Opportunistic Traffic Scheduling in Cellular Data Networks

Utpal Paul*, Milind Madhav Buddhikot†, Samir R. Das*
*Computer Science Department, Stony Brook University, Stony Brook, NY 11794-4400, U.S.A.
† Alcatel-Lucent Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974-0636, U.S.A.

***Abstract*—The cellular data networks are experiencing a serious capacity crunch in the face of exponential increase in mobile data traffic volume. New traffic management techniques are needed to improve network and user perceived performance. In this work, we consider the existence of a higher-layer, agent-based scheduling system that could potentially delay scheduling of low priority flows at peak loads. The priorities are assumed to be user or application tagged, either automatically or manually. The general goal is to potentially move the low priority flows in time and space opportunistically to reduce the overall resource needs. We develop and evaluate two scheduling schemes – one based on a straightforward greedy method that requires real-time load monitoring and the other based on model-based estimation of future traffic loads and subscriber mobility based on historical data. Simulation results using a large-scale cellular network trace data collected inside a nationwide network show the potential of these approaches in reducing base station resource requirements. This indirectly demonstrates that if providers can incentivize subscribers to tag certain flows as low priority, they can potentially accommodate a significant number of additional subscribers in the same network without expending any additional resource.**

I. Introduction

Broadband cellular data networks have become the most common means of mobile data access. This is fueled by the availability of low-cost smart phones, tablet, e-reader and netbook devices with plethora of mobile applications. A recent study shows that for the last five years, the traffic volume in cellular data network has doubled in every year and this trend is expected to continue in the coming days [1]. Accommodating the exponential increase in traffic volume has now become a challenge for the service providers as increasing network capacity involves major capital investment in terms of new spectrally efficient technology, additional spectrum and/or additional base stations.

As reported in previous studies [2], [3], the load under a base station in a cellular data network fluctuates during the day, following a diurnal pattern, very high during the peak period (mid day) and typically low during the off-peak period (late night). The difference of traffic volume between peak and off-peak period is also very high. Due to this high variance of load, base station resources are under utilized for a significant period during the day. Dynamic resource allocation can address this situation. But this does not quite solve the issue of capacity limitation, as during the peak period all the base stations must allocate all available resources to accommodate the traffic. We take the approach of addressing this issue in the higher layer by shifting some traffic load from peak to off-peak periods in an opportunistic fashion. The goal is to reduce the peak-to-average ratio of traffic load in the network, ideally flattening the load curve as much as possible. This enables the provider to accommodate more users in the network without investing in capacity improvement. The basic idea is not unlike recent efforts in developing smart electric grids [4], [5] where there is an interest in reducing peak load by shifting load towards off-peak hours when electricity is cheaper. But in our case, the temporal shift of the load also prompts the opportunity of spatial movement considering the mobility of the subscriber. This should reduce the variance of load under a base station and also lower the peak of the load curve while serving the same total load. The service provider can potentially use the spare resource to accommodate more users in the network.

In this paper, we consider a model where a fraction of data requests specified by the subscribers can be delayed. Requests of this category are treated with low priority. Interactive applications or most of the short-lived flows like mail reading or http browsing may not fall into this category. Possible examples of low priority flows are certain types of media download/upload, P2P flows etc., that can tolerate a reasonable amount of delay without hurting the user experience any significantly. The subscriber can specify a deadline by which he/she wants the service to be provided and network makes a best-effort to find appropriate spot in time and space to fulfill the request. The rest of the traffic in the network is treated as high priority and they are served immediately. This model allows the network to move around the low priority traffic both spatially and temporally, and schedule them with the availability of spare resource of the currently associated base station of the corresponding subscriber.

The above approach provides two clear benefits: (i) Moving low priority traffic from a congested space-time point to another that has the capacity to carry the traffic. This allows the high priority traffic to be served with better performance. (ii) Reducing peak resource requirements by removing the load from the peak period to off-peak period; and thus indirectly allowing the network to make room for more users.

Our focus in this paper is to evaluate the model described above and analyze the feasibility of two simple approaches to schedule the low priority traffic: (i) greedy scheduling approach and (ii) modeling-based approach. We evaluate the waiting time of the low priority traffic and also analyze the effect on high priority traffic. We also investigate how much the model can reduce the resource requirement in the network.

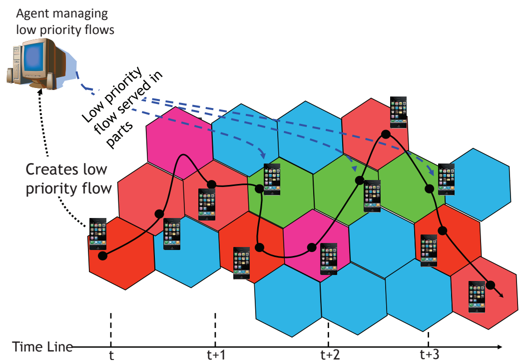


Fig. 1. Overall idea of opportunistic scheduling of low priority flows. The color in each cell indicates the congestion level in that cell. Red means highly congested and green means no congestion. The trajectory of a subscriber is shown. The low priority flow is served when the congestion level at the cell is low.

