Lookback Scheduling for Long-Term Quality-of-Service Over Multiple Cells

Hatem Abou-zeid*, Hossam S. Hassanein*, Stefan Valentin†, and Mohamed F. Feteiha*
*Queen's University, Canada, h.abouzeid@queensu.ca, hossam@cs.queensu.ca, feteiha@cs.queensu.ca
†Bell Labs, Alcatel-Lucent, Germany, stefan.valentin@alcatel-lucent.com

***Abstract*—In current cellular networks, schedulers allocate wireless channel resources to users based on short-term moving averages of the channel gain and of the queuing state. Using only such short-term information, schedulers ignore the user's service history in previous cells and, thus, cannot meet long-term Quality of Service (QoS) guarantees when users traverse cells with varying load and capacity. We propose a new scheduling framework, which extends conventional short-term scheduling with long-term QoS information from previously traversed cells. We demonstrate our scheme for relevant channel-aware as well as for channel and queue-aware schedulers. Our simulation results show high gains in long-term QoS while the average throughput of the network increases. Therefore, the proposed scheduling approach improves subscriber satisfaction while increasing operational efficiency.**

***Keywords*—Multi-cell scheduling, base station cooperation, long-term QoS, proportional fairness, exponential scheduling rule.**

I. Introduction

Global mobile traffic is growing at unprecedented rates, driven by the large screens of Smartphones and Tablets coupled with online media streaming [1]. At the same time, traffic is becoming more unevenly distributed in space and time [2]. Depending on the current traffic situation, users experience a mix of good and bad service while traversing the network. Such varying QoS is expected to increase with upcoming small cell deployments [3], which will result in users traversing a larger number of cells per session. Coping with this spatially varying service quality for mobile users is targeted in this paper.

A closer look at current cellular networks reveals three important characteristics of current schedulers. First, the accurate computation of scheduling weights plays a key role in providing Quality of Service (QoS) guarantees to mobile users. Second, current schedulers compute weights by averaging Physical layer (PHY) data rate and queue length over time intervals in the order of seconds [4, Ch. 6]. Third, this weight computation excludes the user's long-term service experience in previously traversed cells. However, focusing only the current cell and ignoring most of the user's service history will provide unsatisfactory QoS to mobile users in the long run.

This work was made possible by a *National Priorities Research Program* (NPRP) grant from the Qatar National Research Fund (Member of Qatar Foundation).

In this paper, we propose Long-term Lookback Scheduling (LLS). This new scheduling framework is based on two components. First, Base Stations (BSs) aggregate QoS indicators, such as a user's PHY data rate and queuing state, over tens or hundreds of seconds. Then, the BSs exchange these values as the users traverse the cellular network. The final scheduling weight is computed by combining these long-term measures with the conventional short-term moving average QoS indicators of the current cell. By doing so, the scheduler can now account for the users' QoS indicators over multiple time scales and multiple cells.

The proposed scheduling approach improves long-term user satisfaction over multiple cells while still capturing instantaneous channel gains and queuing states. For example, if a user with a poor service history enters a new cell, the proposed scheduler will prioritize this user over another incoming user who previously received good service. However, if a user with a poor scheduling history has a bad channel state, other users will be scheduled to efficiently use wireless channel resources. By enabling resource allocation over multiple cells, LLS reduces the negative effects of uneven traffic distribution without sacrificing spectral efficiency.

Although LLS is a general approach for various QoS indicators and schedulers, we demonstrate this framework for two practical examples. First, we modify the Proportional Fair (PF) scheduler [5] to include long-term user average rates. While this shows the positive effect of LLS on channel-aware scheduling, we also investigate the potential for channel and queue-aware schedulers. Here, we choose the Exponential (EXP) scheduler as it was shown to have good performance with delay sensitive traffic, by keeping queues stable if it is possible to do so [6]. Both scheduler extensions trade-off QoS indicators at different time scales, without requiring central coordination or excessive signaling. This indicates that LLS can be practically applied in existing cellular networks.

We present the details of the proposed LLS scheme in Sec. IV after outlining the system description in Sec. III. The resulting performance analysis is presented in Sec. V, followed by our conclusions.

