Cross-layer Content/Channel Aware Multi-User Scheduling for Downlink Wireless Video Streaming

Peter E. Omiyi [1], Maria G. Martini [2]
WMN Research Group, CISM Faculty
Kingston University, Kingston upon Thames, London, UK
[1] P.Omiyi@kingston.ac.uk
[2] M.Martini@kingston.ac.uk

***Abstract*—A new cross-layer content/channel-aware downlink scheduling algorithm for wireless video streaming is presented. A novel mathematical model has been developed for the performance analysis of the proposed scheduling algorithm. Using novel content-aware and standard performance metrics, the performance of the scheduling algorithm is evaluated and optimized. The results show that there is an inherent conflict between content-awareness for enhanced video quality and channel-awareness for high channel capacity. The scheduling algorithm can be tuned to maximize the throughput of the most significant video packets, while minimizing the capacity penalty due to quality/capacity trade-off. The results also suggest that the level of content-awareness required for optimum performance at the scheduler and the achieved capacity, are highly sensitive to the delay constraint. The lower the delay constraint, the higher the importance of content-awareness and the lower the capacity. Under the parameters of the investigation in this paper, the results show that the capacity is 3.7 times more for a delay constraint of 10ms than for a zero delay constraint.**

***Index Terms*—Scheduling, video streaming, SVC, content-aware, cross-layer.**

I. Introduction

Supporting multimedia applications and services over wireless networks is challenging due to constraints and heterogeneities such as limited bandwidth, limited battery power, random time-varying channel conditions, different protocols and standards, and quality of service (QoS) requirements. Cross-layer design methodologies have been proven beneficial for addressing these challenges and providing high-quality performance to the end user in wireless multimedia communications. In this framework, information collected at different layers of the protocol stack can be used for the optimization of the whole transmission system.

Typically cross-layer design is performed by jointly designing two layers of the protocol stack. For instance, in [1] and [2], scheduling plays an important role in providing quality of service (QoS) support to multimedia communications in various kinds of wireless networks, including cellular networks, mobile ad hoc networks, and wireless sensor networks. The authors propose a scheduling algorithm at the medium access control(MAC) layer for multiple connections with diverse QoS requirements,where each connection employs an adaptive modulation and coding scheme at the physical (PHY) layer over wireless fading channels. Each connection is assigned a priority, which is updated dynamically based on its channel conditions, and each time the connection with the highest priority is scheduled. This approach focuses on the joint design and operation of the MAC and PHY layers.

The work in [3] includes in the analysis MAC-PHY and APP layers, presenting as an example a MAC-application layer optimization strategy for video transmission over 802.11a wireless LANs based on classification.

This paper applies a cross-layer approach to the problem of scheduling video streams to multiple users on the downlink of a cell or sector of a wireless mobile network, where scalable video coding (SVC) is employed[1]. Traditionally, there are two main classifications of scheduling algorithms: channel aware / channel unaware schedulers and content-aware/content-unaware schedulers. Channel-unaware schedulers make no use of channel state conditions such as power level and channel error and loss rates. These basically focus on fulfilling delay and throughput constraints. These include Round-Robin, weighted fair queuing (WFQ) and priority based algorithms. Such algorithms assume perfect channel condition, no loss and unlimited power source. However, due to the nature of wireless medium and the user mobility, these assumptions are not valid.

The BS downlink scheduler could use instead channel information, such as in [4–7], including the Carrier to Interference and Noise Ratio (CINR), which is reported back from the mobile receiver. Most of channel-aware algorithms assume that channel conditions do not change within the frame period. It is also assumed that the channel information is known at both the transmitter and the receiver. In general, schedulers favor the users with better channel quality to exploit the multiuser diversity and channel fading.

The above strategies are, traditionally, non-content aware and the quality of service of the received video is measured in generic terms of packet delay, packet loss rate or data rate. Video quality, however, is not a simple function of the data rate, delay or data loss but rather is affected differently by different segments of the video stream. This is emphasized in SVC where a video stream comprises of multiple quality layers. In multi-user video transmission, this introduces a type of multi-user content diversity that can be exploited by content-aware scheduling policies in optimizing the utilization of the network resource. Examples of content-aware methods are found in [8–16].

In [8, 9], a concept of incrementally additive distortion among video packets is used to determine the importance of video packets for each user. Essentially, the increase in distortion due to the loss of a video packet is a function of all other video packets that are dependent on it and cannot be decoded if it was not sent or could not be decoded. This information is used to drop video packets in the event of congestion over the wireless interface, beginning with the

[1]This work was partially supported by the European Commission under FP7 project INFSO-ICT-214625 OPTIMIX.

least important video packet. This buffer management strategy is combined with various scheduling approaches that set the priorities across users in order to either maximize throughput, ensure proportional fairness or minimize late packets. Scheduling across users, however, does not explicitly exploit the relative importance of video packets.