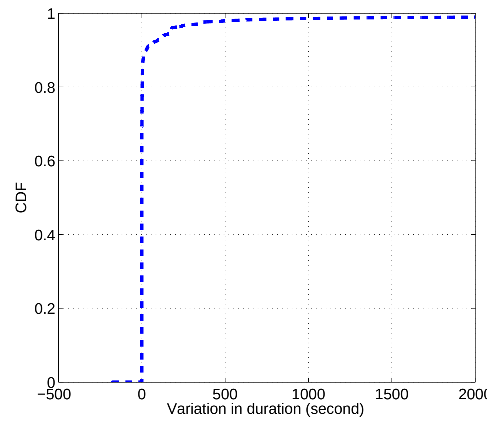


Fig. 2. CDF of variation of flow duration in the simulation from the trace of each flow. Here all the flows are treated with same priority.

The evaluation is done using a trace-driven simulation on a large-scale data traffic collected at the core of a nation-wide 3G network for our analysis. The data set spans one week in 2007 and consists of all data traffic associated with the entire subscriber base (in the order of hundreds of thousands) in a nation-wide network with thousands of base stations. All generated data packet headers (but not including user payloads) and various signaling and accounting packets are captured, archived and later post processed using a tool we have developed to gather all the flow level information. This is the same data set we used in our earlier works [3], [6].

The rest of the paper is organized as follows. We describe our model and scheduling strategy in Section II. We present our analysis using the greedy scheduling based approach in Section III and describe the modeling based approach in Section IV. We describe the implications both from the perspective of network provider and subscriber in Section V. We discuss related work in Section VI and conclude in Section VII.

II. Overall Approach

A. Model Description

We use an abstract model of the base station behavior to help us analyze opportunistic scheduling. A number of base stations covers a geographic region. A subscriber moves in the network and associates to a single base station at any time instant. When a subscriber creates a flow, the associated base station allocates radio resource (channel) for that flow. Here flow means TCP or UDP flow (upload or download). If the subscriber moves from one cell to another, hand off takes place and all the ongoing flows are now served by the new base station. In the current cellular network design, all regular flows are treated equally (i.e., high priority in our terminology), meaning that they need to be served immediately. Base stations need to be equipped with enough resource to accommodate all such flows, especially in the peak period. A flow arriving in a congested situation either gets dropped or served with a low performance.

In our model, we introduce a second category of flow priority. These flows have low priority and can be delayed and scheduled opportunistically on the resource availability. Figure 1 shows the overall idea. A subscriber creating a flow tags it with low priority and informs an agent running higher in the hierarchy inside the network. This agent also tracks individual base station loads at a suitable granularity and is responsible for scheduling the low priority flows. The agent uses its knowledge of the base station loads and also the mobility of corresponding subscribers for the actual scheduling decisions. Note that low priority flows may need to wait for being scheduled by the agent and even after the start, it may need to be defered/suspended and resumed later for multiple times until it is completely served. This means that low priority flows may be served in chunks, as demonstrated in Figure 1. The subscriber can also specify a time window for a low priority flow within which the flow needs to be completed. The agent considers this time window as a deadline for that low priority flow and tries to schedule the flow accordingly.

B. Approach

A fraction of the flows created by the subscribers is indicated as low priority flows. We consider two different approaches for scheduling the low priority flows.

Greedy Scheduling Approach is a simple approach where the agent continuously monitors the load of the base station where the corresponding subscriber of the low priority flow is associated with and starts the flow whenever there is any spare capacity available. The idea is very simple, but keeping track of each of the base stations' load for a large number of low priority flows may create an extra overhead of messege passing on the network. We discuss about this approach in more details with the results of our simulation run in Section III.

Modeling Based Scheduling Approach schedules the low priority flows using the predictive model of base station load and subscriber mobility. It models the mobility pattern of each subscriber in the network to predic his location. It also models the load on each base station in the network and determines the opportunistic time spots of each base station. Using these two models, low priority flows are scheduled where the scheduling problem is formulated and solved as a network flow problem. This approach is comparatively more complex and can suffer from modeling error. We discuss about this approach in more details in Section IV.

We investigate how the model helps to reduce the resource requirement in the network. To evaluate this, we assign lower capacity to each base station in our simulation run and estimate the effect on both high and low priority flows. The goal here is to schedule low priority flows so that high priority flows are benefitted and low priority flows suffer a reasonable delay. We evaluate this with both approaches mentioned.

C. Data Set

Our data set provides a range of information for each flow created by subscribers including the start time, flow duration in seconds, number of bytes transferred. It also provides the information of the corresponding subscriber and the base station where the flow is initiated. It keeps track of the mobility information of each subscriber irrespective of flow creation. This lets us know whenever a subscriber changes his current base station.

We model the capacity, C_j of a base station j as the maximum aggregate throughput of the base station observed during the span of the data trace. This estimation of capacity of the base station may not be accurate. But at the absence of the accurate information of the underlying physical resource, it can provide us an indication of the capacity. This will also make our model simple. We calculate the capacity for each base station in the network. Here the capacity of the base stations are not equal and may indicate much less than the actual physical resource assigned. Thus the results presented in our evaluation indicates a lower bound of improvement.

