Multi-User Video Streaming Over Multi-Hop Wireless Networks: A Distributed, Cross-Layer Approach Based on Priority Queuing

Hsien-Po Shiang and Mihaela van der Schaar

Abstract*— Emerging multi-hop wireless networks provide a low-cost and flexible infrastructure that can be simultaneously utilized by multiple users for a variety of applications, including delay-sensitive multimedia transmission. However, this wireless infrastructure is often unreliable and provides dynamically varying resources with only limited Quality of Service (QoS) support for multimedia applications. To cope with the time-varying QoS, existing algorithms often rely on non-scalable, flow-based optimizations to allocate the available network resources (paths and transmission opportunities) across the various multimedia users. Moreover, previous research seldom optimizes jointly the dynamic routing with the adaptation and protection techniques available at the medium access control (MAC) or physical (PHY) layers. In this paper, we propose a distributed packet-based cross-layer algorithm to maximize the decoded video quality of multiple users engaged in simultaneous real-time streaming sessions over the same multi-hop wireless network. Our algorithm explicitly considers packet-based distortion impact and delay constraints in assigning priorities to the various packets and then relies on priority queuing to drive the optimization of the various users' transmission strategies *across the protocol layers* as well as *across the multi-hop network*. The proposed solution is enabled by the scalable coding of the video content (i.e. users can transmit and consume video at different quality levels) and the cross-layer optimization strategies, which allow priority-based adaptation to varying channel conditions and available resources. The cross-layer strategies – application layer packet scheduling, the policy for choosing the relays, the MAC retransmission strategies, the PHY modulation and coding schemes – are optimized per packet, at each node, in a distributed manner. The main component of the proposed solution is a low-complexity, distributed, and dynamic routing algorithm, which relies on prioritized queuing to select the path and time reservation for the various packets, while explicitly considering instantaneous channel conditions, queuing delays and the resulting interference. Our results demonstrate the merits and need for end-to-end cross-layer optimization in order to provide an efficient solution for real-time video transmission using existing protocols and infrastructures. Importantly, our proposed delay-driven, packet-based transmission is superior in terms of both network scalability and video quality performance to previous flow-based solutions that statically allocate resources based on predetermined paths and rate requirements. In addition, the results provide important insights that can guide the design of network infrastructures and streaming protocols for video streaming.

***Index Terms*— Cross-layer optimized video streaming, distributed video transmission, dynamic routing, multi-hop wireless networks, priority queuing.**

Manuscript received May 15, 2006; revised Dec. 18, 2006. This work was supported by grants from NSF Career Award, Intel IT Research, UC Micro and HP Research.

The authors are with the Department of Electrical Engineering (EE), University of California Los Angeles (UCLA), Los Angeles, CA (e-mail: {hpshiang, mihaela}@ee.ucla.edu).

Digital Object Identifier 10.1109/JSAC.2007.070513.

I. Introduction

IN THIS paper, we focus on multiple users (referred to interchangeably as applications, peers or stations) transmitting delay-sensitive video bitstreams across the same multi-hop wireless local area network (WLAN). Such wireless infrastructures often provide dynamically varying resources with only limited support for the Quality of Service (QoS) required by real-time multimedia applications. Hence, efficient solutions for multimedia streaming must accommodate time-varying bandwidths and probabilities of error introduced by the shared nature of the wireless medium and quality of the physical connections. In the studied distributed transmission scenario, users need to proactively collaborate in sharing the available wireless resources, in order to ensure that the various multimedia applications are provided with the necessary QoS. Such collaboration is needed due to the shared nature of the wireless infrastructure, where the cross-layer transmission strategy deployed by one user impacts and is impacted by the other peers.