II. Related Work

Prior work in BS coordination for scheduling has focused on instantaneous cooperation to achieve short-term objectives, i.e. BSs coordinate their transmissions periodically to minimize interference, balance load, or perform joint transmissions to a user such as in Coordinated Multi-Point (CoMP) [7].

In [8], Frank *et al.* propose a scheduling scheme for the 3GPP LTE uplink that accounts for inter-cell interference, and by avoiding high interference situations for users at the cell edge, they improve the average spectral efficiency. Bu *et al.* propose a Load Balancing scheme that improves proportional fairness over the network by controlling the association of users among neighboring BSs [9]. Users are not associated to the BS that gives the strongest signal strength, but instead to the BS that satisfies a network-wide proportional fairness criterion. This scheme is extended in [10] where partial frequency reuse (an inter-cell interference mitigation mechanism) is jointly optimized with the load balancing in a multi-cell network. More recently, in [11] the authors consider the case where a user is served by multiple BSs simultaneously and propose a scheme that provides instantaneous fairness over the network.

LLS differs fundamentally from the above multi-cell coordination approaches. Instead of adjusting scheduling based solely on current user conditions and needs, we propose incorporating the long-term service history of the users in prior cells into the scheduling framework. We do so to improve the long-term QoS for users as they traverse the network.

III. System Description

In this section we introduce our system model, performance metrics, as well as the traditional schedulers that we use in the proposed LLS framework.

A. Channel and Traffic Models

We study a network with a Base Station set $\mathcal{M}$ and a user handheld set $\mathcal{N}$. An arbitrary user is denoted by $i \in \mathcal{N}$ and an arbitrary BS by $m \in \mathcal{M}$, where the number of BS is $|\mathcal{M}| = M$ and the number of users $|\mathcal{N}| = N$.

Each BS covers a hexagonal cell. All the users traverse the network according to the Random Way Point (RWP) mobility model with a constant speed S, zero pause time between the waypoints, and no wrap-around. Omitting the wrap-around creates a traffic hotspot in the center of the network, which we are interested in to enable the study of an uneven load distribution. Fig. 1 shows the 19 cell network modeled in this paper along with 3 exemplary user motion paths.

We model the wireless downlink as typical for studies on macro-cell Long Term Evolution (LTE) systems. The path loss is calculated according to [12] as $\text{PL} = 128.1 + 37.6 \log_{10} d$ with the user-BS distance d in km. The BSs have omnidirectional antennas and a log-normal distribution with a variance of 8 dB accounts for slow fading. Fast fading is modeled as i.i.d. Rayleigh-fading, and link adaptation is modeled using Shannon's equation where the SNR is clipped at 20 dB to account for a maximum modulation order of 64 Quadrature Amplitude Modulation (QAM). Two traffic models are considered in this paper. The first is full buffer, meaning that each user has download traffic at any point in time. This model is used to compare the performance bounds for schedulers that

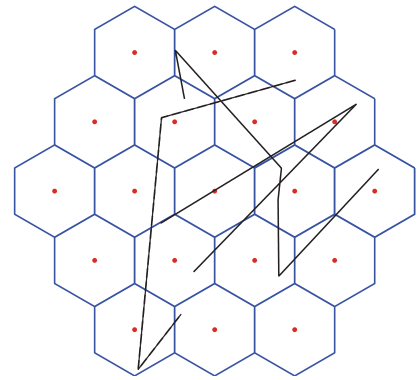


Fig. 1. The studied network with 19 cells and 3 sample motion paths.

are unaware of BS queues. The second traffic model includes user queues at the BS as shown in Fig. 2. For each of these queues, packets arrive at a constant rate λ_i. This model is used to evaluate queue-aware schedulers and represents non-real time video streaming traffic arriving at the BS with a constant bit rate.

To account for the buffering of such traffic, we also consider playback buffers at the user terminals. Here, the video stream will play at a constant rate R_{Stream} when the buffer is sufficiently filled. The video stream will freeze (or stall) when the buffer becomes empty if the user has not been scheduled sufficiently. The stream will remain frozen until the playback buffer is re-filled to the playback threshold, which corresponds to the behavior of modern stream protocols such as HTTP Live Streaming (HLS) [13].