We also model the flows and their resource requirement. Our data set keeps track of the total number of bytes transferred and duration for each flow. From this information, we calculate the *average throughput*, T_i^{avg} for each flow i. Along with that our data set also provides the *maximum throughput*, T_i^{max} achieved for each flow, i during its life-time under a base station, j. We consider this as an indication of channel quality for that flow under that particular base station. Our flow-model suggests that each flow, scheduled under a base station, is served with a fraction of its maximum throughput. That is, a flow i scheduled under base station j is served with a throughput of σT_{ij}^{max}, where $0 < \sigma \leq 1$. At an instance of time, all the flows under a base station j are served with same fraction which is obtained using this formulation: $\sum_i \sigma T_{ij}^{max} \leq C_j$, $0 < \sigma \leq 1$. The value of σ needs to be changed based on the availability of the resource, specifically at the arrival and departture of any flow under the base station. We assume that when a flow is scheduled with some resource, it is served constantly with the specified throughput until that rate is changed or the size of the data transmission for that flow is over the total number of bytes as specified in the trace.

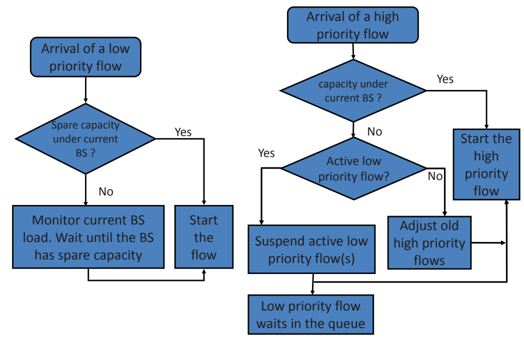


Fig. 3. Flow chart describing the greedy scheduling approach.

This may not be realistic as in real protocol physical resource such as channels which are assigned to mobile devices can be time shared among multiple flows and acquisition of physical resources should depend on the packet generation behavior of the flow.

To demonstrate the effectiveness of our capacity and flow model, we do a sanity checking by a simulation run. We use the data set without any priority enforcement of the flows and with full capacity of the base stations observed from the trace. The goal is to simulate the flows with the arrival time and transmitted bytes as specified in the trace and investigate the variation of flow duration in the simulation with respect to the flow duration in the date set for each flow. Figure 2 plots the distribution of the variation of flow duration. More than 80% of the flows follow the same timeline as in the trace and among the rest of the flows, most of them are deviated marginally. This indicates the effectiveness of the model of our simulation.

Low priority flows are introduced under a base station based only on either of the scheduling approaches discussed in next two sections. Only the spare capacity after the high priority flows are distributed among all the currently assigned low priority flows. More on this will be discussed for each approach in the respective sections.

III. Greedy Scheduling Approach

In this section, we describe the greedy scheduling approach. We also develop a trace driven simulator to evaluate the approach using our data set described before. Our goal here is to quantify the benefit that this model provides in terms of reducing resource requirement using the greedy approach.

A. Approach

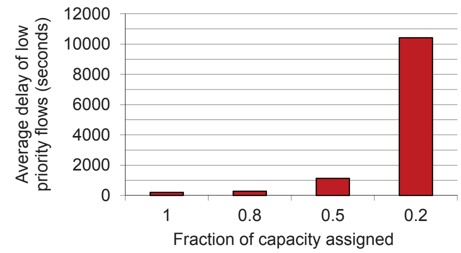
The greedy scheduling approach can be described as follows. This is also shown as flow charts in Figure 3.

- High priority flows are treated just as it is in the current network design. It is started immediately after it arrives. All the active high priority flows under a base station share the total capacity so that each of these flows

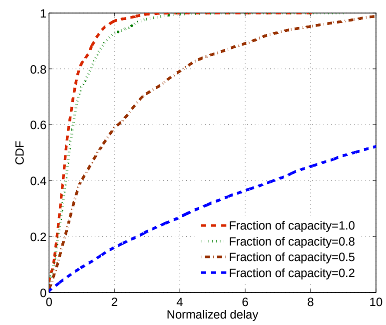
is served with its own required capacity based on the availability.
- Upon arrival of a low priority flow from a subscriber, the agent checks whether there is any spare capacity under the base station where the subscriber is currently associated. If there is no capacity available, the flow is stored in a queue where all such low priority flows wait to be served. Otherwise, the flow starts immediately with the capacity available.
- An active low priority flow is deferred in case of an arrival of an high priority flow under the same base station having no spare capacity. The deferred low priority flow is stored in the queue with its current status.
- For each of the low priority flows waiting in the queue, the agent always checks for any spare capacity under the base station where the corresponding subscriber is associated with. Whenever the agent finds an opportunity under a base station, it starts a low priority flow from the waiting list with the capacity available.
- When a subscriber having active flows hands off from one base station to another, all the active high priority flows (if any) of that subscriber are first accommodated under the new base station. It may require to defer a number of active low priority flows under the new base station to accommodate the migrating high priority flows. If the mobile subscriber has any low priority flow being served by the old base station, the agent decides about that flow based on the available capacity of the new base station.