Prior research on multi-user multimedia transmission over multi-hop wireless networks has focused on centralized, flow-based resource allocation strategies based on a pre-determined rate-requirement [1], [2]. These solutions are not scalable to the network size or the number of users and attempt to solve the end-to-end routing and path selection problem as a combined optimization using algorithms designed for Multi-Commodity Flow [3] problems. Such an optimization ensures that the end-to-end utility function (benefit) is maximized while satisfying constraints on individual link capacities. For instance, in [4], a dynamic routing policy based on queuing backpressure is proposed, which ensures that the average delay is bounded for the various users as long as the transmission rates are inside the capacity region of the network. However, the flow-based optimization does not guarantee that explicit packet-based delay constraints are met for video applications. These network layer research papers do not consider the real-time adaptation to time-varying channel conditions, video characteristics and encoding parameters (that influence packet-based delay constraints). Importantly, they do not take into account the loss tolerance provided by video applications, which can be exploited by the wireless network to support a larger number of users. Therefore, these solutions often lead to inferior network efficiency and suboptimal resulting qualities for the video users.

Alternatively, the majority of the video-centric research does not consider the protection techniques available at lower layers of the protocol stack (MAC, PHY) and/or optimizes the video transport using purely end-to-end metrics, thereby excluding the significant gains of cross-layer design [5], [6], [7]. Recent results on the practical throughput and packet loss analysis of multi-hop wireless networks have shown that the incorporation of appropriate utility functions (that take into account specific parameters of the protocol layers such as the expected retransmissions, the error rate and bandwidth of each link [7], as well as expected transmission time [8]) can significantly impact the actual performance. In [9], an integrated cross-layer optimization framework was proposed that considers the video quality impact. However, the solution proposed in [9] considers only the single user case, where a set of paths and transmission opportunities are statically pre-allocated for each video application. This leads to a sub-optimal, non-scalable solution for the multi-user case, which ignores important problems such as inefficient routing and time allocation to avoid interference among neighboring nodes. In summary, while significant contributions have been made to enhance the separate performance of the various OSI layers, no framework exists that integrates distributed and adaptive routing and resource allocation with cross-layer optimization for efficient multi-user multimedia streaming over multi-hop wireless networks.

In this paper, we propose such an integrated cross-layer solution for multiple video users. Our solution relies on the users' agreement to collaborate by dynamically adapting the quality of their multimedia applications to accommodate the more important flows/packets of other users. Unlike commercial multi-user systems, where the incentive to collaborate is minimal and there are often free-riders, we investigate the proposed approach in an enterprise network setting where users exchange accurate and trustable information about their applications (e.g. packet priorities). In our setting, the importance of the packets is determined based on their contribution to the overall distortion of a particular video as well as their delay deadlines. This information is encapsulated in the header of each transmitted packet and is used by intermediate nodes to drive the cross-layer transmission strategies. Moreover, our priority queuing approach also enables path diversity gains due to the delay-optimized dynamic routing, since the packets of the same application may be transmitted over different paths between the source and destination peers.

To increase the number of simultaneous users as well as to improve their performance given time-varying network conditions, we deploy scalable video coding schemes that enable a fine-granular adaptation to changing network conditions and a higher granularity in assigning the packet priorities. In our set-up, each user has a distinct source-destination pair. We assume a directed acyclic multi-hop overlay network [10] that can convey (in real-time) information about the expected delay for each priority class from a specific node to the destination. Each receiving node performs polling-based contention-free media access [19] that dynamically reserves a transmission opportunity interval in a service interval (SI). The network topology and the corresponding channel condition of each link are assumed to remain unchanged within the SI. Each node maintains a queue containing video packets from various users and correspondingly determines the transmission strategies based on the network information feedback from the neighbor nodes of the next hop. At intermediate nodes, we select the next hop based on a shortest-delay policy similar to the Bellman-Ford routing algorithm [12]. However, in our approach, we explicitly consider the packet deadlines and their priorities. Based on this intermediate node selection, we determine the expected delay for the packet and relay this information via the overlay network to the previous nodes.

The main contributions of this paper are listed below.