In [10, 11], the subflow concept is introduced in which a video flow (bitstream) is divided into several subflows based on their delay constraints as well as based on the relative priority in terms of the overall distortion of the decoded video. This is combined with a prioritized scheduling function of the 802.11e WLAN MAC. One drawback of this approach is its limitation to the 802.11e MAC which allows a limited number of priority classes. This results in limited subflow differentiation of the video stream and thus limited gains from the multi-user content diversity. Furthermore, multi-user channel diversity is not exploited.

In [13], content-aware packet reordering per user and multi-user content-aware priority based scheduling are also proposed. User priorities are determined in each time-slot based on maximizing a multi-dimensional utility function, which aims to minimize the distortion of users and maximize the (error-free) per-user data rate. Content-based long-term fairness is also included in the utility function. However, the optimum multi-user transmission profile is intractable and an approximate sub-optimal solution is used by the scheduler. A similar approach is considered in [14].

In this paper, a multi-user, content-aware priority based scheduling algorithm is proposed, where packet priorities are selected in order to jointly exploit multi-user content and channel diversity. This work is, therefore, most similar to [13]. However, in the proposed algorithm, the priority of each packet is determined as a polynomial function of the quality of the channel(s) available to the user in each time-slot and the content-dependent importance of the packet. The proposed approach avoids the multi-dimensional optimization at every time-slot inherent in [13], in which the computational complexity increases with the number of users and available channels. Instead, a single prioritization parameter is used, which determines the trade-off between content-awareness for enhanced video quality and channel-awareness for high channel capacity.

Another contribution of the work presented in this paper consists of new performance metrics for content-aware scheduling. The proposed performance metrics, unlike the traditional objective video quality metrics [15, 17–22], provide a direct measure of the effectiveness of the scheduling strategy in handling the different segments of the video stream according to their relative importance. Finally, a novel mathematical model and analysis of the proposed scheduling algorithm is presented. This is used to obtain results evaluating the performance of the proposed scheduler.

The rest of the paper is organized as follows. Section II describes the proposed scheduler, while Section III, presents the key performance metrics and the derivation of their analytical expressions. Numerical results and conclusions are presented in Section IV and Section V, respectively.

II. Joint Multi-user Content-Aware and Channel-Aware Prioritized Scheduler

Figure 1 shows a schematic of the joint content/channel aware scheduling approach. Downlink (DL) scheduling decisions are made periodically every DL frame, with consecutive DL frames separated by uplink (UL) frames in which UL data, CSI and hybrid ARQ feedback are sent. There is a fixed number B of multidimensional resource slots per DL frame. A GOP may span several DL frames. A resource slot comprises of a fixed number of symbols (time) and a fixed number of resource units in one or more resource dimensions, such as a fixed number of sub-carriers (frequency), codes and spatial layers. The set of resource slots per DL frame is denoted as $\mathcal{B}$.

The number of bits that can be sent by a user on an allocated resource slot depends on the specific channel conditions for the user on that resource slot. The group of data bits sent on a resource slot by a user constitute a packet. The channel quality $q_{i,c}$ of a resource slot c for user i is defined as the ratio of the spectral efficiency of the transmission of user i on the resource slot to the maximum spectral efficiency possible on any resource slot. The number of bits per transmitted packet is a product of the channel quality, the maximum spectral efficiency and the duration of the resource slot. For ease of notation, $q_{i,c}$ is denoted as q in most of the paper.

Each packet has a relative importance or significance, denoted as v, which is a function of factors that impact on the perceived video quality at the end user, such as the total number of bits that will be lost or cannot be decoded at the video decoder as a result of its loss. A target maximum delivery delay t is associated with each packet, in order to enforce a minimum bit rate requirement. The latter is imposed on the video stream on a per-GOP basis to ensure that all the most significant bits in each GOP are received on time to be used by the decoder. The decoder uses these most significant bits together with error concealment to provide acceptable video quality, without interrupting video playback.

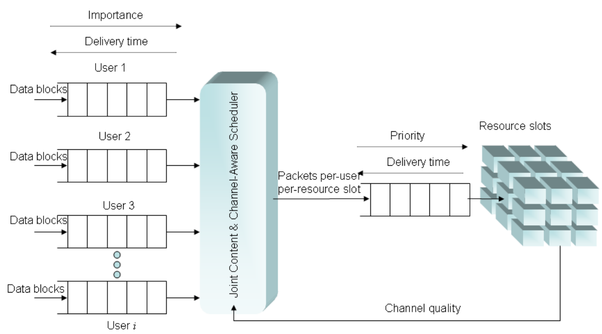

Fig. 1. Schematic of joint content/channel aware scheduling

The resource slot allocation process in each DL frame is iterative with a maximum of B iterations, with one resource slot allocated in each iteration. At any iteration n, the number of unallocated resource slots is $B - n$ and comprises the set $\mathcal{B}_n \in \mathcal{B}$, where n is a positive integer and $\mathcal{B}$ is the set of all resource slots. At iteration n the resource slot allocation is as follows.

1) For each user, bits are first grouped in order of their delivery times $\tau + t$ at the decoder, where τ is the time of arrival of the bit at the scheduler, with the bits with the earliest deadline at the head of the buffer.
2) Among all bits with the same deadline for each user, bits are arranged in order of relative importance v, with the most important blocks at the head of the buffer.

