Towards a Semantic-aware Radio Resource Management

Luis Guillermo Martinez Ballesteros, Cicek Cavdar, Pietro Lungaro, Zary Segall
Mobile Services Lab
KTH Royal Institute of Technology
Kista, Sweden
lgmb@kth.se, cavdar@kth.se, pietro@kth.se, segall@kth.se

***Abstract*— In this paper semantic-aware model for radio resource management in wireless networks is introduced and studied through simulation. By semantic-awareness, the network can selectively manage the radio resource allocation based on the evaluation of transferred content, and its associated processing, and prioritize users that are close to experience interruptions, in order to improve the wireless resource utilization and the user's Quality of Experience (QoE). Different radio resource management (RRM) strategies are proposed and investigated, considering buffer capacity at the terminals and the experience of the users in time while watching a video and waiting for resource allocation. The simulation results show that the users can reduce the total duration, frequency and length of the interruptions during a playback by applying semantic-awareness in the radio resource allocation, which might affect positively user's QoE.**

Keywords-Radio Resource Management; resource allocation

I. INTRODUCTION

Recent introduction of new generation of wireless infrastructures is being accompanied by an increase in both the number of users and their interest in multimedia content. This growth has been driven in the last decade by the popularity of multimedia content (e.g. video-sharing websites, social networks, video on demand sites, mobile IPTV, etc.), that according to the tendency will generate much of the mobile traffic growth through 2016, showing, at the same time, the highest growth rate of any mobile application ([1]). Before this scenario, a common approach to reach the goal of high quality information delivery has been the implementation of resource management schemes and scheduling algorithms to optimize resource allocation and traffic distribution as function of network parameters ([2]-[13]). Solutions have evolved from a perspective mainly centred on the evaluation of network based constraints (e.g., Signal to Noise Ratio or instant data rates) deprived of knowledge about the transferred content [3], to a perspective where inherent characteristics of the content are considered to improve network performance. In some cases ([11] [12]), the video distortion level is used to calculate the rates to deliver a multimedia content in a gradient-based scheduling and resource allocation scheme. Even though these studies consider the evaluation of content status to allocate resources, their objective is to maximize the average peak signal-to-noise ratio (PSNR) of all video users, which not always impacts positively the QoE. In [13], a resource allocation scheme that considers both the average rate achieved so far and the future expected rate is proposed with the goal of maximizing sigmoid function of the average bit rate. Prediction does not consider what happens to the content in the terminal by establishing a direct relation between the bit rate and the QoE. In ([2][9]), authors improve system throughput by allocating resources according to predefined utility functions to measure the QoE and QoS respectively, without considering how the content is processed at the terminal side. However, the idea of maximizing performance through infrastructure improvements and adjustment of network parameters is usually not optimal with respect to user perceived quality for multimedia applications [2]. In this paper, we want to investigate the effect of using semantic information (i.e., buffer capacity, player data rate, waiting times) provided by users terminals on the radio resource allocation in the downlink transmissions (base station (BS) to device) in mobile networks and its impact on the user's perception. In particular, some QoE related parameters (i.e., duration, the length and the frequency of the interruptions) are quantified to provide an initial measure of the effect of incorporating semantic-awareness to wireless infrastructures.

The rest of the paper is organized as follows: In Section II, we present the semantic-aware proposed architecture and RRM schemes considered in this study. In Section III, we describe the simulation settings and performance measurements considered in the paper. In Section IV, we present the results obtained with the simulation scenario. We conclude the paper in section V.

II. SYSTEM MODEL

A. Network Description and Service Model

Semantic-aware networks are infrastructures with the capacity to selectively manage the information flow depending on its importance from an application point of view. Unlike the concept of content-awareness, where the network management is considering the type of content in a static way, semantic-aware approach requires infrastructures with the capability to exchange information dynamically with the terminal and evaluate the content related information provided by users (i.e., both specific details regarding transferred content and the status of their processing in the terminal) in order to selectively manage the information flow, and distribute resources depending on the

 ISBN: 978-1-61208-313-1

applications performance. In this regard, two elements are essential for the operation of the semantic-aware engine: a semantic-aware resource manager (SRM) and a semantic client (SC) (Figure 1).