We have developed an event-driven queueing simulator to study the impact of opportunistic scheduling. To classify flows into high and low priority flows we take the following approach. We assume that short-lived flows are of immediate need and cannot be delayed (e.g., http browsing or email reading). On the other hand, the subscriber could be incentivized to delay long-lived flows (e.g., large download or P2P traffic). We consider flows longer than 1500 sec as long-lived flows where as the overall average flow duration among all the flows is 150 sec. In our data set, around 12% of flows are such long-lived flows. A random subset of flows which is about 8% of all flows in the network is chosen as low priority for our simulation.

Based on the available capacity of the serving base station, the throughput of high priority flows varies. As mentioned in the previous section, each high priority flow is assigned a fraction, σ of its maximum throughput. In our simulation, a low priority flow is started under a base station only if all the current high priority flows are being served with their maximum throughput, that is, $\sigma = 1$ and the base station still has some spare capacity. The spare capacity of the base station is distributed among the active low priority flows under that base station. The number of low priority flows under a base station is incremented as long as each of the active low priority flows under the base station achieves at least its average throughput, T_i^{avg}. Arrival of a new high priority flow under a base station may need to suspend zero or more low priority flows depending on the capacity situation. In the



(a) Average delay of low priority flows.



(b) CDF of normalized delay of low priority flows.

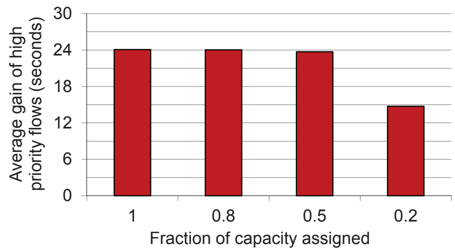
Fig. 4. Effect on the low priority flows for the greedy scheduling approach. 75% of the long-lived flows which are 8% of all flows are assumed low-priority.

case when there is no active low priority flow under the base station, the new flow is accommodated only by adjusting the throughput of other high priority flows, if required. On the other case, a number of active low priority flows under the base station is suspended to start the new high priority flow with its maximum throughput.

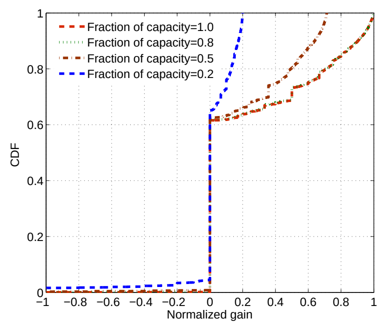
Our goal of this evaluation is to investigate how the model and the greedy approach can reduce the resource requirement in the network. To do this, we study what would happen to the incoming flows if the base station capacities were reduced. We do the simulation study for capacities such as 20%, 50%, 80% of the actual capacity of the base stations as determined from the trace and also provide the 100% results as the base case. The idea is to study the impact on the flows with reduced capacity base stations. If the impact is acceptable, e.g., low priority flows are not delayed substantially and only few high priority flows are impacted, this would indicate that more subscribers could be accommodated with the provisioned capacity.

B. Simulation Results

With flow prioritization and opportunistic scheduling, it is possible that high priority flows end early relative to its actual end time in the trace. This is because they are expected receive more capacity during service. Low priority flows on the other



(a) Average gain of high priority flows.



(b) CDF of normalized gain of high priority flows.

Fig. 5. Effect on the high priority flows for the greedy scheduling approach. 75% of the long-lived flows which are 8% of all flows are assumed low-priority. The rest are high priority.

hand are likely to be deferred, possibly multiple times, and thus would end late relative to its actual end time in the trace. We use the term 'delay' for a low priority flow to indicate the difference between its end times in the simulation run and the actual trace (end time in simulation – end time in trace). We use the term 'gain' for the high priority flows to indicate the same thing, but in the opposite direction (end time in trace – end time in simulation).

Figure 4(a) shows the effect of greedy scheduling on the low priority flows for different (reduced) capacity assignments of the base stations. Note that the average delay of low priority flows is 1200 sec (20 min) when the capacity of base stations is made half of the actual. This is comparable to the original flow duration of long-lived flows in the data set as evident in Figure 4(b) showing the delay of each flow normalized by its flow duration specified in data set. On the other hand, Figure 5 shows the gain of the high priority flows in actual and normalized fashion. Note that more than half of the flows are unimpacted and and over one third of the flows show gain in varying degrees depending on the capacity of the base stations. A negligibly small fraction of high priority flows are negatively impacted for capacities 100%, 80% and 50%. This fraction is only noticeable (about 5%) for the 20% capacity case.

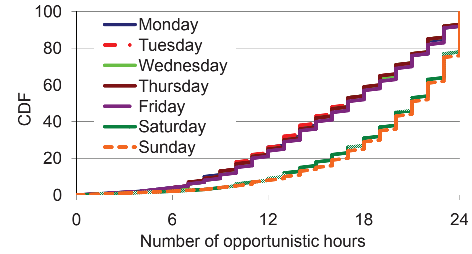


Fig. 6. CDF of the number of opportunistic hours in a base station in a day.

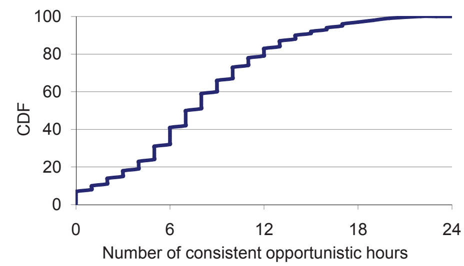


Fig. 7. CDF of the number of 'consistent' opportunistic hours in a base station in the weekdays.

