Efficient Lookahead Resource Allocation for Stored Video Delivery in Multi-Cell Networks

Hatem Abou-zeid* and Hossam S. Hassanein†
*Electrical and Computer Eng. Dept. Queen's University, Canada, h.abouzeid@queensu.ca
†School of Computing, Queen's University, Canada, hossam@cs.queensu.ca

***Abstract*—Novel transmission mechanisms are imperatively needed to cope with the exponential growth of mobile traffic and its associated power consumption. To address such challenges, we present *lookahead* video delivery schemes that *jointly* improve the streaming experience and reduce BS power consumption. This is accomplished by exploiting knowledge of *future* wireless rates users are anticipated to face. Such an approach is useful for delivering stored videos that can be strategically buffered in advance at the users' devices. For instance, a user leaving the cell center may have content prebuffered efficiently before poor channel conditions prevail. This will save energy as transmission will not be needed during poor conditions. In this paper, we first formulate Lookahead Resource Allocation (LRA) as a multi-cell optimization problem that leverages predictions of user mobility rates. Then, we present centralized and distributed algorithms that closely follow the benchmark results of the optimal solution. Numerical results demonstrate that significant improvements in video streaming and BS power consumption are achievable by the LRA strategies.**

I. Introduction

Mobile media has undoubtedly become the predominant source of traffic in wireless networks. This is being driven further by video sharing through social networking websites. Much of this is pre-recorded video content such as movies, TV shows, and short clips delivered from popular sites such as YouTube and Netflix. As video content dominates the overall network traffic, it is shall become the major contributor to the network energy consumption. With Base Stations (BSs) accounting for more than 50 percent of the network energy consumption [1], efficient video transmission is vital for both the end user and the provider. Consequently, novel paradigms for efficient video delivery are imperative to save energy and reduce operational costs.

In this paper, we investigate how predictions of user rates can be exploited for energy efficient transmission of stored videos that can be buffered at the user devices. The predictability of a wireless channel is generally possible due to the correlation between location and channel capacity [2]. Therefore, if a user's future location is known, upcoming data rates can be anticipated from radio maps that store average values of historic signal strengths at different geographic locations. Such maps can also be updated in real-time from User Equipment (UE) measurements [3], [4]. While such predictions are particularly plausible for users in public transportation, trains, or vehicles on highways, studies on human mobility patterns reveal a high degree of temporal and spatial regularity suggesting a potential 93% average predictability [5]. Furthermore, a plethora of navigation hardware and software is also available in today's smart phones for user reporting of current location and target destinations.

Being aware of a user's upcoming rate allows the network to plan spectrally efficient rate allocations without violating user streaming demands. For instance, if a user is moving towards the cell edge or a tunnel, the network can increase the allocated wireless resources allowing the user to buffer more video content. *Pre-buffering* this additional data then provides smooth video streaming since the user can consume the buffer while being in poor radio coverage. Additionally, by not serving users in such conditions, the network-wide spectral efficiency increases since valuable channel resources can be provided to other users instead. The idea is to grant users more air-time access at their highest achievable data rates and less access when they are at lower achievable rates. This allows the BS to transmit more data in less time, and consume lower transmit energy. Therefore, by leveraging knowledge of the users' future data rates the BS can devise lookahead resource allocation strategies that reduce BS transmission resources and enhance the streaming experience.

We summarize the main contributions of this paper in the following:

- We develop Lookahead Resource Allocation (LRA) that leverages rate predictions for efficient, enhanced delivery of stored videos. The problem is formulated as a multi-objective Linear Program (LP) which provides a benchmark solution. The formulation captures the trade-off between overall network streaming quality and BS power consumption.
- To efficiently solve the aforementioned LRA problem, we develop centralized and distributed algorithms (operating over multiple BSs). Results indicate that the proposed algorithms performs close to the LP benchmark, at a fraction of the memory and computation requirements.

We compare the performance of the LRA approaches through simulations and observe that BS power can be reduced by almost 50% while video degradations are simultaneously reduced by 40% compared to traditional allocation approaches. Our results demonstrate that LRA strategies are a promising mechanism for energy-efficient video delivery.

The rest of this paper is organized as follows. Section II reviews pertinent literature, while Section III introduces the system model and notation. In Section IV we describe the multi-objective LP formulation of LRA, while Section V

978-1-4799-3083-8/14/$31.00 ©2014IEEE

presents the proposed LRA algorithms. In Section VI, we evaluate the gains in video quality and BS power savings of the proposed schemes. Finally, we conclude the paper in Section VII.

