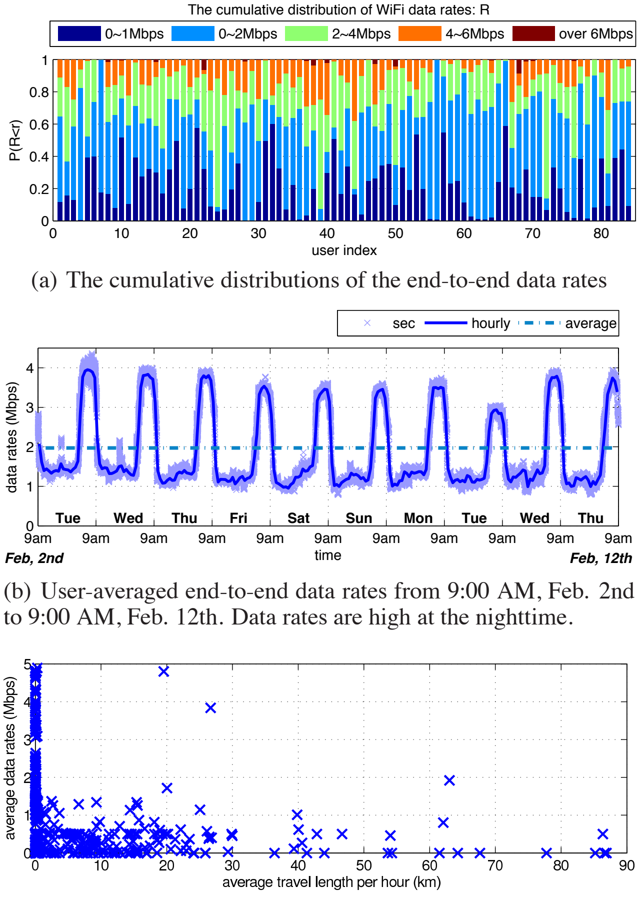
3. OFFLOADING EFFICIENCY

In this section, we report the simulation results using the traces we discussed in Section 2. Since we have detailed records of user connectivity and data rates during the connectivity for every three minute interval, they can be used to simulate the offloading of input traffic with diverse patterns.

3.1 Simulation Method

For each user, we generate input data traffic with specific arrival and size patterns. A data file for uploads arrives during typical active hours (9:00 $\sim$ 24:00) to the phone of the user with a random inter-arrival time and a random size selected from input distributions (typically exponential or Weibull) of a mean a for inter-arrival times and a mean b for file sizes. We say that b/a is *traffic intensity*. Each file is associated with a deadline typically assigned by its user or application program. Upon arrival, each file is scheduled for uploads in a FIFO manner. The transmission time of a data transfer is determined by the measured data rate experienced by the user at the time of the transfer (this is taken from the user log trace). If the transfer cannot be finished before its deadline, the file is assumed to be uploaded through 3G networks and is simply removed from the queue. We develop a MATLAB simulator which follows a simulation model depicted in Figure 8.

We define *offloading efficiency* to be the total bytes trans-



(a) The cumulative distributions of the end-to-end data rates

(b) User-averaged end-to-end data rates from 9:00 AM, Feb. 2nd to 9:00 AM, Feb. 12th. Data rates are high at the nighttime.

(c) Per-hour mobility vs. Data rate

Figure 7: End-to-end data rates per user, day and hourly mobility.

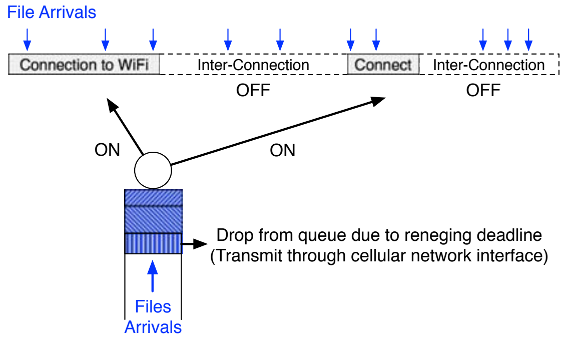


Figure 8: A simulation model of a user. The data in the user queue is serviced only when a mobile phone is connected to a WiFi network. When the delay deadline of the file in the queue expires, the file is removed from the queue and uploaded through 3G networks.

Table 2: Input data to the experiment for Figure 9. We use the projection from [2] on the amount of mobile data traffic, their constituent types and proportion mobile data traffic in year 2014. We assign artificial deadlines to different types of data from short to long deadlines. The mean inter-arrival times are estimated from the estimated monthly volumes. DL : Deadline.