C. Critique of Greedy Scheduling

While the greedy scheduling approach is straightforward to explain, the approach as described requires the agent to monitor the load on the base stations in a continuous basis, looking for scheduling opportunities. This naturally requires a significant amount of control information to be passed around among the base stations and the agent. This could be a significant overhead on the network, especially during the peak periods. Managing all the low priority flows in the network by a single agent may also be a scalability issue. Also, a low priority flow may suffer from a large number of suspend/resume operation incurring an extra processing overhead on the network. This can potentially introduce thrashing. Much of these issues can, however, be addressed via well-known techniques, such as choosing more granular measurement/scheduling intervals to reduce control overhead and choosing load thresholds to make scheduling decisions for low priority flows to reduce thrashing. But these can also negatively impact the performance advantage.

IV. Modeling Based Scheduling Approach

We propose a modeling based approach to address the practical limitations of implementing the greedy approach that requires continuous load monitoring. The modeling based approach relies on the hypothesis that human mobility and network load are predictable and thus models for them can be created using historical trace data and off-line analysis. These models are useful in scheduling low-priority flows. This strategy completely eliminates the need for any real time monitoring. To establish the usefulness of this approach, we first evaluate how much predictability exists in the load and mobility that can be gainfully exploited.

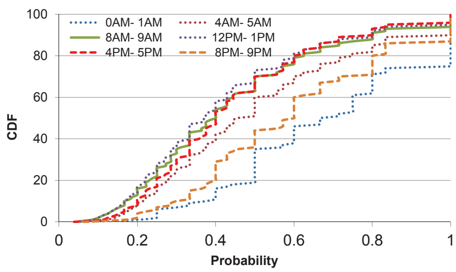


Fig. 8. CDF of probability of a subscriber being in the most likely location.

A. *Profiling Base Station Load*

We first determine how frequently periods of opportunity arise where the base station load is low (low is defined as 25% of the capacity, where the capacity is defined as the maximum load the base station has seen in the entire period of the trace). We do this study in the granularity of an hour. The Figure 6 shows the CDF of the number of these opportunistic hours of a base station in each day. Note that a typical base station has at least 16 opportunistic hours in a day and the behaviour is similar among all the weekdays. Weekends, as can be expected, provide more number of opportunistic hours. Now an obvious question is: Is the set of opportunistic hours of a base station 'consistent' (i.e., same hour of day across days)? The Figure 7 shows the CDF of the number of consistent opportunistic hours of each base station among all the 5 weekdays. We kept the weekends out of this as the nature of load in weekends is different from the weekdays. We see that a typical base station has 7 opportunistic hours that are consistent among days. This analysis of base station load indicates that a plenty of scheduling opportunities exists for low priority flows and much of it is predictable.

B. *Profiling Subscriber Mobility*

We model the subscriber trajectories to find out the probability of a subscriber being at a specific location, that is, in a specific cell at a given time instance. To do this, we list different cells where a particular subscriber is observed in the trace during a particular time period in all 5 weekdays and calculate the total duration spent in each of these different cells. Specifically, if the length of the time period is l_t, then for each time period, t_i of the day, a subscriber is observed for $5 \cdot l_t$ time as profiling is done using the 5 weekdays in our data set. For each location, j where the subscriber is observed in time period t_i, we add the time durations d_{ijk} which is the duration the subscriber is present in location j at time period t_i in kth weekday of our data set for all k days. We calculate the probability of that subscriber being in location j in time period t_i as the ratio of $\sum_k d_{ijk}$ and $5 \cdot l_t$. The distribution is created for each subscriber for each time period in a day. Figure 8 shows the CDF of probability of a subscriber being in the most likely location at different time periods. Here we consider 1 hour time period. We indicate the location

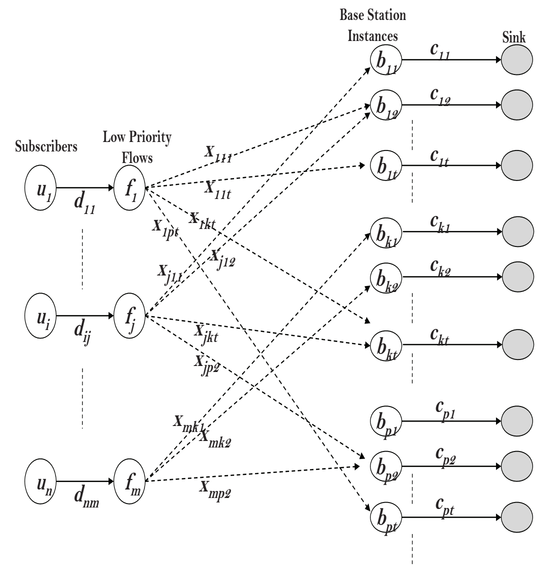


Fig. 9. Network graph used to solve the scheduling problem using the Modeling Based Approach.

with highest probability at a time period as the subscriber's most likely location at that time. Note that a typical subscriber is found in his most likely location with probability 0.4-0.6. We observe that this probability increases during the off-peak period. This analysis indicates that a subscriber's location can be predicted with reasonable accuracy.