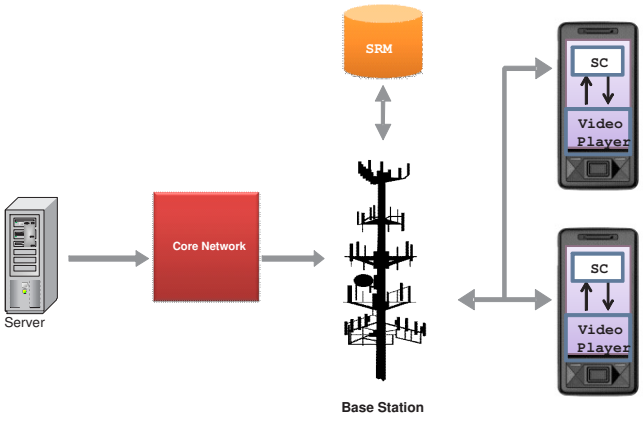

Figure 1. Example of the main elements composing the Semantic-aware system.

The SRM is centrally located in the base station and collects the reports provided by the user terminals. It keeps track of the terminals and the processing content current status. The SCs report information regarding the buffer capacity at the terminal, the player data rate consumption, and the users waiting times to the semantic-aware resource manager. SCs are software applications with collecting and sensing functionalities installed in the mobile terminals. Once the gathered data is passed to the SRM, this selects the time instants more appropriate for allocating resources and delivering content to individual users according to the operator's goal, by applying a scheduling policy.

At the fixed side of the infrastructure, the BS is connected to a multimedia server with capacity to store different multimedia files of size S_v bits. With regard to the wireless side of the architecture, we assume that the total user population of size N_u is uniformly distributed and downloading multimedia files, v. Each user $l \in \{1,\ldots, N_u\}$ is interested in receiving a maximum of M items, all of them with size S_v and duration in seconds T_v. Once a user l requests content, information will be downloaded at an instantaneous data rate from the BS to the user l at time t, $R_l(t)$. Downloaded information is placed into a buffer B_l of infinite capacity before being effectively consumed by user l. An initial buffering time of b seconds is considered. This time b, counted only once by a multimedia file, corresponds to the interval between the first request time and the time at which the effective consumption of bits by user l from the buffer B_l starts. Once the time b has elapsed, the playback starts and bits from the buffer B_l will be consumed at time t with a data consumption rate C_l bps. C_l is a constant value that depends on the size and the duration of time that needs to be spent to watch the multimedia file v requested by user l, S^l_v and T^l_v respectively. So,

$$C_l = \frac{S^l_v}{T^l_v} \quad (1)$$

Content processing will continue until the file stored in the server has been consumed by user l. Duration of processing one multimedia file will determine how long users will last in the system, and the request for a new multimedia file will stay active until M multimedia files has been processed by user l.

B. Semantic-aware Resource Management Schemes

In our implementation, all proposed semantic-aware schemes follow a similar procedure to assign resources every slotted time interval n of duration Δt. The SRM starts by detecting how close the users are of experiencing a lack of resources in the buffer B_l that can affect the correct processing of the information and the user perception. Proximity of shortage is measured in terms of the video watching time left in the buffer at time t, T_b seconds, given the $C_l(t)$ rate. So,

$$T_b = \frac{B_l(t)}{C_l} \quad (2)$$

This identification process based on the evaluation of the T_b value lead to a classification of the users in two queues, one with those users with imminent shortage and other filler with those with no imminent shortage, called X and Y, respectively. If size of queue X is equal to one, the user l in that queue receives the resource with no consideration of the users present in the queue Y. In contrast, if the length of the queue X is more than one, users in the queue will be ranked in descending order considering the utility function of the RRM scheme. Then the scheduler allocates the resource to the user in the top of the ranking. If queue X is empty, the procedure described before will be executed with the users placed in the queue Y. Supported by T_b other values extracted from the semantic information, the scheduler will look first at the users of X group. If size of X is equal to one, that user l receives the resource. In contrast, if there is more than one in the set X, users will be ranked in descending utility order considering the criteria of the semantic aware RRM scheme. Then the scheduler allocates the resource to the user in the top of the ranking. Different RRM used in this study, including the reference case, and the utility functions linked to them are described below:

1) Proportional Fair (PF): This RRM strategy represents the reference case or no semantic-aware scheme. In this case, users are not classified according to T_b value. By contrast, there is no buffer capacity consideration. At each time interval n, this scheme assigns the resource to the user with the largest ratio $\left(R_l/A_l\right)$ where R_l is the instant download data rate achievable by user l and A_l is the average download data rate of user l.