	Video	Data	P2P	Audio (VoIP)	Total
Ratio [2]	64.0 %	18.3 %	10.6 %	7.1 %	100 %
Data/month	4.48 GB	1.28 GB	740 MB	500 MB	7 GB
Avg. IAT	1 hour	2 hours	2 hours	1 hour	-
Traffic vol.	10 MB	5.7 MB	3.3 MB	1.1 MB	-
Traffic dist.	Weibull (k=0.5)	←	←	Exponential	-
On-the-spot	0 sec.	0 sec.	0 sec.	0 sec.	-
DL:short	30 min.	30 min.	0 sec.	0 sec.	-
DL:medium	1 hour	1 hour	0 sec.	0 sec.	-
DL:long	6 hours	6 hours	0 sec.	0 sec.	-

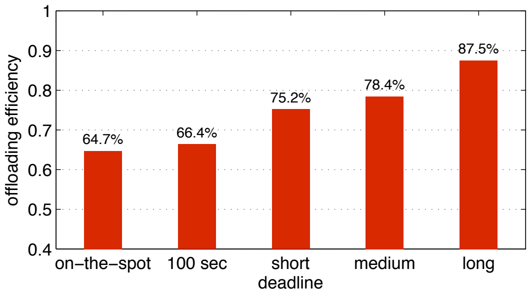


Figure 9: Offloading efficiency of delayed transfers with various deadlines when all the transfers are opted for delayed transfers.

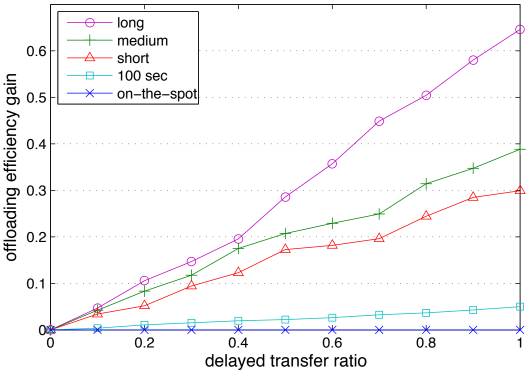


Figure 10: Offloading efficiency gains over on-the-spot offloading achieved by delayed transfers as the delayed transfer ratio varies.

ferred through WiFi divided by the total bytes generated.

3.2 Traffic Model

To understand the impact of offloading in relieving the future traffic demands, we use the projection data released from CISCO [2] on the amount of mobile traffic demands by year 2014. It is predicted that an average user consumes about 7GB per month and the contribution of various data types to this traffic is summarized in Table 2. We assign three different types of offloading deadlines to each data type from short to long deadlines. The short deadline is 30 minutes while long deadline is 6 hours. Note that as we shall see later in Section 3.4, most transfers finish well before their deadlines. We assume that the inter-arrival time distribution is exponential and the distributions of the arrival traffic volumes of video, text and data are Weibull and that of the audio (VoIP) data is exponential. The means of all the distributions are deduced from the estimated monthly traffic of each type. Audio and P2P (e.g., data sharing in the proximity) data are assumed to be not delay-tolerant so zero delay deadlines are assigned to them.

3.3 3G Network Traffic Reduction

In this section, we measure the amount of traffic offloading to WiFi from 3G networks. Figure 9 shows the offloading efficiency of on-the-spot and delayed offloading. We added the results using 100 second delay deadlines for comparison with the results in [6]. In this experiment, we assume that all transfers of video and data use delayed transfers. It is surprising that on-the-spot offloading (without any delays) can achieve extremely high offloading efficiency already. Note that on-the-spot offloading is what is currently being performed by smartphones today. If most of mobile data volume comes from smartphones, WiFi can offload more than 65% of traffic even today. This is much larger than 10 to 30% on-the-spot offloading efficiency reported by [6]. This difference is because average users in our traces spend much more time in WiFi zones than those from bus or war-driving traces in [6] as we discussed it in Section 2.2.

As we increase delay deadlines, offloading efficiency increases substantially. For long deadlines, the efficiency increases to 88% indicating most of mobile data can be offloaded to WiFi. With 100 second or less deadlines used in [6], the additional gain of delayed transfers over on-the-spot is only 5%. That is substantially smaller than 20 to 33% gain reported by [6]. This difference also comes from that their traces contain much shorter inter-connection times as buses and cars travel much faster than average users on the street or offices. To have substantial gain using short delays, users must need to experience very short inter-connection times (as if they are in a car).