C. *Scheduling Low Priority Flows*

We formulate the problem of scheduling the low priority flows as Network Flow problem [7] using the profiles created for each base station and subscriber. Here, we assume that each low priority flow also has a deadline by which it needs to be finished. We construct a network graph as shown in Figure 9 using the following steps:

- Each subscriber having at least an unserved low priority flow is respresented as a node u_i. Each of these nodes is connected to nodes, marked as f_j representing low priority flows created by the corresponding subscriber. The weight on this directed edge, denoted as d_{ij} is the estimated number of bytes to serve the low priority flow.
- We create an instance b_{kt} for each base station b_k at time period t. Each of these nodes is connected to a sink with a directed edge with weight c_{kt} denoting the spare capatiy available under base station b_k at time period t. This is obtained from the profile created for each base station.
- Node f_j representing a low priority flow created by subscriber u_i is connected to different base station instances based on the mobility of the subscriber and the deadline of the flow. This means that f_j is connected to b_{kt} if the subscriber u_i is likely to be under base station b_k with reasonably high probability at time period t (t is within the specified deadline of the flow). The weight of

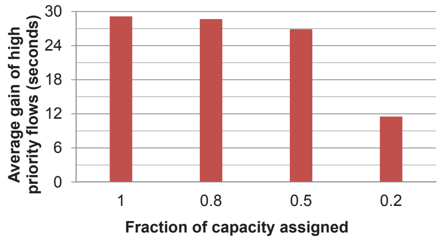


Fig. 10. Gain of high priority flows for the modeling based approach.

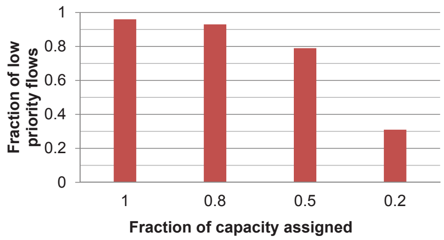


Fig. 11. Fraction of low priority flows finishing within deadline.

this edge, denoted as x_{jkt} is a function of the throughput achieved and the duration of the subscriber's stay under base station b_k in period t. This is an estimation of the number of bytes the subscriber u_i can transmit during his stay under base station b_k in period t. This is also modeled from the historical data.

The agent managing the low priority flows constructs this graph periodically with all the low priority flows waiting with their current states and get the scheduling by solving this as a special case of Network Flow problem where there are multiple sources and multiple sinks [8]. The subscriber nodes u_i act as source nodes. This formulation makes sure that base stations do not get overloaded and low priority flows are scheduled with in the specified deadline. This also enables that low priority flows can be scheduled in chunks. As the scheduling is done by the agent periodically, it can cover up any modeling or predicting error of subscriber mobility and other. The mobile device with the subscriber informs the agent about its current location (i.e., associated base station) whenever there is a handoff. Based on the location of the subscriber and the computed schedule, the agent starts any low priority flow that might be waiting.

D. Evaluation

We evaluate the predictive modeling based approach in the similar manner as we have done it for the greedy scheduling approach: assigning a lower capacity to the base stations and analyzing the effects on both high and low priority flows to demonstrate the reduction in resource requirement by the approach. Before going into the real evaluation, we model the load of each base station in the network to predict the spare capacity at each time slot (one hour in our case). As our evaluation includes assigning lower capacities to the base stations, we also need to model the base station load for each of the lower capacity assignment. For each such capacity assignment we simulate the network with all flows in the data set and model the spare capacity in each time slot for each base station. We also model the mobility of each subscriber by calculating the probability of the subscriber being under a base station at a specific hour. For this modeling purpose, we only use the data set of 5 weekdays from our week-long data. Week-end data is not deemed statistically meaningful as there are only two days and their nature substantially differ from the weekdays.

For a meaningful evaluation, we will need a long term trace. Since the trace is relatively short (only 5 weekdays), for evaluation purposes we synthetically augment the trace using established statistical techniques. The augmented data considers all the subscribers and base stations from the original trace data for the weekdays. The data generation is based on the probability of a subscriber creating a flow under a base station at a time instance. Specifically, while augmenting the data set, a flow of a subscriber is randomly selected at a time instance from the pool of flows the subscriber has created in the original data set at that time instance in any of the 5 weekdays.

Just like in the earlier case, we identify a fraction of the long-lived flows which are 8% of all the flows in the synthetic data set as the low priority flows. For each of these flows, a deadline is specified picked randomly from an window of 1 to 4 hours beyond the arrival of the flow. Note that the window of deadline is comparable to the original flow duration as we are only considering long-lived flows as low priority. We apply our approach to schedule the low priority flows using the models created from our original data set. At the beginning of each hour, we get a schedule of all the low priority flows that are waiting.[1] Note that any low priority flow arriving in the middle of an hour, will only be scheduled at the beginning of the next hour. This situation can be improved by choosing a smaller scheduling interval. After the schedule computation, the agent will start a low priority flow according to the schedule at the location of the corresponding subscriber only if the base station's current real load is lower or equal to the predicted load of that base station at that time instance. At each scheduling event, we consider all the low priority flows: either scheduled or newly arrived. This helps the approach to overcome any modeling error.