II. Related Work

Exploiting user mobility trajectories (and the associated rate predictions) to optimize video delivery has been recently investigated with promising results. Yao et al. in [6] develop a rate adaptation algorithm that proactively switches to the predicted transmission rates based on a stored bandwidth map. The work in [7] uses similar maps to adapt video quality for smooth playback. With such an approach, a user headed to a tunnel will prebuffer several low quality segments in advance. While these works use radio maps to improve streaming, they do not optimize resource allocation based on these maps. This is investigated in our own work [8] where the multi-user rate allocation problem is solved over a time horizon to improve overall streaming smoothness over multiple users. The aforementioned works, however, focus on enhancing user experience but do not address energy efficieny.

The work in [9] and [10] are closer to this paper. In [9] rate predictions are used to minimize system utilization and avoid streaming delays. The authors present a non-convex formulation for the multi-user single cell case, and develop optimal algorithms for the single user case. In [10] we discuss the potential energy savings that can be achieved by a *mobility-aware* wireless access framework. An architecture is presented with the composite functional elements and their interaction is discussed. This paper differs from these earlier works in several aspects, 1) we formulate a detailed multi-user, multi-cell LRA problem as a multi-objective LP that captures a trade-off between overall network streaming quality and BS power consumption, and 2) we present centralized and distributed multi-cell algorithms that follow the pareto-optimal curve of the benchmark solution.

III. System Model

We use the following notational conventions: $\mathcal{X}$ denotes a set and it's cardinality $|\mathcal{X}|$ is denoted by X. We use bold letters to denote matrices, e.g. $\mathbf{x} = (x_{a,b} : a \in \mathbb{Z}_+, b \in \mathbb{Z}_+)$.

A. Network Overview

Consider a network with a BS set $\mathcal{K}$ and an active user set $\mathcal{M}$. An arbitrary BS is denoted by $k \in \mathcal{K}$ and a user by $i \in \mathcal{M}$. Users request stored video content that is transported using an HTTP-based progressive download mechanism. We assume that the wireless link is the bottleneck, and therefore the requested video content is always available at the BS for transmission.

B. Link Model and Resource Sharing

Time is divided in slots of equal duration τ, and is denoted by $n \in \mathcal{N}$, where the set of considered time slots is $\mathcal{N} = \{1, 2, \cdots, N\}$. In each slot, the wireless channel can be shared among multiple users during which the achievable data rate is assumed to be constant for each user. A typical value of such a coherence time τ is 1 s for vehicle speeds up to 20 m/s, during which average wireless capacity is not significantly affected. The achievable data rate depends on the path loss model $\text{PL}(d) = 128.1 + 37.6 \log_{10} d$, where the user-BS distance d is in km [11]. As we are interested in long-term allocation planning, we only consider average rate variations based on path-loss and neglect the variations from fast fading. The feasible link rate is then computed using Shannon's equation with SNR clipping at 20 dB. Therefore, a user i at slot n, will have a feasible data transmission of

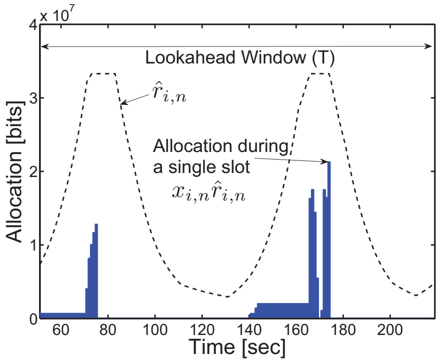


Fig. 1. Sample user allocation during N.

$$r_{i,n} = \tau B \log_2(1 + P_{\text{rx}_{i,n}} / N_o B) \quad \text{[bits]} \qquad (1)$$

where P_{rx}, N_o and B are the received power, noise power spectral density, and the transmission bandwidth respectively.

User link rates are assumed to be known for the upcoming N slots, which we call the *lookahead window*. The future link capacities are determined by computing the expected received power based on the knowledge of the future user locations, and then substituting in (1). This will generate a matrix of future link rates as defined by $\hat{\mathbf{r}} = (\hat{r}_{i,n} : i \in \mathcal{M}, n \in \mathcal{N})$. Fig. 1 illustrates an example of $\hat{r}_{i,n}, \forall n$ for a user traversing two BSs along a highway. In this paper, we assume that knowledge of $\hat{\mathbf{r}}$ is error free to provide the bounds of the potential gains.