The offloading efficiency of 88% for long deadlines is certainly unrealistic. It is not true that all transfers of video and data in Table 2 are opted for delayed transfers with such a long delay (6 hours). It is possible that despite pricing incentives, users may opt for on-the-spot offloading only. To see the effect of this, we measure the performance as the ratio of delayed transfers over the total data traffic (called *delayed transfer ratio*) is varied. In this plot, we are interested in the gain achieved by delayed transfers over on-the-spot. Again the gain achieved by 100 second deadlines is very minimal

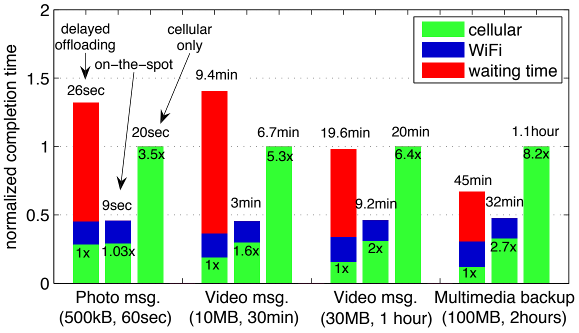


Figure 11: Comparison of average completion times of offloading methods for various types of applications normalized to the time taken using only cellular networks. Parameters in the bracket show the size of files and applied deadlines (more visible in color).

from 2 to 3% for 30 to 50% delayed transfer ratios. The gain for one hour deadline is about 13 to 21% with 30 to 50% delayed transfer. This result indicates that since on-the-spot offloading is already very good, for delayed transfers to achieve substantial gain, their deadlines must be fairly long (e.g., longer than several tens of minutes).

3.4 Completion Time

Deadlines of 30 minutes or one hour seem very long for some applications. However, our results indicate that most transfers finish well before these deadlines. In Figure 11, we measure the average completion times of transfers with various traffic types. For this experiment, we set the data rate of 3G networks to 200 Kbps which we typically get through our iPhones for uplink. We measure the completion times for (a) delayed offloading, (b) on-the-spot offloading and (c) no offloading (3G network only). The result in each traffic type is normalized by the completion time of no offloading. Photo messages with 60 second deadlines finish in 26 seconds on average, 6 seconds more than no offloading. The break-even point where the completion time of delayed offloading becomes the same as that of no offloading, occurs when video messages with 30 MB of one hour deadline are transmitted. When that happens, the amount of 3G network usage of delayed offloading is half of that of on-the-spot. At this point, users using delayed offloading may experience the same delay as no offloading while the cost of delivery by the carriers is only half of that of on-the-spot and about 20% of no offloading. This happens because delayed offloading delays its transfer until it has a WiFi connectivity. Since WiFi offers higher data rate, more use of WiFi leads to shorter completion time. Although delayed offloading has a longer completion time than on-the-spot offloading, it uses 3G network far less, which is translated into cost reduction for carriers and price reduction for users. With larger file sizes or longer deadlines, delayed offloading achieve faster completion time and more cost reduction than no offloading.

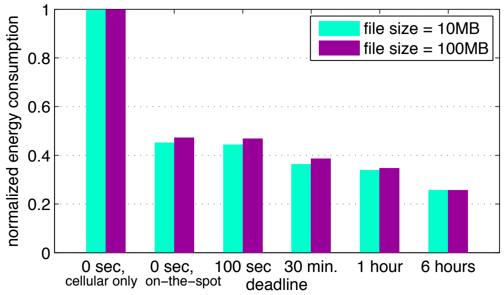


Figure 12: Normalized energy consumption of delayed transfers of 10MB and 100MB files with one hour deadline. File sizes and intervals are assumed to be exponentially distributed.

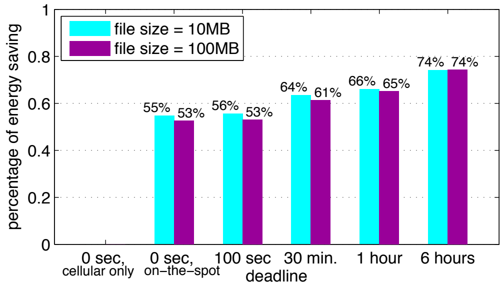


Figure 13: The amount of energy saving gain over no offloading for 1 hour delayed transfers. File sizes and intervals are assumed to be exponentially distributed.