3) The best quality resource slot for each user is selected from the set $\mathcal{B}_n$, where the selected resource slots are not necessarily unique to each user. The best quality resource slot of a user is the slot from the set $\mathcal{B}_n$ that maximizes the number of bits that can be transmitted without error and is a function of the channel quality q.
4) From this set, the scheduler selects the set of users with the lowest head-of-line (HOL) deadline, and from these selects the set of users with the highest HOL priority $p = \alpha v + (1-\alpha)q$. This means that from the set of users with the same HOL importance the scheduler selects the set of users that have the best channel quality. Similarly, from the set of users with the same channel quality q the scheduler selects the set of users with the highest HOL importance.
5) The parameter α in the priority equation is used to tune the system trade-off between exploiting channel diversity to maximize channel throughput and maximizing the per user video quality.
6) From this set, the scheduler selects the user i whose best resource slot c is the best resource slot for the fewest number of other users.
7) This slot c is assigned to user i, and is removed from the set $\mathcal{B}_n$.

The above process is repeated in the next iteration $n = n+1$, if there are still unassigned blocks and available resource slots $(B - n > 0)$.

III. Mathematical Model

Two novel and key performance metrics proposed in this paper for evaluating the performance of content-aware scheduling are defined as follows. The first is the normalized significance throughput $\breve{S}_{\mathrm{v}}(v, \alpha)$, which is defined as the throughput of packets with significance greater than v normalized by the offered load of packets with significance greater than v. The second is the probability that a packet of significance greater than v is delayed beyond its delay constraint $\mathcal{P}_{\mathrm{v}}(v, \alpha)$.

Another metric used is the normalized throughput per resource slot $S_{\mathrm{tot}}(\alpha)$ for a given value of α, which is the throughput per resource slot normalized by the maximum number of bits that can be sent in a resource slot. The normalized throughput $S_{\mathrm{tot}}(\alpha)$ is maximized as α tends to zero and packets are prioritized according to channel quality rather than source significance. The channel utilization $\eta(\alpha)$ is a measure of the deviation from the maximum throughput as α increases from zero. The main objective of the subsequent mathematical analysis is to derive analytic expressions for $\breve{S}_{\mathrm{v}}(v)$, S_{tot} and $\mathcal{P}_{\mathrm{v}}(v, \alpha)$.

In the real system, only bits that are to be transmitted are grouped into packets in the real system, with the number of bits per packet dependent on the channel conditions of the resource slot on which the packet is sent. Therefore, the number of data bits is known before the allocation process, but the number of packets and the number of bits per packet are only known after the allocation process. The traffic model assumed the analysis represents the view of the system after the allocation process. Therefore, the packets considered in the analysis include both real packets that are assigned to resource slots and transmitted, and virtual packets that represent an arbitrary grouping of the bits that are not transmitted but remain queued by the scheduler. The distribution of the number of blocks in virtual packets is assumed without loss of generality to be the same as the distribution in real packets. The number of packets per DL frame from multiple users is assumed to follow an arbitrary distribution with mean λ_{S}.

The scheduling priority is determined by $p = \vartheta + \omega$, where $\vartheta = \alpha v$, $\omega = (1-\alpha)q$ and $0 \leq \alpha \leq 1$. The priority of a randomly selected packet has pdf and cumulative density function of $f_{\mathrm{p}}(p)$ and $F_{\mathrm{p}}(p)$, respectively, where the pdf $f_{\mathrm{p}}(p)$ is given by

$$f_{\mathrm{p}}(p) = \int_{-\infty}^{\infty} f_{\vartheta}(p-\omega) f_{\omega}(\omega) d\omega \tag{1}$$

The probability that a packet has a priority higher than p is $1 - F_{\mathrm{p}}(p)$.

It is assumed that the channel quality q and significance v of a randomly selected packet follow a uniform distribution in the interval $(0, 1]$. Therefore, ϑ and ω follow uniform distributions in the intervals $(0, \alpha]$ and $(0, 1-\alpha]$, respectively. These assumptions do not impact on the generality of the subsequent analysis and any other distribution can be assumed.

In order to derive the pdf of $f_{\mathrm{p}}(p)$, define random variables x and y that are uniformly distributed over $(0, \beta]$ and $(0, 1-\beta]$, respectively, where $0 < \beta \leq 1-\beta \leq 1$ $(\beta \leq 0.5)$. This implies that $x = \vartheta$ and $y = \omega$, when $\beta = \alpha$, while the converse is true ($y = \vartheta$ and $x = \omega$) when $\beta = 1-\alpha$. Let $f_{\mathrm{U}}(z; u)$ denote the pdf of a random variable z that is uniformly distributed over the range $(0, u]$, then the pdfs of x and y are denoted as $f_{\mathrm{U}}(x; \beta)$ and $f_{\mathrm{U}}(y; 1-\beta)$, respectively. The pdf $f_{\mathrm{p}}(p)$ is computed as follows.