B. Schedulers

The BS receives the users' Channel Quality Information (CQI) periodically at every Time Transmission Interval (TTI) and, then, makes scheduling decisions for the upcoming TTI.

1) Max-rate Scheduling: The Maximum Rate (MR) rule schedules the user with the highest instantaneous CQI, as observed from the previous TTI [5]. This maximizes the sum throughput of the network but makes no effort to serve users fairly.

2) Proportional Scheduling: The Proportional Fair (PF) scheduling rule [5] aims for high throughput while maintaining fairness among the users. The intuition of the algorithm is

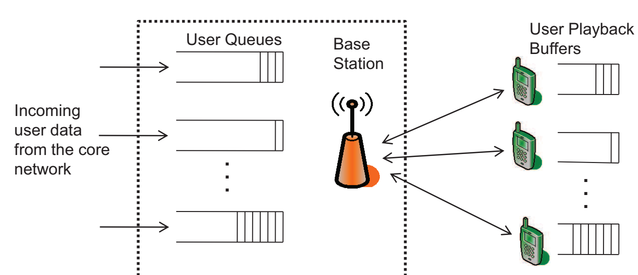


Fig. 2. Studied system with the traffic model for buffered media streaming.

to schedule users when they are at their peak rates relative to their average rates. In any TTI t, PF schedules the user $i^* = \arg\max_{\forall i \in \mathcal{N}} w_i(t)$ where the user weights are calculated $\forall i \in \mathcal{N} : w_i(t) = r_i(t)/R_i(t)$. Here, $r_i(t)$ refers to the instantaneous data rate in the last time slot while

$$R_i(t+1) = \frac{1}{T_w} r_i(t) p_i(t) + \left(1 - \frac{1}{T_w}\right) R_i(t) \tag{1}$$

is the moving average of the data rate. Here, $p_i(t)$ is a binary variable, which is equal to 1 when user i is scheduled at slot t and equal to 0 otherwise. Variable T_w denotes the time window, over which the moving average is computed.

3) Exponential Scheduler: The Exponential (EXP) [6] scheduler is queue-aware and intends to minimize the delay of buffered data in user queues. If a user queue gets large relative to the other users, its scheduling priority will be exponentially increased. Put formally, this scheduling rule chooses user

$$i^* = \arg\max_{\forall i \in \mathcal{N}} \frac{r_i(t)}{R_i(t)} \exp \frac{a_i q_i(t) - \frac{1}{N}\sum_{i=1}^{N} a_i q_i(t)}{1 + \sqrt{\frac{1}{N}\sum_{i=1}^{N} a_i q_i(t)}} \tag{2}$$

where $q_i(t)$ is the delay of the head-of-line packet awaiting transmission in the user queue, and is proportional to the queue length. The parameter a_i is an implementation parameter that reflects the strictness of the scheduler in prioritizing flow i.

C. Performance Metrics

We study the following four performance metrics:

- T_{Net}: the average network throughput which is measured during the downlink as the sum of the average data rate taken over all users of the network.
- J_{Net}: Jain's fairness index for user throughput and is computed as $(\sum_{i=1}^{N} T_i)^2 / N \sum_{i=1}^{N} T_i^2$ where T_i is the individual user throughput. We use this metric to compute the long-term throughput fairness of the network.
- $T_{\text{Slot}}^{10\%}$: the average 10th percentile user throughput w.r.t. time. We developed this metric to quantify the users' short-term starvation. For each user, we compute the received average throughput during a slot of one second, for several hundred seconds. Thereafter, we compute the 10th percentile throughput of these slots for each single user. If a user has a low $T_{\text{Slot}}^{10\%}$ this indicates that several time slots had poor throughput and, thus, user starvation. Finally, we average over all users.
- F^{LT}: the average amount of freezing experienced by all users in the network expressed as a percentage of the playback duration.