3.5 Energy Saving

There exists a fundamental trade-off between energy consumption and delay in smartphone applications as smartphones have multiple radio interfaces with different transmission rates and availability [15,16]. While 3G networks are more widely available than WiFi, their data rates are much less than WiFi. Therefore, by delaying transmission until WiFi is available, there are opportunities to reduce the transmission time as we have seen it in Section 3.4. The reduced transmission times are directly translated into battery power saving for smartphones because energy consumption of WiFi per second is almost the same as 3G networks [19]. The transmission time of a transfer is different from its completion time as transmission times account for only the time that radio interfaces are used to complete the transfer. Thus, the transmission time is the time after subtracting the waiting time from the completion time in Figure 11. We assume that power consumption during the waiting time is negligible using smart WiFi perception technology [5,10,20]

Figures 12 and 13 examine power consumed for delayed transfers of 10 MB and 100 MB files with various delay deadlines which is directly translated from their transmission times. The values are normalized to the energy consumption of no offloading. On-the-spot offloading already achieves 55% energy saving over no offloading because of reduced transmission time through use of WiFi. However, in order for delayed transfers to achieve substantial energy saving gain over on-the-spot offloading, the deadlines must be substantially long. With 100 second deadlines, the saving gain over on-the-spot is extremely limited. One hour deadlines achieve about 20% gain.

3.6 Impact of Traffic Types

In this section, we further evaluate the detailed impact of varying input traffic characteristics to offloading efficiency. We especially focus on the inter-arrival time distributions of input files and and file size distributions. To test diverse traffic types, we first vary traffic intensity. For instance, traffic intensities of text and video messages would be different. For the same traffic intensity, we also vary traffic burstness (i.e., 1/ inter-arrival time or simply the number of files generated per unit time) and the file size distributions. We test deterministic, exponential, and heavy-tailed file size distributions. In our simulation, we conducted simulations for the traffic intensities 0.1, 50, 500 and 5000 KB/min. For each traffic intensity, we test two different cases of traffic burstness and file size.

Offloading efficiency for less bursty traffic, shown in the upper three plots of Figure 14, uniformly ranges from 0.7 to 1 depending on the stringency of delay deadline, but irrespective of average file sizes and file size distributions. Specifically, for the traffic generated with the average rate of 5MB per minute, even just 2 hours of delay tolerance enables us to offload about 80% of data traffic from the current cellular network. This clearly shows a benefit of a combination of delay-tolerance and user mobility which increase the total system capacity significantly. It is intuitive that more bursty traffic induces lower offloading efficiency. The bottom three plots of Figure 14 show such a case that the files are generated every hour (thus, for the same traffic intensity, a file with larger size is generated), where a slight decrease of offloading efficiency is observed. However, such a decrease is visible only for short deadlines, and for long deadlines, the performance difference is not considerable. Note that heavy-tailed inter-arrival distributions are reported to appropriately model the time interval between consecutive e-mails [8]. The bottom plots of Figure 14 show the performance for the case.

It is known that applications often generates traffic whose file-size distributions are heavy-tailed in many cases. See [4] for the video file size distributions in YouTube. Intuitively, more heavy-tailed traffic leads to lower offloading efficiency since file size far larger than the mean can be generated with non-negligible probability. Figure 15 depicts the offloading efficiency for a varying heavy-tail degree in the file size distribution controlled by the k value of Weibull distribution with the mean set to 100 MB. The inter-arrival times have an exponential distribution with one hour average. Recall that smaller $k < 1$ generate more heavy-tailed traffic, and when $k = 1$, it boils down to the exponential distribution. We observe that even for very heavy-tailed traffic, the offloading efficiency is at least 20%, and over 40% of files with two hour deadlines can be offloaded through WiFi except the extreme case, $k = 0.1$.

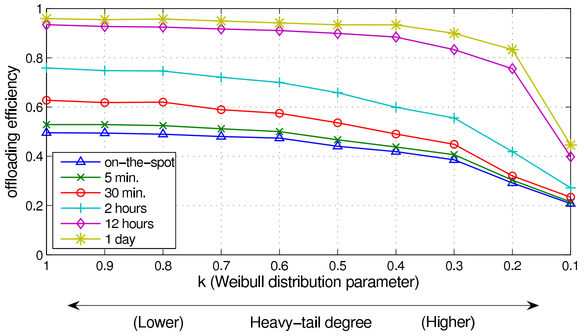


Figure 15: Offloading efficiency varying k parameter of Weibull distribution. Mean file size is 100MB and the traffic generation interval is 1 hour. When $k = 1$, Weibull distribution is exponential. As k decreases, it is more likely that a huge file arrives at the system.