$$f_{\mathrm{p}}(p) = \int_{-\infty}^{\infty} f_{\mathrm{U}}(p-y; \beta) f_{\mathrm{U}}(y; 1-\beta) dy. \tag{2}$$

The pdf $f_{\mathrm{U}}(y; 1-\beta) = 1$ in the range $0 \leq y \leq 1-\beta$, and is zero, otherwise. Therefore,

$$f_{\mathrm{p}}(p) = \int_{0}^{1-\beta} f_{\mathrm{U}}(p-y; \beta) dy \tag{3}$$

For $0 \leq p \leq \beta$

$$f_{\mathrm{p}}(p) = \frac{1}{\beta(1-\beta)} \int_{0}^{p} 1 dy = \frac{p}{\beta(1-\beta)}. \tag{4}$$

For $\beta \leq p \leq 1-\beta$

$$f_{\mathrm{p}}(p) = \frac{1}{\beta(1-\beta)} \int_{p-\beta}^{p} 1 dy = \frac{1}{1-\beta}. \tag{5}$$

For $1-\beta \leq p \leq 1$

$$f_{\mathrm{p}}(p) = \frac{1}{\beta(1-\beta)} \int_{p-\beta}^{1-\beta} 1 dy = \frac{1-p}{\beta(1-\beta)}. \tag{6}$$

The final equation for $f_{\mathrm{p}}(p)$ can be, concisely, expressed as

$$f_{\mathrm{p}}(p) = \begin{cases} \frac{p}{\beta(1-\beta)} & \text{for } 0 \leq p \leq \beta \\ \frac{1}{1-\beta} & \text{for } \beta \leq p \leq 1-\beta \\ \frac{1-p}{\beta(1-\beta)} & \text{for } 1-\beta \leq p \leq 1 \\ 0 & \text{Otherwise,} \end{cases} \tag{7}$$

where $\beta \leq 0.5$. Therefore, for $\alpha \leq 0.5$, the pdf $f_{\mathrm{p}}(p)$ is given by substituting $\beta = \alpha$ in (7), while for $\alpha > 0.5$ $(1-\alpha \leq 0.5)$, the pdf $f_{\mathrm{p}}(p)$ is is given by substituting $\beta = 1-\alpha$.

The cdf $F_{\mathrm{p}}(p)$ is computed as follows. For $0 \leq p \leq \beta$, the cdf is the area of the triangle with base length p and height $f_{\mathrm{p}}(p)$

$$F_{\mathrm{p}}(p) = \frac{p^2}{2\beta(1-\beta)}. \tag{8}$$

For $\beta \leq p \leq 1-\beta$, the cdf is sum of the area of the triangle with base length and height of β and $\frac{1}{1-\beta}$, respectively, and the area of the square with base and height of $p-\beta$ and $\frac{1}{1-\beta}$, respectively.

$$\begin{aligned} F_{\mathrm{p}}(p) &= \frac{\beta}{2(1-\beta)} + \frac{p-\beta}{1-\beta} \\ &= \frac{\beta}{2(1-\beta)} + \frac{2p-2\beta}{2(1-\beta)} = \frac{2p-\beta}{2(1-\beta)}. \end{aligned} \tag{9}$$

For $1-\beta \leq p \leq 1$, the cdf is unity less the area of the triangle with base length $1-p$ and height $f_{\mathrm{p}}(p)$

$$F_{\mathrm{p}}(p) = 1 - \frac{(1-p)^2}{2\beta(1-\beta)}. \tag{10}$$

Concisely, the cdf $F_{\mathrm{p}}(p)$ is

$$F_{\mathrm{p}}(p) = \begin{cases} \frac{p^2}{2\beta(1-\beta)} & \text{for } 0 \leq p \leq \beta \\ \frac{2p-\beta}{2(1-\beta)} & \text{for } \beta \leq p \leq 1-\beta \\ 1 - \frac{(1-p)^2}{2\beta(1-\beta)} & \text{for } 1-\beta \leq p \leq 1 \\ 0 & \text{Otherwise,} \end{cases} \tag{11}$$

where $\beta = \min(\alpha, 1-\alpha)$.

After the allocation process packets are arranged in sequence according to their priorities, such that packets with priority greater than p form a 'run' of a random number of packets. The mean number of packets with priority greater than p is:

$$E[l(p)] = \lambda_{\mathrm{S}}(1 - F_{\mathrm{p}}(p)). \tag{12}$$

Assuming that the length of the run is geometrically distributed, then the parameter π_p of the geometric distribution is given by

$$\pi_p = \frac{1}{E[l(p)] + 1}. \tag{13}$$

If the number of packets l in the run with priority higher than p is less than the available number of resource slots B, then the number of packets with priority higher than p that are transmitted is equal to l. Otherwise, the number of packets with priority higher than p that are transmitted is equal to B. Therefore, the throughput of packets with priority greater than p is

$$\begin{aligned} S_{\mathrm{p}}(p) &= \pi_p(1-\pi_p)\sum_{l=1}^{B} l(1-\pi_p)^{l-1} + B\pi_p \sum_{l=B+1}^{\infty}(1-\pi_p)^l \\ &= \frac{1-\pi_p}{\pi_p}.[1-(1-\pi_p)^B(\pi_p B+1)] + B.(1-\pi_p)^{B+1} \\ &= E[l(p)].[1-(1-\pi_p)^B(\pi_p B+1)] + B.(1-\pi_p)^{B+1}. \end{aligned} \tag{14}$$