1. **Packet-based vs. flow-based/layer-based solutions**

We introduce a novel video streaming approach based on priority queuing that enables us to optimize the cross-layer transmission strategies per packet. The proposed cross-layer adaptation differs from existing solutions for multimedia transmission over multi-hop networks, where the path (or limited multiple paths) is predetermined for the entire bitstream or layer [9]. Moreover, the MAC retransmission and PHY link adaptation are often not considered for these flow-based/layer-based solutions [2]. Our approach is based on a multi-path routing algorithm that determines the next relay per packet. The proposed priority and delay-driven approach allows us to avoid global optimizations based on pre-determined rate requirements or path selections, which are not adaptive to network changes, the number of users or streamed video content characteristics.

2. **Distributed solution based on dynamic routing vs. conventional centralized solutions**

Existing research [2], [6] poses the problem of multi-user resource allocation and cross-layer adaptation over ad-hoc wireless networks as a static, centralized optimization that maximizes the utility (e.g. video quality) of the various users given pre-determined channel (capacity) constraints [15] and video rate requirements. These solutions have several limitations. First, the video bitstreams are changing over time in terms of required rates, priorities and delays. Hence, it is difficult to timely allocate the necessary bandwidths across the wireless network infrastructure to match these time-varying application requirements. Second, the delay constraints of the various packets are not explicitly considered in centralized solutions, as this information cannot be relayed to a central resource manager in a timely manner. Third, the complexity of the centralized approach grows exponentially with the size of the network and number of video flows. Finally, the channel characteristics of the entire network (the capacity region of the network) need to be known for this centralized, oracle-based optimization. This is not practical as channel conditions are time-varying, and having accurate information about the status of all the network links is not realistic.

Alternatively, in our solution, we optimize the cross-layer strategies (dynamic routing, MAC retransmission limit, and PHY modulation and coding scheme) per packet at the various intermediate nodes, in a distributed manner, which allows us to efficiently adapt to changes in the video bitstream, channel characteristics, and network resource. This approach is well suited for the informationally decentralized nature of the investigated multi-user video transmission problem. We also discuss the required information/parameters exchange among networks/layers for implementing such a distributed solution.

3. **Priority queuing with interference consideration**

Our solution aims at minimizing the packet loss rate of the packets in higher priority video classes based on the proposed priority queuing analysis. The analysis is performed for network environments with and without transmission interference consideration. To cope with the interference problems that exist in multi-hop networks due to the broadcast nature of the wireless medium, we adopt a polling-based, contention-free MAC that allocates transmission opportunities at each node to the various classes/packets based on their priorities [19]. To analyze the expected waiting time for the various packets in the presence of interference, we apply a novel virtual queuing method based on the "service-on-vacation" queuing model.

4. **Bottleneck identification**

Using our priority queuing analysis, we can estimate the expected packet loss at the transmitter side. This information can be used by the application layer to decide how many quality layers are transmitted or to adapt its encoding parameters (in the case of real-time encoding) to improve its video quality performance given the current number of users, priorities of the competing streams and network conditions, but also, importantly, to alleviate the network congestion. Note that our analysis provides this network bottleneck identification for each priority class, which is used in our solution to simplify the routing decision strategies. Furthermore, this information can be exploited to improve the network infrastructure such that it can support various multimedia application scenarios under different levels of network congestion.

The rest of this paper is organized as follows. Section II introduces the multi-user video streaming specification (video priority classes, network specification, cross-layer parameters etc.) and subsequently gives the cross-layer optimization problem formulation and highlights the need for a distributed per-packet solution. In Section III, we present our distributed solution which involves dynamically selecting relays that minimize the end-to-end packet loss probability of the higher priority video packets of the various users. In Section IV, we present the queuing delay analysis required in the proposed solution to determine the expected delay at each node. Based on the expected delay, a relay will be dynamically selected. In this section, we do not consider the effect of interference, as is the case in wireless networks where the nodes can simultaneously transmit and receive in orthogonal channels. Subsequently, in Section V, our analysis is extended to a wireless network environment where the transmission is performed in the same channel, and thus the interference needs to be considered. In Section VI, we show that the proposed distributed routing algorithm converges to a steady-state under certain assumptions. Finally, Section VII presents our simulation results, and Section VIII concludes the paper.