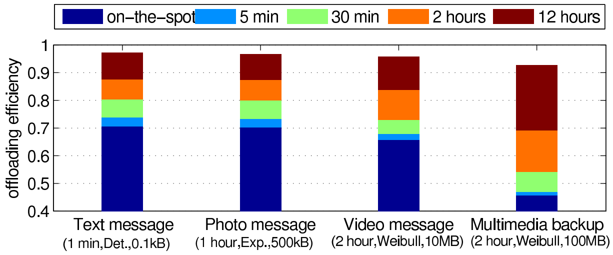


Figure 16: Offloading efficiency of candidate applications (text, photo, video-clip messaging and multimedia backup). Inter-arrival time, type of file size distribution (Det: deterministic, Exp: exponential), mean file size are denoted in the bracket.

To get more realistic offloading efficiency, we set the input parameters of various application data considering the property of the data. We set the inter-arrival time of 0.1 KB text messages to 1 min constant, 500 KB photo messages to an exponential distribution with a mean 60 minutes, 10 MB video messages to a Weibull distribution with a mean 120 minutes and $k = 0.5$, and 100 MB of multimedia backup to a Weibull distribution with a mean 120 minutes and $k = 0.5$. Figure 16 shows the result. Text and photo messages can be offloaded instantly at the rate of 70%. Video messages and multimedia backup can be offloaded around 70% with deadlines of thirty minutes and 2 hours, respectively.

3.7 Impact of WiFi Deployment

We investigate the impact of WiFi density and deployment strategies on offloading efficiency. To test them, we use the current deployment observed in our traces as a baseline, and thin out density by progressively eliminating WiFi APs according to two different strategies: *activity-based* and *random*. In the activity-based strategy, we measure the *connection time of an AP* which is the sum of time duration that