The throughput of packets with significance greater than v is computed as follows. Given that $p = \alpha v + (1-\alpha)q$ and that the maximum value of q is unity, then packets with priorities $p > \alpha v + (1-\alpha)$, always have a significance greater than v. Furthermore, packets with priorities $p \leq \alpha v$ always have a significance less than or equal to v, while packets with priorities in the range $\alpha v < p \leq \alpha v + (1-\alpha)$ have a significance that occasionally exceeds v. The proportion of packets with significance greater than v in the range $\alpha v < p \leq \alpha v + (1-\alpha)$ is

$$P_{\mathrm{v}}(v) = \frac{\lambda_{\mathrm{v}}(v) - \lambda_{\mathrm{p}}(\alpha v + (1-\alpha))}{\lambda_{\mathrm{p}}(\alpha v) - \lambda_{\mathrm{p}}(\alpha v + (1-\alpha))}. \tag{15}$$

The throughput of packets with significance greater than v is the sum of throughput of packets priorities greater than $\alpha v + (1-\alpha)$ and fraction of packets with priorities in the range $\alpha v < p \leq \alpha v + (1-\alpha)$ that have a significance that occasionally exceeds v

$$\begin{aligned} S_{\mathrm{v}}(v,\alpha) &= S_{\mathrm{p}}(\alpha v + (1-\alpha)) \\ &\quad + P_{\mathrm{v}}(v)[S_{\mathrm{p}}(\alpha v) - S_{\mathrm{p}}(\alpha v + (1-\alpha))]. \end{aligned} \tag{16}$$

The offered load of packets with significance greater than v is

$$\lambda(v) = \lambda_{\mathrm{S}}(1 - F_{\mathrm{v}}(v)) = \lambda_{\mathrm{S}}(1-v). \tag{17}$$

A normalized significance throughput is defined as the throughput of packets with significance greater than v normalized by the offered load of packets with significance greater than v and is:

$$\breve{S}_{\mathrm{v}}(v,\alpha) = S_{\mathrm{v}}(v,\alpha)/\lambda_{\mathrm{v}}(v). \tag{18}$$

The normalized significance throughput $\breve{S}_{\mathrm{v}}(v,\alpha)$ is effectively the probability that a packet with significance greater than v in the queue in any given DL frame is transmitted. Therefore, $1 - \breve{S}_{\mathrm{v}}(v,\alpha)$ is the probability that the packet is not transmitted but queued at least until the next DL frame. Assuming that the this probability is approximately independent in consecutive DL frames, then the probability that a packet of significance greater than v is delayed beyond its delay constraint t is computed as follows.

$$\begin{aligned} \mathcal{P}_{\mathrm{v}}(v,\alpha) &= 1 - \sum_{\tau=0}^{t}[1-\breve{S}_{\mathrm{v}}(v,\alpha)]^{\tau}\breve{S}_{\mathrm{v}}(v,\alpha) \\ &= [1-\breve{S}_{\mathrm{v}}(v,\alpha)]^{t+1}. \end{aligned} \tag{19}$$

The throughput of packets with channel quality greater than q is the sum of throughput of packets priorities greater than $\alpha + (1-\alpha)q$ and fraction of packets with priorities in the range $(1-\alpha)q < p \leq \alpha + (1-\alpha)q$ that have a quality that occasionally exceeds q. It is derived in a similar manner as $S_{\mathrm{v}}(v,\alpha)$ to give

$$\begin{aligned} S_{\mathrm{q}}(q,\alpha) &= S_{\mathrm{p}}(\alpha + (1-\alpha)q) \\ &\quad + P_{\mathrm{q}}(q)[S_{\mathrm{p}}([1-\alpha]q) - S_{\mathrm{p}}(\alpha + (1-\alpha)q)], \end{aligned} \tag{20}$$

where

$$P_{\mathrm{q}}(q) = \frac{\lambda_{\mathrm{q}}(q) - \lambda_{\mathrm{p}}(\alpha + (1-\alpha)q)}{\lambda_{\mathrm{p}}((1-\alpha)q) - \lambda_{\mathrm{p}}(\alpha + (1-\alpha)q)}.$$

The normalized throughput per resource slot for a given value of α is the throughput in bits normalized by the maximum number of bits that can be sent in a resource slot is

$$S_{\mathrm{tot}}(\alpha) = \frac{1}{B}\int_0^1 \frac{\mathrm{d}S_{\mathrm{q}}(q,\alpha)}{\mathrm{d}q} q\,\mathrm{d}q. \tag{21}$$

IV. Numerical Results

The results presented in this section assume a DL frame duration of 1 ms. The maximum number of resource slots B per DL frame is 2048, with each resource slot having a bandwidth of 180 KHz. These parameters are similar to those proposed for LTE. However, we assume a much larger number of resource slots (blocks) per DL frame than in an LTE sub-frame, with only 100 in the latter. This can be achieved, for example in LTE advanced by using carrier aggregation to increase the total available bandwidth (and number of sub-carriers) and MIMO to increase the number of possible simultaneous transmissions on the same sub-carriers. Furthermore, we assume a maximum spectral efficiency of 2 bits/s/Hz in each resource slot, which maps to a maximum of 360bits per resource slot.