BS air-time is shared among the active users during each slot n. We define the resource allocation matrix $\mathbf{x} = (x_{i,n} \in [0, 1] : i \in \mathcal{M}, n \in \mathcal{N})$ which gives the fraction of time during each slot n that the BS bandwidth is assigned to user i. The rate received by each user at each slot is the element-wise product $\mathbf{x} \odot \hat{\mathbf{r}}$. A sample allocation $x_{i,n}, \forall n \in \mathcal{N}$ for a user is illustrated in Fig. 1, where the bars indicate the proportion of $\hat{r}_{i,n}$ allocated to that user. Note that since a user can traverse multiple cells during N, BS cooperation is needed to make the allocation plan, which is assumed possible via an inter-BS interface such as the X2 in Long Term Evolution (LTE).

User-BS association is based on the strongest received signal, and is assumed to be known during N. We define the set $\mathcal{U}_{k,n}, k \in \mathcal{K}, n \in \mathcal{N}$ which contains the indices of all the users associated with BS k at time slot n.

C. BS Power Consumption Model

The BS downlink power consumption is based on the linear load dependent power model [12]. BS power is proportional to the BS load, with a fixed power required at minimum load. For BS k at slot n, this can be expressed as follows:

$$p_{k,n} = P_0 + (P_m - P_0)\,\mathrm{BS}^{\mathrm{load}}_{k,n}, \qquad 0 < \mathrm{BS}^{\mathrm{load}}_{k,n} \leq 1, \tag{2}$$

where P_m and P_0 are the power consumption at the maximum and minimum non-zero load, and BS load is computed as $\mathrm{BS}^{\mathrm{load}}_{k,n} = \sum_{i \in \mathcal{U}_{k,n}} x_{i,n}$.

IV. Efficient Lookahead Video Transmission: Problem Formulation

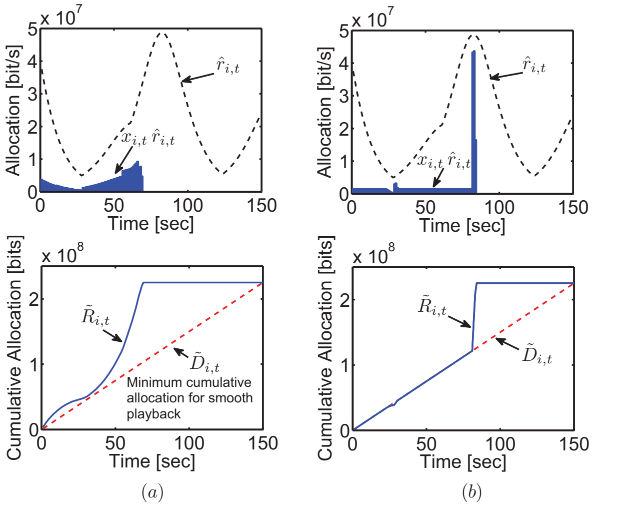
As opposed to live streaming, stored videos can be delivered ahead of time and cached at the UE, after which transmission can be momentarily suspended while the user consumes the buffer. In this section, we formulate the Lookahead Resource Allocation problem for stored video delivery that exploits user rate predictions over multiple cells. The objective is to minimize network-wide BS power consumption, and offer a power-quality pareto optimal trade-off.

A. Lookahead BS Power Minimization

Consider a user requesting a stored video at slot $n = 1$, with a streaming rate of V [bps]. We denote the minimum cumulative video content required for smooth streaming at each time slot as $\tilde{D}_{i,n} = V\tau n$, which is represented with a dashed line in Fig. 2. The cumulative allocation made to a user i by slot n is denoted by $\tilde{R}_{i,n} = \sum_{n'=1}^{n} x_{i,n'}\, \hat{r_{i,n'}}$. If $\tilde{R}_{i,n} \geq \tilde{D}_{i,n}\, \forall\, n$ then user i will experience smooth playback. Fig. 2 illustrates how BS transmission time can be minimized by leveraging user rate predictions. We can see that traditional RA schemes, unaware of the users future channel conditions, will continue to serve the user even when in poor channel conditions. On the other hand, a *lookahead* scheme that is aware of the user's future rate, will wait to make bulk transmissions at times of high channel conditions, while making the minimal transmissions that ensure $\tilde{R}_{i,n} \geq \tilde{D}_{i,n}$ at other times. This achieves lower airtime usage, resulting in lower power consumption or more resources for other services.