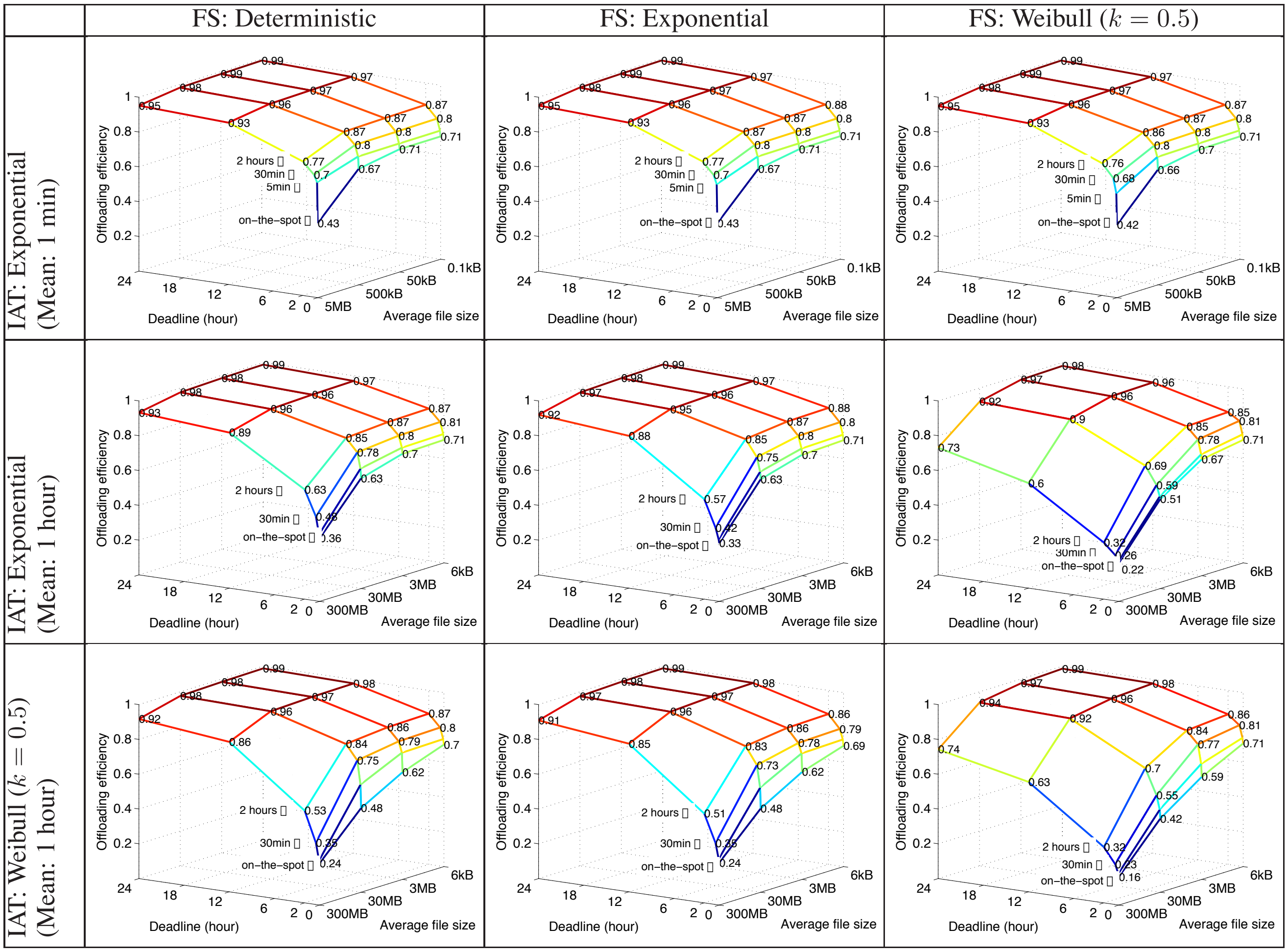


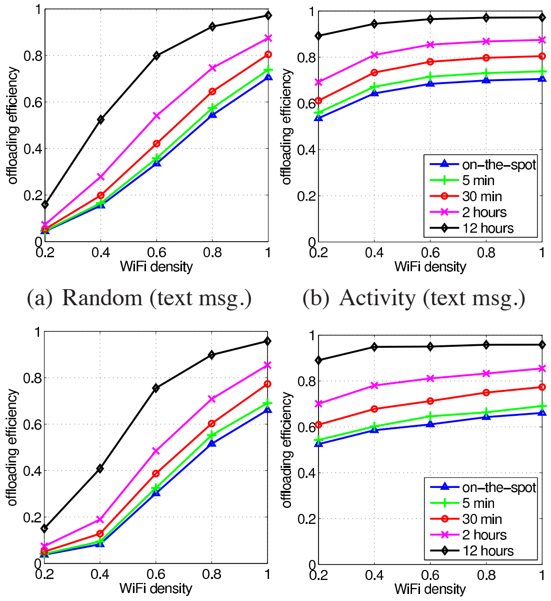
Figure 14: Offloading efficiency for different traffic intensity, file size distributions, and delay deadlines. IAT (Inter-Arrival Time) follows exponential or Weibull whose mean in the bracket determines traffic burstness. FS (File Size) follows deterministic, exponential or Weibull distribution whose mean is specified in the file size axis in the figures. (More visible in color)

each user spend in the coverage of the AP. The activity-based strategy eliminates WiFi APs in the increasing order of their connection times until a target density is reached.

We set the density of the current deployment measured from the trace to 1. The random strategy randomly eliminates APs with equal probability. Figure 17 shows the offloading efficiency for two considerably heterogeneous traffic types, text messaging and multimedia data backup whose traffic parameters are the same as those in Figure 11. The activity-based strategy naturally outperforms random, but it is interesting to see that even after reducing its density by half the activity-based strategy reduces offloading efficiency by only a small percentage while the random strategy has about a 50% performance drop. It implies that careful deployment plans can yield substantial improvement in the capacity even with a small increase in density. We leave the investigation of the optimal strategy of WiFi deployment for delayed offloading as future work.

4. RELATED WORK

Balasubramanian et al. [6] develop several techniques combining 3G networks and WiFi for reducing the total cost of data transfer. The proposed techniques are similar to Breadcrumb [15]. The authors use a city wide measurement data of 3G and WiFi network availability obtained from 20 transit buses in a city and war-driving in two other cities. The work focuses on gains achieved by on-the-spot offloading or delayed offloading with a very short delay deadline (up to 100 seconds). Based on these traces, they report about 10 to 30% on-the-spot offloading efficiency and about 20 to 33% offloading gain of delayed offloading over on-the-spot. Since their traces are taken during driving, they contain a lot of short connection and inter-connection times with WiFi which contribute to the substantial gains of delayed transfers with short deadlines (also low efficiency of on-the-spot offloading). However, their work is limited in reflecting the achievable gains of offloading for average users with regular



(a) Random (text msg.)
(b) Activity (text msg.)
(c) Random (video msg.)
(d) Activity (video msg.)