Figure 2 shows a contour plot of the normalized throughput $S_{\text{tot}}(\alpha)$ versus normalized offered load λ_{tot} and prioritization coefficient α. Figure 2 shows that the normalized throughput

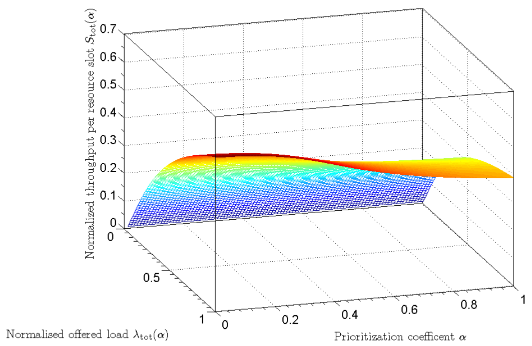

Fig. 2. Normalized throughput per resource slot versus normalized offered load and prioritization coefficient.

$S_{\text{tot}}(\alpha)$ increases with normalized offered load λ_{tot} with the rate of increase declining at high values of λ_{tot}. Furthermore, as the prioritization coefficient α is decreased $S_{\text{tot}}(\alpha)$ increases. A noticeable increase in $S_{\text{tot}}(\alpha)$ is observed as α crosses the 0.5 mark at high values of λ_{tot}, which is the point at which the scheduler begins to favor channel quality over source significance.

Figure 3, Figure 4 and Figure 5 show plots of the normalized significance throughput $\breve{S}_{\text{v}}(v,\alpha)$ versus significance v for different values of α and offered load λ_{tot}. It is observed in all three figures that for α tending to zero, where prioritization is based solely on channel quality, $\breve{S}_{\text{v}}(v,\alpha)$ is constant for any given offered-load. As α increases, the content-awareness at the scheduler rises, and $\breve{S}_{\text{v}}(v,\alpha)$ is observed to increase as a function of significance v. The higher the value of α the higher the value of $\breve{S}_{\text{v}}(v,\alpha)$ for any given offered-load, and the faster the rate of increase of $\breve{S}_{\text{v}}(v,\alpha)$ with v. Furthermore, the lower the offered load the higher the normalized significance throughput $\breve{S}_{\text{v}}(v,\alpha)$ for any given values of α and v.

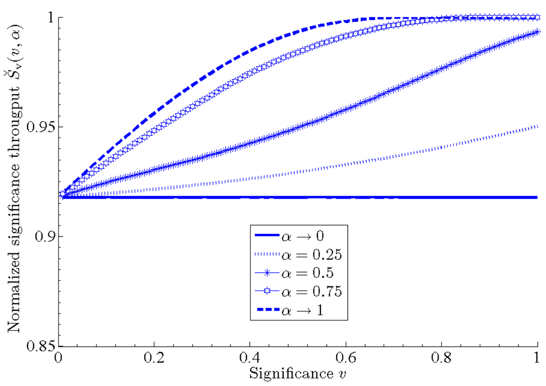

Fig. 3. Normalized significance throughput versus significance for a normalized offered load=0.2.

Figure 6, Figure 7 and Figure 8 show plots of the normalized significance throughput $\breve{S}_{\text{v}}(v,\alpha)$ versus offered load λ_{tot} for different values of α and significance v. It is observed in all

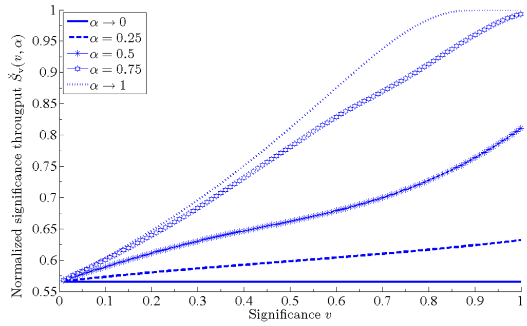

Fig. 4. Normalized significance throughput versus significance for a normalized offered load=0.6.

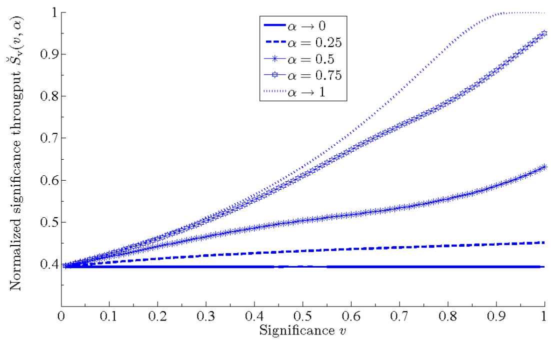

Fig. 5. Normalized significance throughput versus significance for a normalized offered load=1.

three figures that the smaller the value of α the faster the rate of decline of $\breve{S}_{\text{v}}(v,\alpha)$ from unity with increasing offered load. Furthermore, the higher the significance the higher the normalized significance throughput $\breve{S}_{\text{v}}(v,\alpha)$ for any given values of α and offered load.