Figure 10 shows the average gain of the high priority flows for different capacity assignments of the base stations. The average gain of high priority flows is around 27 seconds when the capacity of base stations is made half. This is similar to what we have observed for greedy approach. This is understandable as in both cases the high priority flows are benefiting in similar fashion with more available resources. As the low priority flows are scheduled by the deadlines, the delay of the low priority flows may not be interesting to analyze.

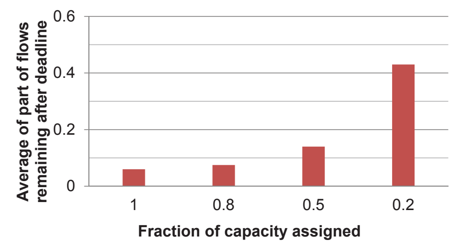
[1]The interval of one hour is chosen to make the computing process simple and tractable. For practical purposes, the scheduling inteval can be smaller.

On the hand, it may be interesting to see what fraction of low priority flows gets finished by the deadline. Figure 11 shows about 80% of the low priority flows are finished by the deadline specified for the case when base station capacity is made half of its original capacity. This shows the potential of the scheduling approach. We also investigate what fraction of each flow served after the deadline. For each of the low priority flows that can not make the deadline, we calculate the fraction of flow size in terms of number of bytes served after the deadline. Figure 12 shows the average fraction of flows remining after the deadline. Note that the remaining portion is not significant (about 15%) even when the capacity is made half.

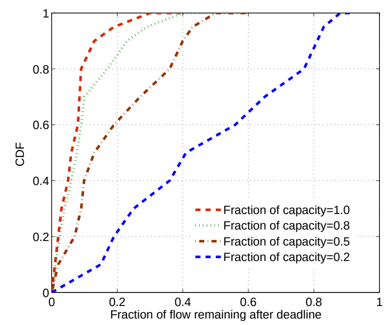
Note that even with the assignment of full capacity, that is, with fraction of capacity=1, a small fraction of low priority flows can not meet the deadline. Our investigation suggests that this is due to the modeling and prediction error. Moreover, the deadlines of the low priority flows are picked randomly and is not correlated to original flow duration.

E. Crtique

The Modeling Based Approach is more practical as it does not require any real-time load measurement and relies on historical data to derive load and mobility estimates. It considers the global network-wide scenario as opposed to the previous greedy approach where each base station is treated in an isolated fashion. Many wireless providers do collect subscriber/base station specific load information in various forms for network monitoring. Thus, off-line use of such data to create profiles as used in the above evaluation is entirely plausible. Scalability of global scheduling can still be an issue. But the network can always be partitioned in smaller parts and scheduling can be done in each of these parts independently to address scalability issues.



(a) Average fraction of flow size remained after the deadline.



(b) CDF of fraction of flow size remaining after the deadline.

Fig. 12. Fraction flow-size of lows priority flows remaining after deadline.

V. Discussion

We now summarize our key observations and identify important practical implications both from the perspectives of the network provider and the subscriber:

1) Better Service: Normally subscribers may experience poor service during the peak periods due to network congestion. Flows can be dropped or served with very poor rate due to congestion. A frustrated subscriber can try multiple times to initiate communication possibly leading to more congestion. Existing networks do not have any built-in mechanism for service differentiation and treats all flows equally. The proposed mechanisms provide a higher-layer, agent based mechanism to provide service differentiation based on a simple prioritization of flows. We show that existing load can be served even with half the capacity with only modest delays on the low priority flows and little or no negative impact on high priority flows, and sometimes with some positive impact.

2) Subscriber Pricing and Incentivization: Providers are moving away from unlimited data plans and replacing them with tiered plans as cellular data networks are becoming more popular. This is evidently focused towards managing the network load better. With the opportunistic scheduling, the providers could provide incentives to subscribers to tag (automatically via apps, or via a profile driven approach, or even manually) flows as low priority. A possible incentive could be that low priority flows are not metered to count as a part of total data usage by the subscriber. This provides a semblance of unlimited data plan to the subscriber and may attract more customers to the provider's network.

3) Reducing Resource Requirement: Our analysis with both the approaches shows that the resource requirement of base stations can be reduced significantly considering only a small fraction of flows with low priority. We believe that this can be reduced even more if the fraction of low priority flows increases. The service provider can utilize the spare capacity to accommodate more high priority flows, in other words, more new subscribers in the network.

VI. Related Work

Our work in this paper has some level of similarity with the broad topic of quality of service scheduling and load balancing, as we propose to move low priority flows both spatially and temporally. This general idea has been widely used where wireless resources are redistributed in form of channel assignment rather than traffic [9], [10], [11]. A large body of work on scheduling approaches on link layer is also available [12], [13], [14], [15]. In contrast, our work deals with the load shifting problem at a higher layer and at the flow level. We focus on scheduling of flows, specially low priority flows either in parts or in whole ignoring the low level issues such as power, interference, radio resources, packet level scheduling. Similar load shifting studies have been done in other contexts, such as power savings (see, e.g., [16]).