IV. Long-term Lookback Scheduling

Conventional schedulers are based on metrics computed at the scheduling BS with no regard to the users' scheduling history in previously traversed cells. In this section we present a scheduling framework that improves the long-term user QoS over multiple cells, without compromising short-term application needs.

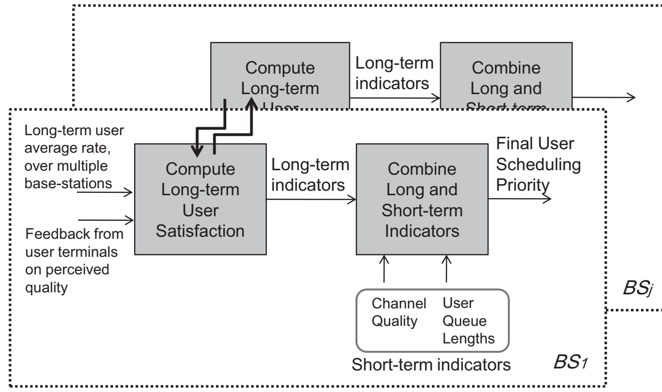


Fig. 3. The proposed Long-term Lookback Scheduling (LLS) framework.

A. Scheduling Framework

The proposed Long-term Lookback Scheduling (LLS) framework is shown in Fig. 3. Like existing schedulers, it uses short-term QoS indicators such as instantaneous channel gain and user queue lengths, which are evaluated at each BS. Our framework introduces a module to compute long-term user satisfaction, which is based on the average rate a user received over multiple cells, or more application specific QoS satisfaction indicators that may be fed back directly from user terminals. These long-term measures are computed over tens or hundreds of seconds and exchanged between BSs when users are handed over as shown in Fig. 3. Such an exchange may be possible via interfaces such as the X2-interface in LTE [4] and will not add significant overhead. This design allows a continuous assessment of user satisfaction over multiple cells without ignoring the short-term indicators a scheduler requires to meet application requirements.

Two main factors influence the overall scheduler design. The first is the definition of utility functions for the short and long-term indicators. Short-term indicators may follow exponential or quadratic utilities that impose thresholds on the acceptable performance limits. This will give them high influence in the overall scheduling decision if they depreciate, and will therefore direct the scheduler to improve the short-term indicators immediately. On the other hand, long-term metrics may be given more slowly varying utilities as the scheduler does not need to react immediately to their depreciation but can gradually improve the long-term indicator. The second design factor is how the short and long-term indicators are combined to achieve the overall scheduling utility. Here simple addition or multiplication may be used. In future work we will investigate the details of these design factors; but our focus in this paper is to illustrate the use of the proposed framework. As an example, we present long-term extensions of two classic scheduling rules.

B. Long-term Lookback Proportional Fair (LL-PF) Scheduler

The LL-PF scheduler is proposed to achieve a mix of short and long-term fairness over multiple cells. It provides the flexibility to serve a variety of applications while maintaining long-term fairness between users. This scheduler is only channel aware, and uses the average user-rate for scheduling decisions. As opposed to traditional rate-based schedulers, LL-PF computes the average user rate over both short and long durations. These two satisfaction indicators are then combined using an exponential function for the short-term user rate, and users are scheduled according to

$$i^* = \arg\max_{\forall i \in \mathcal{N}} \frac{r_i(t)}{R^{\mathrm{LT}}_{i,m(t)}(t)} \exp \frac{\alpha_i}{R_i^{\mathrm{norm}}(t)} \tag{3}$$

where R_i^{norm} is the short-term average rate (1), normalized by the highest user rate value leading to $R_i^{\mathrm{norm}} \in [0, 1]$. By choosing the exponential function for short-term utility, the user priority will increase exponentially as $R_i^{\mathrm{norm}}(t)$ decreases. Thereby, we ensure that users do not starve. In (3),

$$R^{\mathrm{LT}}_{i,m(t)}(t+1) = \frac{1}{T_w} r_{i,m(t)}(t) p_i(t) + \left(1 - \frac{1}{T_w}\right) R^{\mathrm{LT}}_{i,m(t-1)}(t) \tag{4}$$

denotes the long-term average user rate, over several base-stations, as computed at BS m. Therein, $m(t), m(t-1) \in \mathcal{M}$ are the respective BS indices in the current and previous time slot. If a user changes the cell, $m(t) \neq m(t-1)$ but the average is still computed as BSs exchange the value of $R_{i,m(t)}(t)$ during handover. Note that the time window T_w is significantly longer than in the computation of the short-term moving average (1). Choosing a large value for T_w, provides user fairness over a longer duration.