Figure 17: Offloading efficiency for various amount of WiFi deployment and different deployment strategies. Irrespective of traffic type, activity-based deployment which might primarily lead to install WiFi APs to users' houses shows clearly higher offloading efficiency than random.

mobility patterns. Especially, their traces ignore the effect of long sojourn times that users spend at home or offices and roaming effects through walking. These slow movements create long connection times during which on-the-spot offloading can be most effective. Furthermore, they create long inter-connection times (on average 25 minutes in our traces) which leads to the long waiting period before WiFi connectivity. Therefore, their results are more meaningful for offloading during driving but less so for studying the offloading efficiency of average users.

There are several more measurement studies focusing on WiFi and 3G network availability over given movement paths. Han et al. [12] suggest a two-pass measurement methodology involving rough search and detailed measurement phases for WiFi APs. Gass et al. [11] present a detailed measurement result by comparing the characteristics of 3G networks and WiFi in a city. Both of these results are based on wardriving by vehicles or by walk.

Ra et al. [16] present an online algorithm called SALSA over mobile smartphones with 3G/EDGE/WiFi interfaces that optimizes energy and delay trade-offs using a Lyapunov optimization framework. SALSA is tested over real 3G/EDGE/WiFi measurement performed using 66 sample walk traces of about one hour length in various areas including campus, shopping mall and airport. Balasubramanian et al. [7] present a different type of energy and delay trade-offs arising from energy the consumption characteristics of multi-modal wireless terminal equipped with WiFi, 3G and GSM mobile network technologies. Based on a measurement study, they develop a energy consumption model for each technology. The model is then used to design an algorithm that schedules (i.e., delays) transmissions to minimize the overall time spent in high energy states (i.e. energy tail) while respecting user-specified delay-tolerance deadlines.

Nicholson et al. [15] propose a scheme that can predict near future WiFi connectivity and quality. The scheme enables mobile devices to schedule their data transfers to harness higher transmission rates of WiFi APs. It exploits users' tendency of following regular movement patterns around the region where static WiFi APs are deployed. The authors show that delaying transmissions according to short-term forecasts can achieve higher data rate as well as lower power consumption.

5. DISCUSSIONS AND CAVEATS

The perhaps biggest surprise in our analysis is 65% traffic reduction currently achievable by on-the-spot WiFi offloading without use of any delay. Assuming most mobile data demands are from smartphone users, this gain is what the carriers are currently achieving. Roughly it implies that about 35% of the projected 7 GB/month per-user usage in 2014 (about 2.5 GB) will be transferred through 3G networks. With additional incentives for delayed offloading, this gain can quickly grow. This means that from user's perspective, with a fixed price plan of 2 GB/month over 3G networks (what is currently adopted by AT&T for iPhone 4G), average users do not overscribe at all. With more creative price plans for delayed transfers, users may even opt for a cheaper monthly data plan and can offload most of excess data traffic.

This paper focuses only on *temporal offloading*. However, allowing delays in applications also enables load balancing. End-to-end data rates at night are much higher than daytime because we tend to experience stable links overnight at home as well as less congestion in the backhaul network. Delayed transfers, especially with long delay deadlines is likely to enable traffic dispersion over time so as to shift the high daytime demand for networking resources to the night time.

Our study makes a number of convenient assumptions. First, we assume that the measured data rates of WiFi in our traces are sustained independent of load in the network. Although the measured data rates account for traffic conditions (e.g., contention and dynamic data rates) existing at the time of connection, we ignore the issue of increased contention in the future as more users use WiFi offloading. Measuring and predicting the exact data rates for the future is very challenging. This factor depends on the trade-off between capacity and demands offered by the current WiFi technology which is still developing, so we do not have a clear answer for how we can incorporate the impact of the increased load on the performance of WiFi offloading. However, our results are still meaningful as they can be viewed as an upper-bound on the performance gain since contention can only increase with more usage. In the other words, our results are meaningful if the carriers can provision enough WiFi resource to sustain the current WiFi data rates.

The main focus of our study is purely performance oriented. We ignore a number of technical and policy issues in our study. First, energy consumption is high if mobile devices constantly scan for WiFi connectivity. A number of solutions (e.g., [5, 10, 20]) for this problem are being developed. Rahmati and Zhong [5] design an intelligent energy-saving algorithm for predicting WiFi availability and device scans for WiFi APs only in areas where WiFi is likely available. Several researchers [10, 20] are developing an energy-efficient location tracking system for mobile phones based on map matching and war-driving or magnetometer and accelerometer sensor readings which consume only a small fraction of power used for GPS. As users tend to maintain regular mobility patterns daily, mobile phones can perform scanning only when they are in a pre-recorded area of WiFi stations.