We consider a multi-user, *multi-cell* scenario, and a lookahead window of N slots. The problem of minimizing network wide BS power without violating streaming requirements is equivalent to minimizing the BS air-time due to the linear load dependent power model of (2). This can be formulated as the LP:

$$\begin{aligned}
\underset{\mathbf{x}}{\text{minimize}} \quad & \sum_{n=1}^{N}\sum_{i=1}^{M} x_{i,n} && \qquad (3)\\
\text{subject to:} \quad & \text{C1:} \quad \sum_{i \in \mathcal{U}_{k,n}} x_{i,n} \leq 1, && \forall\, k \in \mathcal{K}, n \in \mathcal{N},\\
& \text{C2:} \quad \tilde{D}_{i,n} - \tilde{R}_{i,n} \leq 0, && \forall\, i \in \mathcal{M}, n \in \mathcal{N},\\
& \text{C3:} \quad 0 \leq x_{i,n} \leq 1 && \forall\, i \in \mathcal{M}, n \in \mathcal{N}.
\end{aligned}$$



(a) (b)

Fig. 2. Buffered streaming allocation using (a) traditional schemes, and (b) rate predictions to mimimize resource utilization.

Constraint C1 expresses the resource limitation at each base station. It ensures that the sum of the air-time of all users associated with a BS k is equal to 1 at every time slot. C2 ensures that the cumulative video content requirement is not violated at each time slot. Finally, C3 provides the bounds for the resource allocation factor. Note that the outer summation over time slots in (3) is to minimize the sum air-time consumed during the window of N slots.

B. Joint Power-Video Degradation Minimization

At medium to high load, it will not be possible to satify constraint C2 in (3) for all users and all time slots. When $\tilde{R}_{i,n} < \tilde{D}_{i,n}$, the user experiences video stalling, or a lower quality video, and therefore the video experience is degraded. We define Video Degradation (VD) as the amount of unfulfilled video demand. For a given user allocation $x_{i,n}$, it is defined at each slot n as the cumulative *positive* difference between the requested demand and the allocation:

$$\mathrm{VD}_{i,n} = [\tilde{D}_{i,n'} - \sum_{n'=1}^{n} x_{i,n'}\, \hat{r}_{i,n'}]^{+} \tag{4}$$

When $\sum_{n'=1}^{n} x_{i,n'}\, \hat{r}_{i,n'} > \tilde{D}_{i,n'}$ it implies that future video content is prebuffered, and $\mathrm{VD}_{i,n} = 0$. On the other hand, if the converse holds, then the user will experience video degradation. Therefore, VD represents the amount of unfulfilled video demand. The average network VD over N slots can be computed as $\mathrm{VD}_{\mathrm{Net}} = \frac{1}{NM}\sum_{n=1}^{N}\sum_{i=1}^{M} \mathrm{VD}_{i,n}$.

The objective now is to exploit the rate predictions to make the optimal pre-buffering allocations to users, that achieve a tradeoff between minimizing the sum user VD and the consumed BS air-time. This is formulated as the following multi-objective optimization problem:

$$\begin{aligned}
\underset{\mathbf{x}}{\text{minimize}} \quad & \sum_{n=1}^{N}\sum_{i=1}^{M}\left(\frac{w_{\mathrm{VD}}}{\tilde{D}}\mathrm{VD}_{i,n} + \frac{w_{\mathrm{Air}}}{MN}\, x_{i,n}\right) && \qquad (5)\\
\text{subject to:} \quad & \mathrm{C1}, \mathrm{C3},
\end{aligned}$$

where w_{VD} and w_{Air} are weights $\in [0,1]$ for VD and air-time minimization respectively, and $\tilde{D} = \sum_{n=1}^{N}\sum_{i=1}^{M}\tilde{D}_{i,n}$, and MN are normalization constants for each objective. When $w_{\text{VD}} = 1$ and $w_{\text{Air}} = 0$, (5) will minimize user degradations irrespective of the consumed BS air-time. As w_{Air} increases, the problem will trade-off the reduction in VD against the additional air-time required. This trade-off is particularly useful for cases when users experience prolonged conditions of poor coverage, during which the air-time spent will not result in quality improvements. Note that (5) is non-linear due to the $(\cdot)^+$ operator in 4. However, since it is piece-wise linear and convex, and the constraints are linear, problem (5) can be reformulated as the following LP, which we refer to as LRA-LP:

$$\underset{\mathbf{x},\mathbf{Y}}{\text{minimize}} \quad \sum_{n=1}^{N}\sum_{i=1}^{M}\left(\frac{w_{\text{VD}}}{\tilde{D}}Y_{i,n} + \frac{w_{\text{Air}}}{MN}\,x_{i,n}\right) \tag{6}$$

$$\begin{aligned}
\text{subject to:}\quad & \text{C1}, \text{C3}, \\
& \text{C4}\colon \tilde{D}_{i,n} - \tilde{R}_{i,n} - Y_{i,n} \leq 0, \quad \forall\, i \in \mathcal{M}, n \in \mathcal{N} \\
& \text{C5}\colon Y_{i,n} \geq 0, \quad \forall\, i \in \mathcal{M}, n \in \mathcal{N}.
\end{aligned}$$

Here we have introduced $Y_{i,n}$ as additional optimization variables which we restrict to have positive values in C5. The value of $Y_{i,n}$ therefore captures the degradation only (i.e. when $\tilde{D}_{i,n} > \tilde{R}_{i,n}$), and remains unaffected if content is pre-buffered.