There are different pieces of work dealing with the priority scheduling in wireless networks. The authors in [17] have proposed a technique to set priority among the source stations in ad hoc network. The authors in [18] have devised an distributed priority scheduling in packet level for the nodes in ad hoc networks. The authors in [19] have modeled the arrival of flows in a base station as a queueing model with priority set between its own flows and flows arrived because of handoff. Our work is different from these set of works as we set priorities on the flows by using application layer information and deal with the opportunistic scheduling of low priority flows. A similar work, but in a different context has been done targeting TCP, where the authors have developed a variation of the regular TCP, called 'TCP-low priority,' in order to to utilize excess network bandwidth distributedly as compared to the fair-share bandwidth in regular TCP [20].

VII. Conclusion

In this paper, we have explored an avenue to reduce the peak load on cellular data networks. The idea is to treat certain flows as low priority and delay scheduling such flows if the base station has reached its capacity limits. Low priority flows are to be scheduled opportunistically based on the available capacity. The main goal of this model is to move traffic from the peak period to off-peak period that potentially reduces the average-to-peak ratio of load under base stations. We have presented two approaches to schedule the low priority flows. The first one is a straightforward greedy approach, but needs continuous monitoring of base station load in order to determine scheduling opportunities. The second one is a modeling based approach where models are created to predict subscriber location (base station) and base station loads based on historical data. This approach reduces the need for monitoring, but can potentially suffer from inaccurate estimates. Our analysis indicates that the capacity requirements at the base station can be reduced significantly – by as much as a factor of two – with only modest delays on the low priority flows. If low priority flows that those that are long-lived and delay tolerant such delays would be perfectly acceptable to the applications, but would be beneficial for addressing the data overloads in base stations. Further, this will help the providers to accommodate more subscribers without increasing network capacity. Our future work will also involve incentive and pricing schemes to make this realistic. The future work will also consider the design of the agent-based system that can perform the opportunistic scheduling proposed in this study.

References

[1] Message from Mobile Wireless Congress 2010. http://www.analysysmason.com/About-Us/News/Insight/The-message-from-MWC-2010/.

[2] R. Beckman, K. Channakeshava, F. Huang, A. Vullikanti, A. Marathe, M. Marathe, and G. Pei, "Implications of dynamic spectrum access on the efficiency of primary wireless market," in *Proc. DySPAN*, 2010.

[3] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding traffic dynamics in cellular data networks," in *Proc. IEEE Infocom*, 2011.

[4] E. Santacana, G. Rackliffe, L. Tang, and X. Feng, "Getting smart," *IEEE Power and Energy Magazine*, 2010.

[5] A.-H. Mohsenian-Rad, V. Wong, J. Jatskevich, and R. Schober, "Optimal and autonomous incentive-based energy consumption scheduling algorithm for smart grid," in *Proc. Innovative Smart Grid Technologies (ISGT)*, 2010.

[6] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding spatial relationship in resource usage in cellular data networks," in *Proc. IEEE NetSciCom*, 2012.

[7] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. MIT Press and McGrawHill, 2001.

[8] Multi-commodity flow problem. http://en.wikipedia.org/wiki/Multi-commodity_flow_problem.

[9] S. K. Das, S. K. Sen, and R. Jayaram, "A dynamic load balancing strategy for channel assignment using selective borrowing in cellular mobile environment," *Wireless Networks*, 1997.

[10] D. Everitt and D. Manfield, "Performance analysis of cellular mobile communication systems with dynamic channel assignment," *IEEE Journal on Selected Areas in Communications*, 1989.

[11] J. V. Leeuwaarden, S. Aalto, and J. Virtamo, "Load balancing in cellular networks using first policy iteration," Helsinki University of Technology, Tech. Rep., 2001.

[12] O. K. Tonguz and E. Yanmaz, "The mathematical theory of dynamic load balancing in cellular networks," *IEEE Transactions on Mobile Computing*, 2008.

[13] J.-W. Lee, R. Mazumdar, and N. Shroff, "Joint resource allocation and base-station assignment for the downlink in CDMA networks," *IEEE/ACM Transactions on Networking*, 2006.

[14] L. Du, J. Bigham, L. Cuthbert, P. Nahi, and C. Parini, "Intelligent cellular network load balancing using a cooperative negotiation approach," in *Proc. IEEE WCNC*, 2003.

[15] S. Das, H. Viswanathan, and G. Rittenhouse, "Dynamic load balancing through coordinated scheduling in packet data systems," in *Proc. IEEE Infocom*, 2003.

[16] C. Peng, S.-B. Lee, S. Lu, H. Luo, and H. Li, "Traffic-driven power saving in operational 3g cellular networks," in *Proc. ACM Mobicom*, 2011.

[17] X. Yang and N. H. Vaidya, "Priority scheduling in wireless ad hoc networks," in *Proc. ACM Mobihoc*, 2002.

[18] V. Kanodia, C. Li, A. Sabharwal, B. Sadeghi, and E. Knightly, "Distributed priority scheduling and medium access in ad hoc networks," *Wireless Networks*, 2002.

[19] J. Keilson and O. Ibe, "Cutoff priority scheduling in mobile cellular communication systems," *IEEE Transactions on Communications*, 1995.

[20] A. Kuzmanovic and E. W. Knightly, "TCP-LP: Low-priority service via end-point congestion control," *IEEE/ACM Transactions on Networking*, 2006.