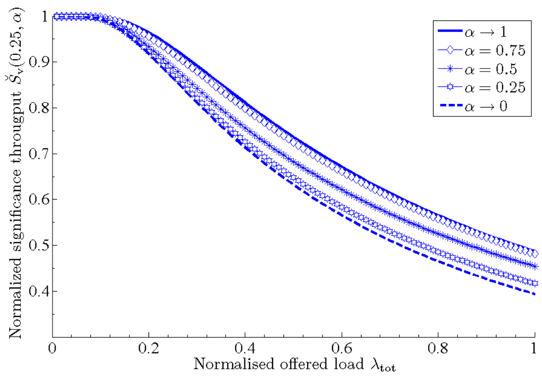

Fig. 6. Normalized significance throughput for a significance greater than 0.25 versus normalized offered load.

In order to support the video QoS, it is required that packets with a significance greater than a target value v are received at the decoder before being out-of-date. This implies that the normalized significance throughput $\breve{S}_{\text{v}}(v,\alpha)$ of packets with significance greater than v exceeds a target value θ_{v}. In order to maximize the number of users, it is required to maximize the normalized throughput $S_{\text{tot}}(\alpha)$. Maximizing $S_{\text{tot}}(\alpha)$ requires minimizing the prioritization coefficient α and maximizing the normalized offered load λ_{tot}, which minimizes $\breve{S}_{\text{v}}(v,\alpha)$. Therefore, there is a trade off between maximizing $S_{\text{tot}}(\alpha)$ and ensuring that $\breve{S}_{\text{v}}(v,\alpha) > \theta_{\text{v}}$. The optimal values of α and λ_{tot} are the values that deliver the highest value of $S_{\text{tot}}(\alpha)$

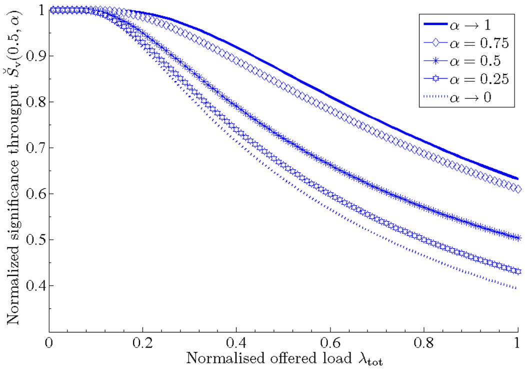

Fig. 7. Normalized significance throughput for a significance greater than 0.5 versus normalized offered load.

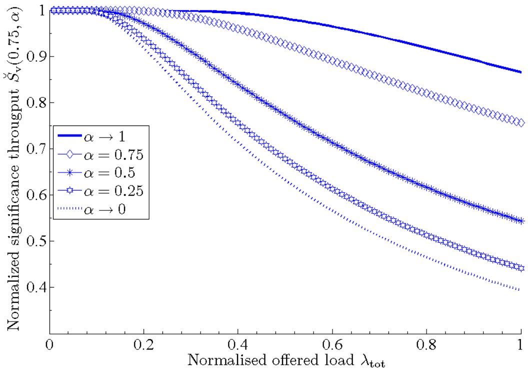

Fig. 8. Normalized significance throughput for a significance greater than 0.75 versus normalized offered load.

that can be supported, while ensuring that $\breve{S}_{\rm v}(v,\alpha) > \theta_{\rm v}$.

Consider packets with significance greater than $v = 0.5$ and a target maximum packet delay of 10 DL frames. In order to ensure that the probability of exceeding this delay is less than 10^{-3}, a value of $\theta_{\rm v} = 0.4663$ is obtained from (19). Using (18) and (21), the optimum values of α and $\lambda_{\rm tot}$ are determined by a direct search as 0.19 and 0.86, respectively, with $S_{\rm tot}(\alpha) = 0.5$ (where 0.5 is the maximum value). Assuming a target maximum packet delay of zero (DL frames), a value $\theta_{\rm v} = 0.999$ is required, which gives optimum values of α and $\lambda_{\rm tot}$ as 0.91 and 0.14, respectively, with $S_{\rm tot}(\alpha) = 0.1348$.

V. Conclusions

A new joint content/channel-aware multiple user DL scheduling algorithm for wireless video streaming has been presented. Using novel content-aware and standard performance metrics, the performance of the scheduling algorithm is evaluated and optimized using a new mathematical model. The delay constraint is observed to impact on the level of content-awareness required at the scheduler for optimum performance, and on the achieved capacity. The results show that the capacity is 3.7 times more for a delay constraint of 10ms than for a zero delay constraint. Given that the lower delay constraint also implies a higher video quality, the capacity/delay-constraint trade-off impacts on the overall quality/capacity trade-off. Therefore, the task of achieving the desired quality/capacity trade-off requires first setting a delay constraint or minimum data rate that is consistent with the desired quality. Under the condition of this delay constraint, the optimum prioritization coefficient for the scheduler is selected that maximizes the throughput of the most significant packets, while minimizing the capacity penalty.