The free parameter α determines the rate at which the exponential factor of the short-term user rate increases. Increasing α will make the scheduler biased towards providing short-term fairness, and a value of 0 will make the scheduler a purely long-term multi-cell proportional fair scheduler, which we presented in [14]. Although the value of α depends on the application preference, we will see in Sec. V that LL-PF maintains a higher level of fairness than PF for any level of α.

C. Long-term Lookback Exponential (LL-EXP) Scheduler

The LL-EXP extends the channel and queue-aware scheduler from Sec. III-B3 to long-term fairness. Our extension keeps the instantaneous user queue size as a short-term scheduling indicator but replaces the average user rate $R_i(t)$ in (2) with the long-term average rate $R^{\mathrm{LT}}_{i,m(t)}(t)$. Depending on the user's trajectory, these long-term averages may be computed over several BSs. Moreover, we introduce a tuning parameter γ to control the scheduler bias towards the queue size short-term indicator. All in all, LL-EXP algorithm schedules user

$$i^* = \arg\max_{\forall i \in \mathcal{N}} \frac{r_i(t)}{R^{\mathrm{LT}}_{i,m(t)}(t)} \exp \gamma_i \frac{a_i q_i(t) - \frac{1}{N}\sum_{i=1}^{N} a_i q_i(t)}{1 + \sqrt{\frac{1}{N}\sum_{i=1}^{N} a_i q_i(t)}}. \tag{5}$$

LL-EXP considers the short-term status of user queues and therefore accommodates delay sensitive traffic. However, as opposed to (2) it also includes the long-term average user rate and can therefore provide long-term fairness. Also by relaxing the duration over which the user average is computed, the scheduler can be more opportunistic in serving users that have a high instantaneous rate $r_i(t)$. Results in Sec. V indicate that this reduces the likelihood of video freezing experienced by users. The parameter γ is similar to α in (3) and influences the rate of increase of the exponential term.

V. Performance Evaluation

In this section we first illustrate our scheduling framework for a very simple scenario. Then we study a more general simulation set-up and present performance studies for each of the long-term lookback schedulers 1) LL-PF is compared to traditional PF, and 2) LL-EXP is compared to EXP.

A. Simple Scenario

Consider the scenario of Fig. 4 where two users watching a video stream are moving towards Cell 3. User 1 is arriving from a congested cell and suffered excessive video freezing, whereas User 2 is coming from a sparsely populated cell and experienced better playback. As Cell 3 is also congested, both users will now be subject to video freezing. In the traditional scheduling approaches both users will suffer equally on arrival at BS_3. This, however, can be changed if BS_3 is made aware of the freezing history of User 1 in its previous cell. Now, BS_3 can increase the scheduling weight of this user to increase its QoS.

A sample result from applying a user's service history in the proposed LLS framework is shown by the bar plot in Fig. 4. Here, we can see that total freezing for User 1 is reduced from 28% to 21% compared to the case of scheduling without LLS (i.e. no service history from prior cells). User 2 on the other hand suffers slightly more freezing in LLS than in the case without LLS. This indicates that, with information from previous cells, LLS can allocate resources to provide a more fair video experience to the users. The total amount of freezing for both users has also been reduced.