II. Multi-User Video Streaming Specification

A. Video Priority Classes

We assume that there are V video users (with distinct source-destination pairs) sharing the same multi-hop wireless infrastructure. In [23], it has been shown that partitioning a scalable embedded video flow (stream) into several priority classes (quality-layers) can improve the number of simultaneously admitted stations in a congested 802.11a/e WLAN infrastructure, as well as the overall received quality. Similarly, in this paper, we categorize the video units (video packets, video subbands, video frames) of the video bitstream into several priority classes. We adopt an embedded 3D wavelet codec [25] and construct video classes by truncating the embedded bitstream [23]. We assume that the packets within each class have the same delay deadline (see e.g. [13], [23] for more detail on how the delay is computed per class). For a video sequence v, we assume there are N_v classes, and these video classes are characterized by:

- $\boldsymbol{\lambda}_v$, a vector of the quality impact of the various video classes. We prioritize the video classes based on this parameter. The video classes are organized in an embedded bitstream in terms of their video quality impact, i.e. $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{N_v}$.
- $\mathbf{R}_v$, a vector of the rate requirements of the various video classes.
- $\mathbf{d}_v$, a vector of the delay deadlines of the various video classes. Due to the hierarchical temporal structure deployed in 3D wavelet video coders (see [13], [29]), the lower priority packets also have a less stringent delay requirement, i.e. $d_1 \leq d_2 \leq \ldots \leq d_{N_v}$. This is the reason why we can prioritize the video bitstream only in terms of the quality impact. However, if the used video coder did not exhibit this property, we need to deploy alternative prioritization techniques $\lambda_k^{video}(\lambda_k, d_k)$ that jointly consider the distortion impact and delay constraints (see the more sophisticated methods discussed in e.g. [28], [33]).
- $\mathbf{L}_v$, a vector of the average packet lengths of the various video classes.
- $\mathbf{P}_v^{succ}$, a vector containing the probabilities of successfully receiving the packets in the various video classes at the destination.

We denote the video classes using f_k, which can be characterized by the elements $\lambda_k, R_k, d_k, L_k, P_k^{succ}$ in the above mentioned vectors.

At the client side, the expected received video quality for video v can be modeled using any desirable video rate-distortion model:

$$Q_v^{rec} = F_v\left(\boldsymbol{\lambda}_v, \mathbf{R}_v, \mathbf{d}_v, \mathbf{L}_v, \mathbf{P}_v^{succ}\right), \tag{1}$$

represented by the function $F_v(\cdot)$ which can be computed as in e.g. [13], [24], [29], based on the successfully received video classes.

We assume that the client implements a simple error concealment scheme, where the lower priority packets are discarded whenever the higher priority packets are lost [13]. This is because the quality improvement (gain) obtained from decoding the lower priority packets is very limited (in such embedded scalable video coders) whenever the higher priority packets are not received. For example, drift errors can be observed when decoding the lower priority packets without the higher priority packets [29]. Hence, we can write:

$$P_k^{succ} = \begin{cases} 0, \text{if } P_{k'}^{succ} \neq 1 \text{ and } f_{k'} \prec f_k \\ (1 - P_k) = E\left[I\left(D_k \leq d_k\right)\right], \quad \text{otherwise,} \end{cases} \tag{2}$$

where we use the notation in [28] - $f_{k'} \prec f_k$ to indicate that the class f_k depends on $f_{k'}$. Specifically, if f_k and $f_{k'}$ are classes of the same video stream, $f_{k'} \prec f_k$ means $k^{'} < k$ due to the descending priority ($\lambda_{k' > \lambda_k}$). This error concealment policy facilitates our priority queuing solution, which will be discussed in Section III. P_k represents the end-to-end packet loss probability for the packets of class f_k. D_k represents the experienced end-to-end delay for the packets of class f_k. $I(\cdot)$ is an indicator function. Note that the end-to-end probability P_k^{succ} depends on the network resource, competing users' priorities as well as the deployed cross-layer transmission strategies vector, which will be discussed in more detail in Section III-C.