The multi-objective problem in (6) although linear, has a large number of constraints and optimization variables, which increase dramatically and as N increases. This can be solved with large-scale LP solvers such as Gurobi [13] but will require significant memory and considerable time. Therefore, (6) can serve as an offline performance benchmark, whereas for real-time implementation, we present two heuristic algorithms in the following section.

V. Lookahead Video Transmission Algorithms

The main idea of the proposed video transmission algorithms is to first to keep track of the *cumulative* rates allocated to users up to slot n. Then, the *future* video degradations users are predicted to experience are estimated, and a heuristic is introduced to make resource allocations $x_{i,n}$ that reduce user VDs without consuming excessive BS resources. We first present an iterative algorithm that requires a central BSs to make the allocation decisions for all the cooperating BSs. Then, we show how the algorithm can be extended to operate in a distributed fashion.

A. Centralized LRA Algorithm

The objective of this algorithm is to jointly minimize VD_{Net} and BS air-time as in the pareto-optimal formulation of (5). It consists of the following steps:

- *Step 1*: Initialize $x_{i,n} = 0$ for all the users and time slots.
- *Step 2*: Compute the *future* $\tilde{\text{VD}}_i$ each user will experience at slot n. This is determined based on the current cumulative allocation at slot n, and a tentative air-time allocation for the upcoming slots $n+1, n+2, \cdots, N$, i.e. $\tilde{\text{VD}}_{i,n} = \sum_{n'=n}^{N}[V\tau n' - \sum_{n''=1}^{n'} x_{i,n''}\,\hat{r}_{i,n''}]^+$, where n' and n'' are dummy variables.

Algorithm 1 Centralized LRA Algorithm

```
Require: $\hat{r}_{i,n}, \mathcal{U}_{k,n}, V, \tau, , M, K, N$
Initialize $x_{i,n}, \tilde{R}_{i,n} = 0 \quad \forall\, i, n$
repeat {allocation iterations}
    Calculate $\text{VD}_{\text{Net}}$ before allocation.
    for all time slots $n$ do
        Reset $x_{i,n} = 0 \quad \forall i.$
        for all base stations $k$ do
            for all users $i \in \mathcal{U}_{k,n}$ do
                Calculate $\tilde{\text{VD}}^{i}_{k,n}$ using (7)
            end for
            Set $x_{i^*,n} = 1$ to $i^*$ that achieves $\tilde{\text{VD}}^{i^*}_{k,n} \leq \tilde{\text{VD}}^{i}_{k,n}$
            only if $\tilde{\text{VD}}^{i^*}_{k,n-1} - \tilde{\text{VD}}^{i^*}_{k,n} > \gamma$
        end for
    end for
    Calculate $\text{VD}_{\text{Net}}$ after allocation.
until {no more decrease in $\text{VD}_{\text{Net}}$}
return $\mathbf{x}$
```

- *Step 3*: Each BS performs a greedy allocation to minimize VD_{Net}. It finds the user that when allocated the full air-time at slot n reduces VD_{Net} the most. To do so, the BSs first compute the *sum* of future VD of all users $\in \mathcal{U}_{k,n}$, that results from allocating the full air-time to user i and nothing to users $i^{'} \in \mathcal{U}_{k,n}\backslash\{i\}$ (the other users in the BS):
$$\tilde{\text{VD}}^{i}_{k,n} = \sum_{i \in \mathcal{U}_{k,n}}\sum_{n'=n}^{N}\text{VD}_{i,n'} \tag{7}$$
where $x_{i,n} = 1$ and $x_{i^{'},n} = 0 \quad \forall i^{'} \in \mathcal{U}_{k,n}\backslash\{i\}$. After computing (7) $\forall i \in \mathcal{U}_{k,n}$, the bandwidth is allocated to user i^*, that achieves $\tilde{\text{VD}}^{i^*}_{k,n} \leq \tilde{\text{VD}}^{i}_{k,n}$. The idea of this allocation metric is to choose the user that when allocated the airtime will result in the lowest overall future BS video degradation. This means that ideally the selected user needs to have a good *current* channel quality, and poor future conditions relative to the other users. When selected the user will achieve the best *reduction* in future BS VD.
- *Step 4*: To introduce the BS airtime trade-off, the user allocation result of step 3 is applied only if the resulting improvement in BS VD before and after allcoation, is larger than a threshold γ, i.e. $\tilde{\text{VD}}^{i^*}_{k,n-1} - \tilde{\text{VD}}^{i^*}_{k,n} > \gamma$. A larger value of γ will introduce more weight to the air-time reduction objective.
- *Step 5*: Repeat steps 2 to 4 for all $n \in \mathcal{N}$.
- *Step 6*: Calculate VD_{Net}.
- *Step 7*: Repeat steps 2-5 until there is no more decrease in VD_{Net}.