B. Simulation Set-up

We evaluate the schedulers in a network of 19 cells with an inter-BS distance $D = 1\,\mathrm{km}$, and a BS transmit power of 40 W

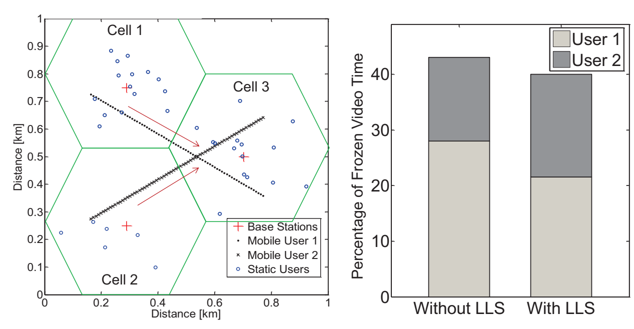
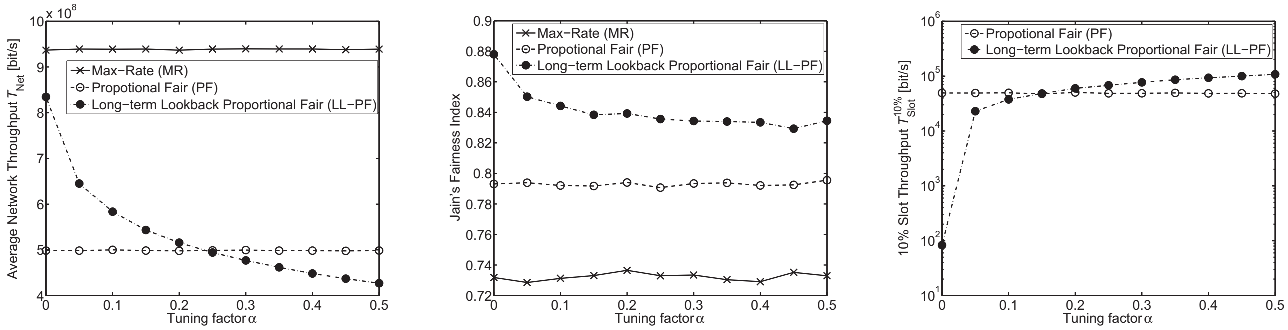


Fig. 4. Long-term Lookback Scheduling improves video quality experience and provides more fair service over multiple cells.



(a) Average network throughput T_{Net}

(b) Jain's fairness index

(c) 10th % slot throughput $T^{\text{slot}}_{10\%}$. $T^{\text{slot}}_{10\%} = 0$ for MR and is therefore not shown.

Fig. 5. Effect of α on the network throughput T_{Net}, Jain's fairness index, and 10th % slot throughput $T^{10\%}_{\text{Slot}}$, at a user speed of $S = 40\,\text{km/h}$, and $N = 250$.

for the downlink. The center carrier frequency is 2 GHz, and a bandwidth of 10 MHz for the full buffer traffic scenario and 5 MHz for the buffered video streaming traffic. Simulations run for 500 s simulated time, with a 200 s warm-up period, and are repeated 10 times.

C. Performance of LL-PF

For this study we assume that all the users have the same value of α, and set the time window T_w to 1 s for computing the short-term average, and to 300 s for the long-term average. We use a full buffer traffic model to study the schedulers' highest performance in a saturated network.

Fig. 5 illustrates the performance of LL-PF for different values of α. With $\alpha = 0$, the LL-PF throughput is much larger than PF and approaches the throughput of the MR scheduler. It also achieves the highest long-term fairness as shown in Fig. 5(b) as it is operating as a pure long-term scheduler. This, however, comes at the cost of a low $T^{10\%}_{\text{Slot}}$ value, indicating short-term starvation as observed in Fig. 5(c).

The extent of this starvation is illustrated in Fig. 6 which shows a sample user throughput over time for the different schedulers. The PF scheduler operates only on short-term and tries to provide a minimal throughput per second for each user. This comes at the cost of a reduced network throughput. On the other hand, LL-PF with $\alpha = 0$ results in times where the user is not served as it providing bursts of data to the user when its channel is best compared to its long-term average. When $\alpha = 0.05$, LL-PF provides a balance between guaranteeing short-term data rates, and provides bursts of high data. Therefore, as α increases $T^{10\%}_{\text{Slot}}$ increases, but network throughput decreases. This effect of α is detailed in Fig. 5. It is important to observe that the LL-PF scheduler can guarantee a higher level of short-term fairness than conventional PF if $\alpha > 0.2$ by achieving a higher $T^{10\%}_{\text{Slot}}$. In addition to this increased flexibility, Fig. 5(b) shows how the long-term fairness of the LL-PF scheduler is always higher than that of the PF scheduler, regardless of the setting of α. Thus, the proposed LL-PF scheduler can provide a range of short-long term objectives while *always* guaranteeing a higher level of long-term fairness over multiple cells.