B. Network Specification

Let $\Re = [\Gamma, \mathbf{C}]$ represent the network specification, where Γ represents the given network graph, and $\mathbf{C}$ represents the interference matrix. The network graph Γ defines the network nodes (including the source nodes, destination nodes and relays) and the available transmission links in the multi-hop wireless network. The interference matrix $\mathbf{C}$ defines whether or not two different links can transmit simultaneously, and will be discussed in Section V in more detail. Besides the V source-destination pairs, we assume the network graph Γ consists of H hops with M_h intermediate nodes (relays) at each h-th hop ($0 \le h \le H-1$). The number of source and destination nodes are the same, i.e. $M_0 = M_H = V$, and each node will be tagged with a distinct number m_h ($1 \le m_h \le M_h$) as shown in Fig. 1. The other parameters in the figure will be defined in the following subsection.

C. Cross-Layer Joint Transmission Strategy Vector

Next, we define the transmission strategies of video units (video packets) at various layers across the network. Let us define the cross-layer joint strategies vector

$$\mathbf{STR} = \left\{ \begin{array}{c} STR_{h,m_h}(\vartheta) \,|\, \vartheta = 1 \ldots N^{tot}, \\ 1 \le m_h \le M_h \text{ and } 0 \le h \le H-1 \end{array} \right\}$$

as a vector of transmission strategies that can be deployed for packets present in the queue at the various nodes. N^{tot} is the total number of packets.

$$\begin{aligned} &STR_{h,m_h}(\vartheta \in f_k) \\ &= \left[\pi_{h,m_h}, \beta_{k,h+1,m_h+1}, \gamma^{MAX}_{k,m_h,m_{h+1}}(\vartheta), \theta_{k,m_h,m_{h+1}}(\vartheta)\right] \end{aligned}$$

represents the cross-layer transmission strategies for a packet ϑ at the intermediate node m_h at the h-th hop. Next, we describe the cross-layer transmission strategies.

- **Application Layer**

The packet headers are extracted at the various relays, to determine the packet priority, delay deadlines and packet lengths required for our cross-layer solution. Based on this information, the packet scheduling π_{h,m_h} should transmit a packet in the highest priority class f_k (i.e. the class with the highest quality impact) that is present in the queue at the node m_h. Thus, the packets with the largest quality contribution are scheduled first for transmission. The packets for which the delay deadline has expired are discarded from the queue. In other words, the higher priority packets are transmitted to the level that the network can accommodate, while the lower priority packets are queued and will be dropped if their delays exceed the delay deadline.

- **Network Layer**

We define β_{k,h,m_h} as the percentage of packets in priority class f_k (fraction of time) to select the node m_h as its relay at the h-th hop. We refer to this term as the *relay selecting parameter*. By assigning relays according to the relay selecting parameter, multiple paths can be chosen for the packets f_k in class , i.e. $0 \le \beta_{k,h,m_h} \le 1$. The relay selecting parameters provide a routing description across the network with multi-path capability. Whenever an intermediate node m_h is not reachable for class f_k, then $\beta_{k,h,m_h} = 0$. Since the total number of intermediate nodes in the h-th hop is M_h, we have $\sum_{m_h=1}^{M_h} \beta_{k,h,m_h} = 1$. Note that since each class f_k has a pre-determined destination (i.e. $m_H = \upsilon$), the relay selecting parameter at the last hop (β_{k,H,m_H}) is equal to '1', if m_H is the destination of the class, and '0', otherwise. Instead of selecting a fixed relay for all packets of class f_k, these video packets select the intermediate nodes m_{h+1} as their relay according to the corresponding $\beta_{k,h+1,m_h+1}$. At the intermediate nodes in the h-th hop, β_{k,h,m_h} are the incoming relay selecting parameters, and $\beta_{k,h+1,m_h+1}$ are the outgoing relay selecting parameters. The proposed dynamic routing solution is based on priority queuing while considering the lower layer goodput (effective transmission rate after factoring in packet losses) of all the possible link choices. We will discuss the relay selecting mechanism in Section III-B in more detail. Note that different paths can be selected for packets in the same class.