Note in the first iteration, $\mathbf{x} = 0$ in the computation of (4), and therefore step 3 will not exploit future VD information.

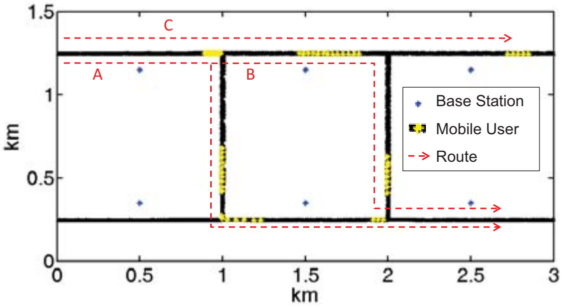


Fig. 4. Road network with mobility routes.

However, subsequent iterations of steps 2-5 will allocate $x_{i,n}$ based on the values of $x_{i,n'}$ $\forall n' = n+1, n+2, \ldots N$ of the previous iteration. As $\hat{r}_{i,n}$ does not change over iterations, the selection in step 3, changes in the direction of decreasing VD. Typically, the algorithm converges within 4 to 6 iterations, as observed in the numerical results in Section VI. The complete procedure is presented in Algorithm 1, which we refer to as LRA-Alg.

B. *Distributed LRA Algorithm*

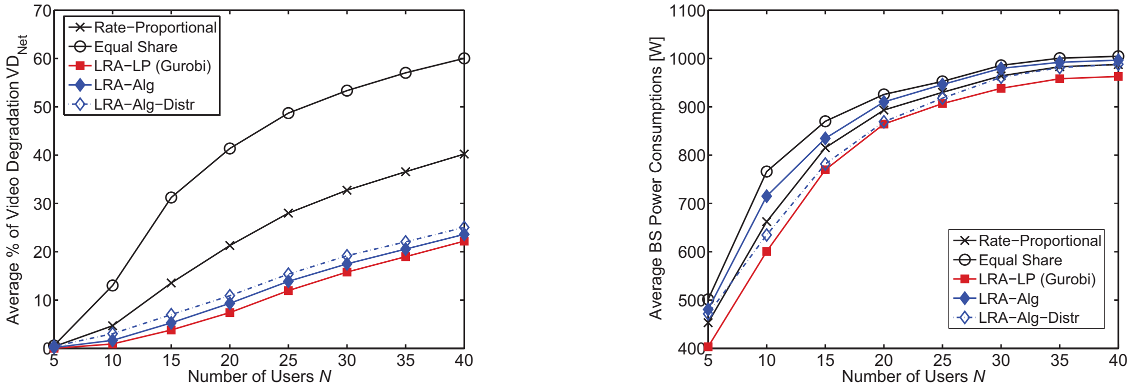
The aim of the distributed LRA algorithm is to allow each BS to perform its own lookahead resource allocation. To account for network-wide rate predictions, each BS will have a rate map (or radio map) of the cooperating region of interest (e.g. several BSs along a highway). At the start of the prediction interval, BSs will exchange the rate predictions of the users currently under their service. Then, instead of initializing $x_{i,n} = 0$ as in step 1 of the LRA-Alg, each BS will perform a temporary allocation for the upcoming N slots, where $x_{i^*,n} = 1$ for the user i^* predicted to have the highest channel rate among all the users at slot n. To do so, the cooperating BS require the initial rate prediction vectors of all the users. Then, with this as the baseline allocation, steps 2 to 5 of LRA-Alg are performed independently at each BS, with no further iterations. The intuition of this distributed procedure is that each BS first makes an initial baseline allocation based on a MaxRate scheme, and thereafter uses the VD metrics to adjust the allocations based on the procedure in LRA-Alg. An important consideration is that for users handed over, the computation of (4) will not be possible in this distributed implementation as the previous values of $x_{i,n}$ are unknown to the target BS accepting the user. This can be circumvented if the user reports its buffer status during handover for cases of prebuffered content, or it's measure of VD in cases of buffer underflow. Alternatively, the serving BS can also report this information to the target BS during handover. We will refer to this algorithm is LRA-Alg-Distr.