2) Buffer based (BB): This RRM strategy tries to allocate resources to the user with the smallest video watching time from the buffer based on the evaluation of T_b by considering current buffer capacity. This scheduler tries to allocate a time slot to the user with the higher imminence of having an interruption. Utility function for grading the users is the inverse of the T_b, $\left(1/T_b\right)$.

 ISBN: 978-1-61208-313-1

3) Inactive online time based (IB): This RRM uses the time a user has been active in the system but with no wireless resource assigned to download bits. We evaluate the total time a user l has been active in the system, with outstanding bits in the server to download, but without a wireless resource assigned to place bits in the buffer or T^l_{wait} in seconds. Utility function for grading the users is the current $\left({}^{1}/_{T^l_{wait}}\right)$.

4) Active online time based (AB): This scheduler looks at the time a user has been selected by the scheduler while it is consuming bits from the buffer B_l or T^l_{dwld}. Utility function for grading the users is the current $\left(T^l_{dwld}\right)$.

5) Mixed criteria based (MB): In this RRM the utility function for the user l is computed as the sum of the individual values T^l_{wait} and T^l_{dwld} divided by the value of T^l_b.

III. INVESTIGATION

A. Simulation Settings

To investigate the performance of using semantic-aware RRM in a wireless infrastructure we performed extensive simulations of an HSDPA network focusing on the downlink connection between one 3-sector BS, with 300m of cell radius, and the user devices requesting for the streaming of content stored in a multimedia server. Propagation model is the 3GPP model, where path loss is $L = 128.1 + 37.6\, log_{10}\,(R), R\ in\ Km$. We assumed that the backbone is lossless and the transmission delay from the media server to the BS is negligible. Maximum BS transmission power $\widehat{P_T} = 20W$, and maximum data rate of 14.4 Mbps. One user is allocated in a time slot of 0.25s. The basic system level assumptions used in the simulations are summarized in Table I. In our system, we played with users densities ranging between 5 and 25 users. In each case the user requests of multimedia content have been modelled with expected inter-arrival time equal to 1 minute. In all cases, users are supposed to pick a video of 57.76 Mbytes, representing a 1080p YouTube-like video of 5 minutes duration. In the simulation, users will watch up to 20 videos of the same size and resolution (homogeneous scenario), being a realization the completion of this number of videos by all users present in the system.

B. Performance Measures

From the operators' perspective, performance is evaluated considering the average Total Duration of Interruption (TDI) obtained by using each one of the proposed schedulers. The maximum and minimum length of interruptions represents the values of the longest and shortest interruptions experienced by user l during the playback. Finally, the average frequency of the interruptions will represent how often a user l can experience cuts during the playback.

IV. RESULTS

Figure 2 demonstrates the average TDI comparison of different RRM schemes for different number of users watching HD videos. With the implementation of the MB scheme there is reduction in the average TDI during the streaming session compared to the PF scheme. Reduction goes from 74%(with 5 users) to 8% (with 25 users). The other scheme that looks at the buffer capacity (BB) also guarantees a reduction in the TDI that goes between 53% (5 users) to 3% (25 users) regarding the PF. In contrast, schemes focused on the evaluation of online times (inactive/active) reduce the performance of the system increasing TDI.

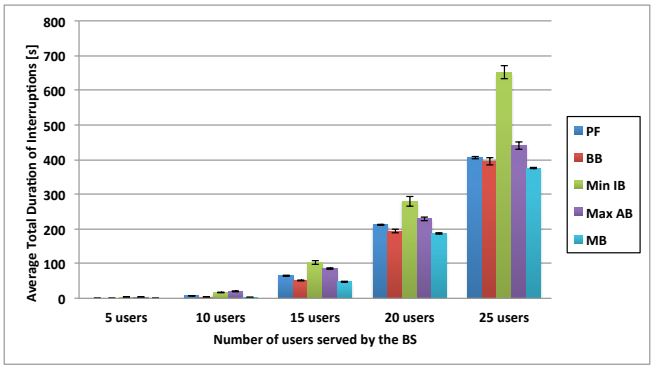

Figure 2. Average TDI by number of user for the different resource schedulers in a HD scenario. Error bars indicate 95% confidence intervals.