- **MAC Layer**

At the MAC layer, we assume the network deploys a protocol similar to that of IEEE 802.11a/e [19], which enables packet-based retransmission and polling-based time allocation. Let $\gamma^{MAX}_{k,m_h,m_{h+1}}(\vartheta)$ represent the maximum number of retransmissions for packet ϑ of priority class f_k over the link (m_h, m_{h+1}) at the h+1-th hop. The optimal retransmission limit is adapted based on the delay d_k deadline of the packet, which will be discussed in more detail in Section III-C.

- **PHY Layer**

Let $\theta_{k,m_h,m_{h+1}}(\vartheta)$ denote the modulation and coding scheme used for packet ϑ of class f_k for transmission over the link (m_h, m_{h+1}) at the h+1-th hop. (This is affected by the packet length). Let $T_{k,m_h,m_{h+1}}\left(\theta_{k,m_h,m_{h+1}}\right)$ and $p_{k,m_h,m_{h+1}}\left(\theta_{k,m_h,m_{h+1}}\right)$ represent the corresponding transmission rate and packet error rate (as shown in Fig. 1). In this paper, the goodput over the link is defined as $T^{goodput}_{k,m_h,m_{h+1}}\left(\theta_{k,m_h,m_{h+1}}\right) = T_{k,m_h,m_{h+1}}\left(\theta_{k,m_h,m_{h+1}}\right) \cdot \left(1 - p_{k,m_h,m_{h+1}}\left(\theta_{k,m_h,m_{h+1}}\right)\right)$.

In Section III-C, we will discuss the various cross-layer strategies in more detail.

D. Problem Formulations

- **Centralized Problem Formulation**

The conventional formulation of the multi-user wireless video transmission problem can be regarded as a cross-layer

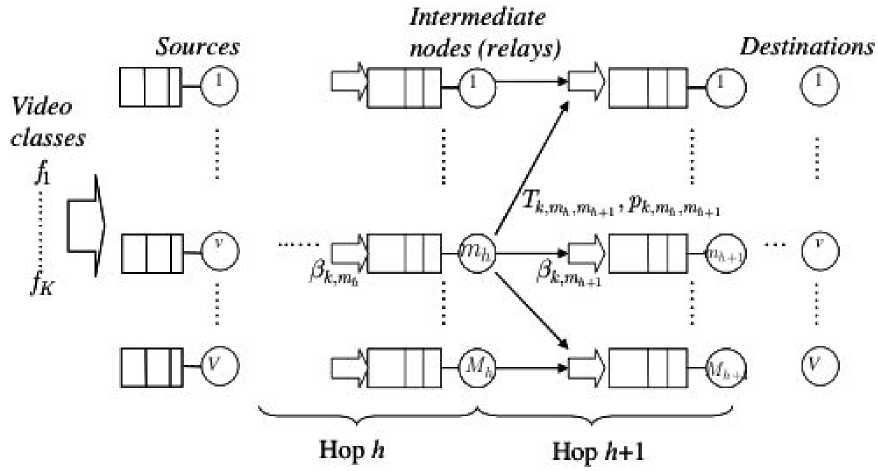

Fig. 1. The multi-hop overlay network model with V video users and K priority classes.

optimization that maximizes the overall video quality:[1]

$$\mathbf{STR}^{opt} = \arg\max_{\mathbf{STR}} \sum_{\upsilon=1}^{V} Q_\upsilon^{rec}\left(\boldsymbol{\lambda}_\upsilon, \mathbf{R}_\upsilon, \mathbf{d}_\upsilon, \mathbf{L}_\upsilon, \mathbf{P}_\upsilon^{succ}, (\mathbf{STR}, \Re)\right), \tag{3}$$

with the constraint that all successfully received packets must have their end-to-end delay D_k smaller than their corresponding delay deadline d_k (i.e. for every $\vartheta, \vartheta \in f_k, D_k(\vartheta) \leq d_k$).