VI. Numerical Results

A. *Simulation Set-up*

We consider the six BS network shown in Fig. 4, where realistic vehicular mobility is generated on the roads using the SUMO traffic simulator [14]. Vehicles traverse the three routes denoted by A, B and C in Fig. 4 with equal probability. Gurobi 5.1 [13] is used to solve the optimization problems, and Matlab was used as a simulation environment. We assume a BS transmit power of 40 W, a center carrier frequency of 2 GHz, and a bandwidth of 10 MHz. The video streaming rate V is set to 3 Mbits, and the lookahead window N to 250 slots, with a slot duration τ of 1 s. BS power consumption at minimum and maximum load is 200 W and 1300 W respectively as presented for macro BSs employing time-domain duty-cycling in the power model of [12].

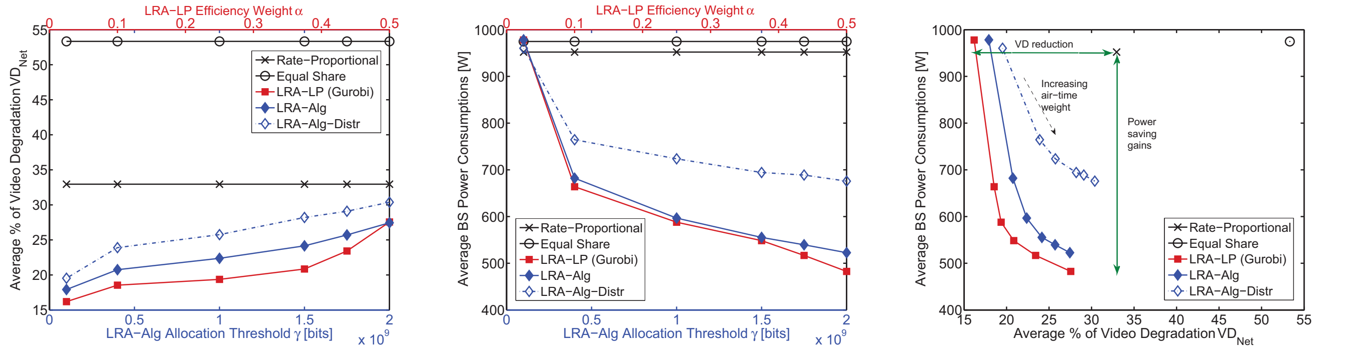
We compare the performance of the LRA schemes against two baseline approaches that do not exploit rate predictions: Equal Share (ES) and Rate-Proportional (RP). In ES, airtime is shared equally among the users at each time slot. The RP allocator is designed to be more spectrally efficient but not completely fair to users. Here, the airtime assigned to each user i at slot n is in proportion to the achievable data-rate $\hat{r}_{i,n}$ of that user, i.e. $x_{i,n} = \hat{r}_{i,n} / \sum_{i \in \mathcal{U}_{k,n}} \hat{r}_{i,n}$.



(a) Video Degradation VD$_{\text{Net}}$: $w_{\text{VD}} = 1, w_{\text{Air}} = 0$ and $\gamma = 0$.

(b) BS Power Consumption: $w_{\text{VD}} = 1, w_{\text{Air}} = 0$ and $\gamma = 0$.

Fig. 3. Video degradation VD$_{\text{Net}}$ and BS Power Consumption for varying number of users.



(a) Video Degradation VD_{Net}.

(b) BS Power Consumption

(c) Pareto-optimal trade-off of VD_{Net}-BS Power Consumption for the different allocation algorithms; $M = 30$.

Fig. 5. Trade-off of VD_{Net} and BS Power Consumption for varying LRA-LP weights ($w_{\text{VD}} = 1 - \alpha, w_{\text{Air}} = \alpha$) and LRA-Alg γ values.