References

[1] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer scheduling with prescribed QoS guarantee in adaptive wireless networks," *IEEE Journal on Areas in Communications*, vol. 23, no. 5, pp. 1056–1066, 2005.

[2] Q. Liu, X. Wang, and G. B. Giannakis, "A cross-layer scheduling algorithm with QoS support in wireless networks," *IEEE Trans. on Veh. Technology*, vol. 55, no. 3, pp. 839–847, 2006.

[3] M. van der Schaar, D. S. Turaga, and R. Wong, "Classification-based system for cross-layer optimized wireless video transmission," *IEEE Trans. on Multimedia*, vol. 8, no. 5, 2006.

[4] Y. Ofuji, S. Abeta, and M. Sawahashi, "Unified packet scheduling method considering delay requirement in forward link broadband wireless access," in *Proc. Vehicular Technology Fall-Conf. 2003*, 2003.

[5] P. Falconio and P. Dini, "Design and performance evaluation of packet scheduler algorithms for video traffic in the high speed downlink packet access," in *Proc. IEEE PIMRC 2004*, Sep. 2004.

[6] D. Kim, B. Ryu, and C. Kang, "Packet scheduling scheme for real time video traffic in WCDMA downlink," in *Proc. 7th CDMA International Conf.*, Seoul, Korea, October 29 - November 1, 2002.

[7] A. I. Perez-Neira, P. Henarejos, V. Tralli, and M. A. Lagunas, "A low-complexity space-frequency multiuser scheduling algorithm," in *Proc. International ITG Workshop on Smart Antennas*, Berlin, Germany, 2009.

[8] G. Liebl, T. Stockhammer, C. Buchner, and A. Klein, "Radio link buffer management and scheduling for video streaming over wireless shared channels," in *Proc. Packet Video Workshop*, Irvine, CA, USA, December 2004.

[9] G. Liebl, M. Kalman, and B. Girod, "Deadline-aware scheduling for wireless video streaming," in *Proc. IEEE ICME 2005*, Amsterdam, Netherlands, July 2005.

[10] M. van der Schaar, Y. Andreopoulos, and Z. Hu, "Optimized scalable video streaming over IEEE 802.11a/e HCCA wireless networks under delay constraints," *IEEE Trans. on Mobile Computing*, vol. 5, no. 6, pp. 755–768, 2006.

[11] J.-L. Hsu and M. van der Schaar, "Cross layer design and analysis of multiuser wireless video streaming over 802.11e edca," *IEEE signal processing letters*, vol. 16, no. 4, pp. 268–271, 2009.

[12] H. Zhang, Y. Zheng, M. A. Khojastepour, and S. Rangarajan, "Scalable video streaming over fading wireless channels," in *Proc. IEEE WCNC 2009*, Budapest, Hungary, Apr. 2009.

[13] P. V. Pahalawatta, R. Berry, T. N. Pappas, and A. K. Katsaggelos, "Content-aware resource allocation and packet scheduling for video transmission over wireless networks," *IEEE Journal on Sel. Areas on Commun.*, vol. 25, no. 4, pp. 749–759, 2007.

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

[15] J. Huang, Z. Li, M. Chiang, and A. K. Katsaggelos, "Joint source adaptation and resource allocation for multi-user wireless video streaming," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 18, no. 5, pp. 582–595, 2008.

[16] M. G. Martini and V. Tralli, "Video quality based adaptive wireless video streaming to multiple users," in *Proc. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, Las Vegas, NV, Mar.–Apr. 2008.

[17] J. Chakareski and P. Frossard, "Rate-distortion optimised distributed packet scheduling of multiple video streams over shared communication resources," *IEEE Trans. on Multimedia*, vol. 8, no. 2, pp. 207–218, 2006.

[18] D. Jurca and P. Frossard, "Video packet selection and scheduling for multipath streaming," *IEEE Trans. on Multimedia*, vol. 9, no. 3, pp. 629–641, 2007.

[19] J. A. Y. Li, A. Markopoulou and N. Bambos, "Content-aware playout and packet scheduling for video streaming over wireless links," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 885–895, 2008.

[20] G. Liebl, T. Schierl, T. Wiegand, and T. Stockhammer, "Advanced wireless multiuser video streaming using the scalable video coding extensions of H.264/MPEG4-AVC," in *Proc. IEEE International Conference on Multimedia and Expo, 2006*.

[21] M. G. Martini, M. Mazzotti, C. Lamy-Bergot, J. Huusko, and P. Amon, "Content adaptive network aware joint optimization of wireless video transmission," *IEEE Communications Magazine*, vol. 45, no. 3, pp. 1–10, 2007.

[22] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal processing: Image communication*, vol. 29, no. 1, Jan. 2004.