Due to the informationally decentralized nature of the multi-users video transmission over multi-hop networks, a centralized solution for this optimization problem is not practical. For instance, the optimal solution depends on the delay incurred by the various packets across the hops, which cannot be timely relayed to a central controller. Instead, we propose a distributed packet-based solution to optimize the quality of the various users sharing the same multi-hop wireless infrastructure.

- **Proposed Distributed Problem Formulation**

Based on the proposed prioritized video classes and deployed error concealment strategy, a distributed cross-layer optimization can be formulated as a per-hop minimization of the end-to-end packet loss rate at the node m_h of the h-th hop:

$$\begin{aligned} STR_{h,m_h}^{opt}(\vartheta^* \in f_k) &= \arg\max_{STR} R_k \cdot P_k^{succ}\left(STR_{h,m_h}(\vartheta^*), \Re\right) \\ &= \arg\min_{STR} P_k\left(STR_{h,m_h}(\vartheta^*), \Re\right) \end{aligned} \tag{4}$$

where we minimize P_k for the selected packet $\vartheta^* \in f_k$ in the queue of the node m_h according to the scheduling π_{h,m_h}, with the delay constraint $D_k(\vartheta^*) \leq d_k$.

Note that in a directed acyclic multi-hop network shown in Fig. 1, the end-to-end packet loss probability P_k can be decomposed based on the hop-by-hop packet loss probability $P_{k,h}$:

$$P_k = 1 - \left(\prod_{h=0}^{H-1}(1 - P_{k,h})\right), \tag{5}$$

where $P_{k,h}$ represents the packet loss probability incurred due to delay deadline expiration during a specific hop h, given that the packet was not lost in the previous hop. In the next section, we present our distributed cross-layer solution of Eq. (4) based on the dynamic routing over such multi-stage overlay structure.

[1]A Max-Min fairness criterion can also be appiled to address the fairness issue, which will affect the prioritization λ_k values accordingly.

III. A Distributed Packet-Based Solution Based on Priority Queuing

In this section, we present our distributed packet-based solution. We show that the packet priorities (determined by λ_k for class f_k) and their delay constraints (d_k) drive the selection of optimal transmission strategies at the different layers in a distributed manner at each hop.

A. Required Information Feedback Among Network Nodes for the Distributed Solution

The proposed distributed approach not only simplifies the proposed cross-layer solution but also makes it adaptive to the varying network characteristics, as it does not require feedback about the entire network status. At each node, the transmission strategies for the prioritized video packets are determined based on the information feedback from the neighboring nodes. In order to implement the mentioned distributed solution for multimedia transmission based on priority queuing, the following two types of information feedback to a node m_h are provided:

- $E\left[Delay_{k,m_{h+1}}\right]$: the expected delay from nodes m_{h+1} to the destination node of the packets of class f_k (this information can be relayed by the overlay infrastructure and is required for the dynamic routing solution, which will be discussed in Section III-B).
- *SINR*: the Signal-to-Interference-Noise-Ratio (SINR) from the nodes m_{h+1} in the next hop that are able to establish a link with node m_h according to the network graph Γ. This information can easily be extracted from existing 802.11 WLAN standards [19].

We provide a block diagram in Fig. 2 that indicates the parameters/information that need to be exchanged across layers/various nodes in the proposed cross-layer transmission solution.