Figure 3 shows the length of the shortest interruption experienced by different number of users when the proposed RRM schemes are used. The figure reflects that schemes PF, MB and BB can guarantee that the shortest interruption in any case will be less than 1.3s. Schedulers that do not consider T_b in its allocation criteria will generate as shortest interruption duration values between 2s to 400s, which will generate a negative impact on the QoE perceived by users in a real scenario. Figure 4 shows the average maximum length of interruptions perceived by different number of users with different RRM schemes. In this case network reflects a better performance when PF, BB and MB are used as resource allocation schemes. Although PF shows the best performance when maximum length of interruptions is considered, observing the frequency of the interruptions in Figure 5, PF shows around 25% more interruptions during the playback than the best of the other considered RRM schemes. This recurrence in the number of interruptions will affect more the user's perceived quality, according to the results obtained in [14]. In summary, semantic-aware schemes that use buffer-related information will reduce the TDI through a reduction in the frequency and the length of interruptions.

 ISBN: 978-1-61208-313-1

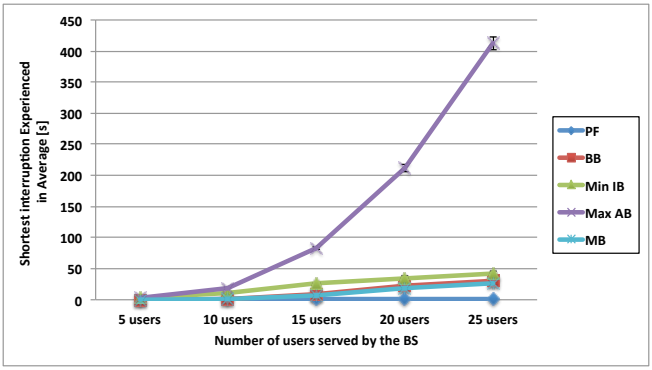

Figure 3. Average minimum length of interruptions by number of users for the different resource schedulers in a HD scenario.

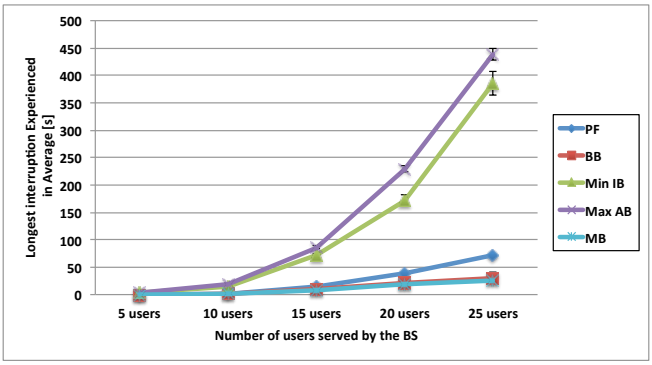

Figure 4. Average maximum length of interruptions by number of users for the different resource schedulers in a HD scenario.

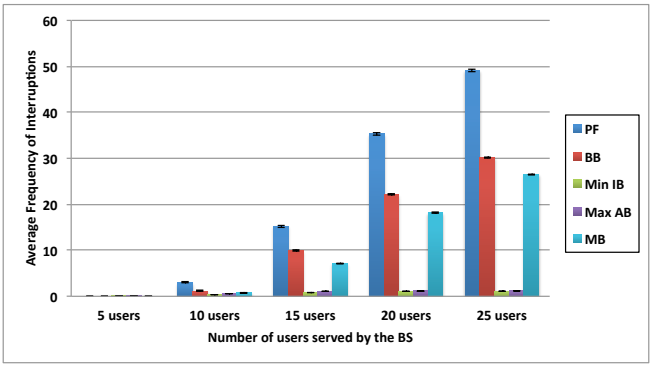

Figure 5. Average frequency of interruptions by number of users for the different resource scheduler in a HD scenario.