We also do not examine issues of security and administration or billing control. As user data are diverted away from the carriers' network, carriers may lose control over the data being offloaded. Despite these issues, we believe that the impact of our work is significant: since our findings conclude that offloading is an effective means for accommodating the current and future traffic growth, our simulation tools can offer important guidance for network providers in deploying and upgrading their networks and also in designing successful and creative price plans. Given the strong performance advantages of WiFi offloading, we foresee that there will be technical solutions as well as policy and price restructuring to address these issues in near future.

Acknowledgements

This work is supported in part by National Science Foundation under grants CNS-0910868, CNS-1016216 and IT R&D program of MKE/KEIT (KI002137, Ultra Small Cell Based Autonomic Wireless Network).

6. REFERENCES

[1] Urban tomography project. http://tomography.usc.edu/.

[2] Cisco visual networking index: Global mobile data traffic forecast update, 2009-2014, February 2010. *http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html*.

[3] Data, data everywhere, February 2010. *http://www.economist.com/specialreports/displayStory.cfm?story_id=15557443*.

[4] A. Abhari and M. Soraya. Workload generation for youtube. *Multimedia Tools and Applications*, 46(1):91–118, 2010.

[5] R. Ahmad and Z. Lin. Context-for-wireless: context-sensitive energy-efficient wireless data transfer. In *Proceedings of ACM MobiSys*, 2007.

[6] A. Balasubramanian, R. Mahajan, and A. Venkataramani. Augmenting mobile 3g using wifi. In *Proceedings of ACM MobiSys*, 2010.

[7] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani. Energy consumption in mobile phones: a measurement study and implications for network applications. In *Proceedings of ACM SIGCOMM IMC*, 2009.

[8] A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, May 2005.

[9] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on the design of opportunistic forwarding algorithms. In *Proceedings of IEEE INFOCOM*, 2006.

[10] I. Constandache, R. R. Choudhury, and I. Rhee. Towards mobile phone localization without war-driving. In *Proceedings of IEEE INFOCOM*, 2010.

[11] R. Gass and C. Diot. An experimental performance comparison of 3G and WiFi. In *Proceedings of ACM PAM*, 2010.

[12] D. Han, M. Kaminsky, A. Agarwala, K. Papagiannaki, D. G. Andersen, and S. Seshan. Mark-and-Sweep: Getting the"inside" scoop on neighborhood networks. In *Proceedings of ACM IMC (short paper)*, 2008.

[13] T. Kaneshige. AT&T iPhone users irate at idea of usage-based pricing, December 2009. *http://www.pcworld.com/article/184589/atandt_iphone_users_irate_at_idea_of_usagebased_pricing.html*.

[14] T. Karagiannis, J.-Y. L. Boudec, and M. Vojnovic. Power law and exponential decay of inter contact times between mobile devices. In *Proceedings of ACM MOBICOM*, 2007.

[15] A. J. Nicholson and B. D. Noble. Breadcrumbs: forecasting mobile connectivity. In *Proceedings of ACM MOBICOM*, 2008.

[16] M.-R. Ra, J. Paek, A. B. Sharma, R. Govindan, M. H. Krieger, and M. J. Neely. Energy-delay tradeoffs in smartphone applications. In *Proceedings of ACM MobiSys*, 2010.

[17] M. Reardon. Cisco predicts wireless-data explosion, Feburay 2010. *http://news.cnet.com/8301-30686_3-10449758-266.html*.

[18] I. Rhee, M. Shin, S. Hong, K. Lee, and S. Chong. On the levy walk nature of human mobility. In *Proceedings of IEEE INFOCOM*, 2008.

[19] A. Sharma, V. Navda, R. Ramjee, V. N. Padmanabhan, and E. M. Belding. Cool-tether: energy efficient on-the-fly wifi hot-spots using mobile phones. In *ACM CoNEXT*, 2009.

[20] A. Thiagarajan, L. Ravindranath, K. LaCurts, S. Madden, H. Balakrishnan, S. Toledo, and J. Eriksson. VTrack: accurate, energy-aware road traffic delay estimation using mobile phones. In *Proceedings of ACM SenSys*, 2009.