D. Performance of LL-EXP

When studying the performance of the LL-EXP, we set $a_i = 1$ to assure equal priority to all users. We also set $R_{\text{Stream}} = 1.5\,\text{Mbps}$, and a traffic arrival rate $\lambda_i = 12\,\text{Mbps}$ at the BS to account for the higher core-network bandwidth. The user terminal buffer playback threshold is set to 5 seconds. These parameters were chosen to simulate users streaming a video-on-demand service such as Netflix.

Our two metrics of interest are the average network throughput T_{Net} and the average long-term video freezing F^{LT} experienced by the users. Fig. 7(a) shows the T_{Net} where we can observe a throughput gain for LL-EXP that increases with network load. This gain arises as a consequence of the long-term average R^{LT}which allows the scheduler to opportunistically exploit good channel conditions of users even when other users have had a low short-term data rate. On the other hand, the traditional EXP scheduler has two short-term indicators in its scheduling criterion as presented in (2). This causes the scheduler to have excessive emphasis on achieving a short-term data rate for each user, and thereby

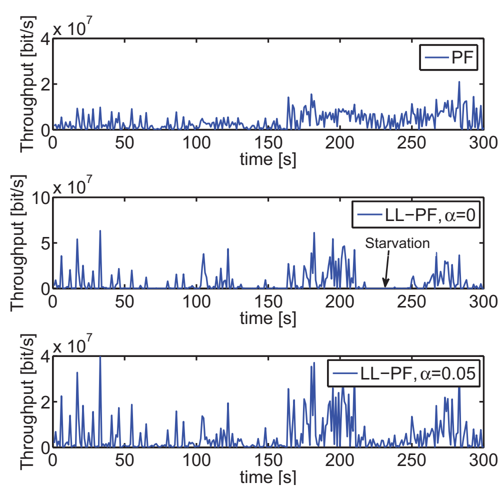
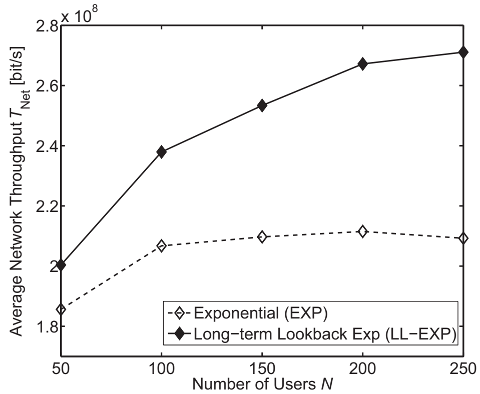
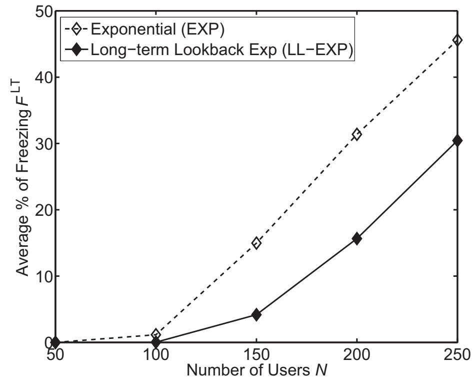


Fig. 6. Sample user throughput vs time for PF and LL-PF schedulers.



(a) Average network throughput T_{Net}



(b) Average % of freezing F^{LT}

Fig. 7. Effect of the number of users N on the network throughput T_{Net} and average video freezing F^{LT} at a user speed of $S = 40\,\mathrm{km/h}$.

prevents the scheduler from opportunistically serving users with high channel conditions. Fig. 7(a) illustrates this, where is it observed that the EXP scheduler throughput saturates quickly with an increasing number of users.