V. CONCLUSION

The concept of Semantic Radio Resource Management has been introduced in this paper as an alternative to improve the user service perception in video streaming services. The considered solution simply requires the introduction of software agents both at the network and terminal side, capable of monitoring applications behaviours. Different RRM strategies were simulated and results show that by using semantic-aware schemes evaluating user's buffer capacity, it is possible to improve the total duration of video stalling, and impact the length and frequency of the interruptions users can experience during the video playback. This indicates a potential of proposed solution to generate improvements in terms of the final QoE perceived by a user in comparison to the "classical" RRM. As future work, the extension of the proposed scheme considering more semantic elements to make resource allocation decisions is planned.

REFERENCES

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2011–2016." Cisco, White Paper, 2012.

[2] S. Thakolsri, S. Khan, E. Steinbach, and W. Kellerer, "Qoe-driven crosslayer optimization for high speed downlink packet access," Journal of Communications, Special Issue on Multimedia Communications,Networking and Applications, Vol. 4, No. 9., 2009, pp. 669–680.

[3] H. Yin and H. Liu, "An efficient multiuser loading algorithm for ofdm-based broadband wireless systems," in Global Telecommunications Conference, 2000. GLOBECOM '00. IEEE, vol. 1, 2000, pp. 103 –107.

[4] G. Aristomenopoulos, T. Kastrinogiannis, V. Kaldanis, G. Karantonis, and S. Papavassiliou, "A novel framework for dynamic utility-based qoe provisioning in wireless networks," GLOBECOM 2010, 2010 IEEE Global Telecommunications Conference, pp. 1–6, 2010.

[5] K. Piamrat, C. Viho, J.-M. Bonnin, and A. Ksentini, "Quality of experience measurements for video streaming over wireless networks," Information Technology: New Generations, 2009. ITNG '09. Sixth International Conference on, pp. 1184 – 1189, 2009.

[6] S. Rabiul Islam and M. Hossain, "A wireless video streaming system based on ofdma with multi-layer h.264 coding and adaptive radioresource allocation," in Image Information Processing (ICIIP), 2011 International Conference on, nov. 2011, pp. 1 –6.

[7] M. Shehada, S. Thakolsri, Z. Despotovic, and W. Kellerer, "Qoe-based cross-layer optimization for video delivery in long term evolution mobile networks," in Wireless Personal Multimedia Communications (WPMC), 2011 14th International Symposium on, oct. 2011, pp. 1 –5.

[8] P. Dutta, A. Seetharam, V. Arya, M. Chetlur, S. Kalyanaraman, and J. Kurose, "On managing quality of experience of multiple video streams in wireless networks," in INFOCOM, 2012 Proceedings IEEE, march 2012, pp. 1242 –1250.

[9] S.-P. Chuah, Z. Chen, and Y.-P. Tan, "Energy-efficient resource allocation and scheduling for multicast of scalable video over wireless networks," Multimedia, IEEE Transactions on, vol. 14, no. 4, aug. 2012, pp. 1324–1336.

[10] H. Adibah Mohd Ramli, K. Sandrasegaran, R. Basukala, R. Patachaianand, M. Xue, and C.-C. Lin, "Resource allocation technique for video streaming applications in the lte system," in Wireless and Optical Communications Conference (WOCC), 2010 19th Annual, may 2010, pp. 1 –5.

[11] P. Pahalawatta, T. Pappas, R. Berry, and A. Katsaggelos, "Content-aware resource allocation for scalable video transmission to multiple users over a wireless network," in Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, vol. 1, april 2007, pp. I–853 –I–856.

[12] X. Ji, J. Huang, M. Chiang, G. Lafruit, and F. Catthoor, "Scheduling and resource allocation for svc streaming over ofdm downlink systems," Circuits and Systems for Video Technology, IEEE Transactions on, vol. 19, no. 10, , oct. 2009, pp. 1549 –1555.

[13] C. Yang and S. Jordan, "Power and rate allocation for video conferencing in cellular networks," in Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on, sept. 2011, pp. 127 –134.

[14] Acision, "Seizing the opportunity in mobile broadband - a global perspective," Acision, Tech. Rep., 2011.

 ISBN: 978-1-61208-313-1