B. Performance Evaluation

Fig. 3(a) illustrates the video performance of the LRA schemes with the objective of minimizing degradations without regard to BS power consumption (i.e. $w_{\text{VD}} = 1, w_{\text{Air}} = 0$ and $\gamma = 0$). Significant gains (upto 45% reduction in VD) are observed compared to the traditional schemes that do not lookahead at future user rates. Additionally, both of the proposed centralized and distributed LRA algorithms achieve close to optimal performance, indicating their effectiveness in minimizing video degradations. Fig. 3(b) shows the corresponding BS power consumption where all schemes have somewhat similar performance, which is expected as power reduction was not considered in this setting.

In Fig. 5(a) and Fig. 5(b) we demonstrate the potential VD-Power trade-off that can be achieved with the proposed LRA video delivery schemes. As α and γ increase, the LP formulation and the LRA algorithms decrease the power consumption at the cost of an increase in VD. This trade-off is summarized in Fig. 5(c) which illustrates the pareto-optimal trade-off between VD_{Net} and BS power. The significant (simultaneous) gains in VD and power are evident over the ES and RP allocation schemes. We observe that with $\alpha = 0.3$ BS power is reduced by almost 50% while VD is simultaneously reduced by 40% compared to RP allocation. The figure also demonstrates how the proposed centralized LRA-Alg scheme closely follows the pareto-optimal benchmark curve of the LRA-LP that is solved offline using Gurobi [13]. Finally, while the distributed scheme (LRA-Alg-Distr) offers considerable gains over RP, its performance deviates from the pareto-optimal benchmark as γ increases.

VII. Conclusion

In this paper, we presented lookahead resource allocation strategies that minimize video degradations of stored streams while simulataneously considering BS power consumption. To provide a benchmark, we first formulated a multi-objective LP that captures the required trade-off between minimizing total video degradation and minimizing network-wide BS power consumption. Then, we presented a centralized heuristic algorithm that closely follows the LP solution. A distributed extension of the algorithm was also developed, for more practical implementation and results demonstrate its effectiveness. All three approaches achieve significant gains in power saving and video quality improvements, thereby demonstrating that LRA strategies are a promising mechanism for energy-efficient content delivery platforms. Future work includes investigating the effects of errors in the rate predictions on LRA.

References

[1] T. Chen, Y. Yang, H. Zhang, H. Kim, and K. Horneman, "Network energy saving technologies for green wireless access networks," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 30–38, 2011.

[2] M. Malmirchegini and Y. Mostofi, "On the spatial predictability of communication channels," *IEEE Trans. Wireless Commun.*, vol. 11, pp. 964–978, Mar. 2012.

[3] J. Johansson, W. Hapsari, S. Kelley, and G. Bodog, "Minimization of drive tests in 3GPP release 11," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 36–43, 2012.

[4] OpenSignal, "The OpenSignal project homepage." http://opensignal.com/, 2013.

[5] C. Song, Z. Qu, N. Blumm, and A. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, pp. 1018–1021, 2010.

[6] J. Yao, S. Kanhere, and M. Hassan, "Improving QoS in high-speed mobility using bandwidth maps," *IEEE Trans. Mobile Comput.*, vol. 11, pp. 603–617, Apr. 2012.

[7] H. Riiser, T. Endestad, P. Vigmostad, C. Griwodz, and P. Halvorsen, "Video streaming using a location-based bandwidth-lookup service for bitrate planning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 8, pp. 24:1–24:19, Aug. 2012.

[8] H. Abou-zeid, H. S. Hassanein, and N. Zorba, "Enhancing mobile video streaming by lookahead rate allocation in wireless networks," in *Proc. IEEE Consumer Commun. and Netw. Conf. (CCNC)*, Jan. 2014, to appear.

[9] Z. Lu and G. de Veciana, "Optimizing stored video delivery for mobile networks: The value of knowing the future," in *Proc. IEEE Int. Conf. on Computer Commun. (INFOCOM)*, pp. 2806–2814, July 2013.

[10] H. Abou-zeid and H. S. Hassanein, "Predictive green wireless access: Exploiting mobility and application information," *IEEE Wireless Commun.*, vol. 20, pp. 92–99, Oct. 2013.

[11] 3GPP, "LTE;E-UTRA; radio frequency system scenarios," Technical Report TR 36.942 V11.0.0, 3GPP, Sept. 2012.

[12] C. Desset, B. Debaillie, V. Giannini, A. Fehske, G. Auer, and H. Holktamp, "Flexible power modeling of lte base stations," in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, pp. 2858–2862, Apr. 2012.

[13] Gurobi Optimization. http://www.gurobi.com/.

[14] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, "Sumo - simulation of urban mobility: An overview," in *Proc. Third Int. Conf. on Advances in System Simulation (SIMUL 2011)*, pp. 63–68, Oct. 2011.