In Fig. 7(b) we also see the significant reduction in freezing of the LL-EXP scheduler, as it is able to support up to 160 users at a $F^{\mathrm{LT}} < 5\%$ whereas EXP can only support around 115. Therefore, the LL-EXP scheduler can improve throughput without sacrificing user experience. This is because the short-term queue length utility increases exponentially when user streaming deadlines are violated. Thus, the proposed LL-EXP scheduler provides throughput and long-term user experience gains.

VI. Conclusion

In this paper, we introduced the Long-term Lookback Scheduling (LLS) framework to increase the QoS of users traversing multiple cells. Our framework extends traditional schedulers by computing weights from information that was acquired over large time windows and previously visited cells. LLS also requires no central coordination and adds only insignificant signaling overhead. This long-term service notion can also be applied to stationary users to achieve long-term QoS objectives. Such a scheme can help avoid user frustration and improve subscriber retention in times of high traffic demand.

Our simulation results show that LLS can improve user satisfaction for both channel-aware and joint channel-queue-aware schedulers. We expect long-term scheduling to be a key approach for enabling constant high user satisfaction with small cells and uneven traffic distribution.

References

[1] G. Maier, F. Schneider, and A. Feldmann, “A first look at mobile hand-held device traffic,” in *Proc. Int. Conf. on Passive and Active Network Measurement (PAM)*, vol. 6032, pp. 161–170, Apr. 2010.

[2] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, “Understanding traffic dynamics in cellular data networks,” in *IEEE Int. Conf. on Computer Commun. (INFOCOM)*, pp. 882–890, May 2011.

[3] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, “A survey on 3GPP heterogeneous networks,” *IEEE Wireless Communications*, vol. 18, pp. 10–21, June 2011.

[4] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-Advanced for Mobile Broadband*. Academic Press, 2011.

[5] P. Viswanath, D. Tse, and R. Laroia, “Opportunistic beamforming using dumb antennas,” *IEEE Trans. Inf. Theory*, vol. 48, pp. 1277–1294, Jun. 2002.

[6] S. Shakkottai and A. Stolyar, “Scheduling algorithms for a mixture of real-time and non-real-time data in HDR,” in *Proc. Int. Teletraffic Congress (ITC)*, pp. 793–804, Sept. 2001.

[7] P. Marsch and G. Fettweis, eds., *Coordinated Multi-Point in Mobile Communications: From Theory to Practice*. Cambridge University Press, 2011.

[8] P. Frank, A. Muller, H. Droste, and J. Speidel, “Cooperative interference-aware joint scheduling for the 3GPP LTE uplink,” in *Proc. IEEE Int. Symp. on Personal Indoor and Mobile Radio Commun. (PIMRC)*, pp. 2216 –2221, Sept. 2010.

[9] T. Bu, L. Li, and R. Ramjee, “Generalized proportional fair scheduling in third generation wireless data networks,” in *Proc. IEEE Int. Conf. on Computer Commun. (INFOCOM)*, pp. 1–12, Apr. 2006.

[10] K. Son, S. Chong, and G. Veciana, “Dynamic association for load balancing and interference avoidance in multi-cell networks,” *IEEE Trans. on Wireless Commun.*, vol. 8, pp. 3566–3576, 2009.

[11] H. Zhou, P. Fan, and J. Li, “Global proportional fair scheduling for networks with multiple base stations,” *IEEE Trans. on Veh. Tech.*, vol. 60, pp. 1267–1879, 2011.

[12] 3GPP, “Physical layer aspect for E-UTRA,” Technical Report TR 25.814 V7.1.0, 3GPP, Oct. 2006.

[13] R. Pantos and W. May, “HTTP live streaming,” Informational Internet-Draft draft-pantos-http-live-streaming-07, IETF, Sept. 2011.

[14] H. Abou-zeid, S. Valentin, and H. Hassanein, “Long-term proportional fairness over multiple cells,” in *Proc. IEEE Int. Workshop on Wireless Local Networks (WLN), collocated with IEEE LCN*, Oct. 2012.