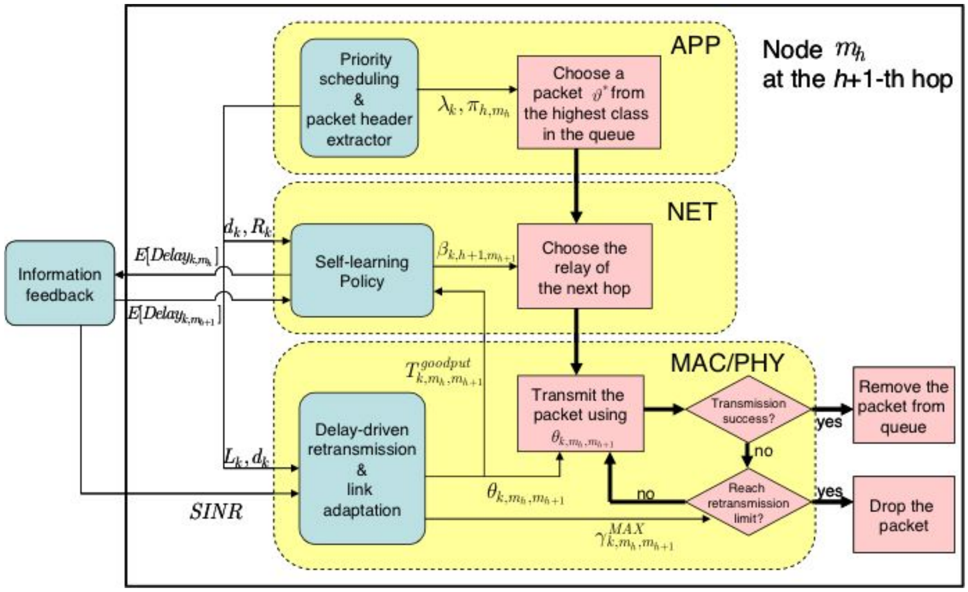

Fig. 2. Integrated block diagram of the proposed distributed per-packet algorithm.

B. Self-Learning Policy for Dynamic Routing

In this section, we provide our dynamic routing solution that minimizes the end-to-end packet loss probability P_k (see Eq. (4)). By definition $P_k = E[I(D_k > d_k)]$ and thus, minimizing P_k is equivalent to minimizing the expected end-to-end delay $E[D_k]$, given a fixed delay deadline d_k for the packets of class f_k.

To minimize the end-to-end delay over the multi-stage overlay structure shown in Fig. 1, we propose a dynamic routing policy to determine the relay selecting parameters. Recall that each node m_h maintains and feeds back to the previous hop the expected delay from itself to the destination $E[Delay_{k,m_h}]$ for each class f_k. $E[Delay_{k,m_h}]$ becomes the cost that will be minimized at each stage, and will be updated at each node using the information feedback from the next hop. Note that $E[Delay_{k,m_h}]$ equals $E[D_k]$, if the node m_h is the source node of the class f_k packets. Specifically, the expectation of delay to the destination of each class can be determined at node m_h as Eq. (6) [30], where $E\left[Delay_{k,m_{h+1}}\right]$ is given by the information feedback obtained from the nodes of the next hop, and the relay selecting parameter $\beta_{k,h+1,m_h+1}$ is chosen such that $E[Delay_{k,m_h}]$ is minimized. $E[W_{k,m_h}]$ is the average queuing delay at the current relay queue, which can be obtained using the priority queuing analysis introduced in Section IV. In a congested network, Eq. (6) is dominated by the second term (the accumulated queuing delay in the rest of the network). Thus, we can simplify this equation as Eq. (7). To determine the relay selecting parameter $\beta_{k,h+1,m_{h+1}}$, we apply the following soft minimum (probabilistic) policy to enable transmission across multiple paths:

$$\beta_{k,h+1,m_{h+1}} = \frac{Coeff_k}{1+\kappa E\left[Delay_{k,m_{h+1}}\right]^{\varphi}}. \tag{8}$$

$Coeff_k$ are normalized coefficients to make sure that the summation of the percentages (fraction) equals to one:

$$Coeff_k = \left(\sum_{m_{h+1}\in \mathbf{M}_{k,h+1}} \frac{1}{1+\kappa E\left[Delay_{k,m_{h+1}}\right]^{\varphi}} \right)^{-1}, \tag{9}$$

where κ and φ are constants. Eq. (8) is inspired from the balking arrival probability in queuing theory [18]. The value of κ is set depending on the arrival rate according to [18]. The term φ weighs the average delay $E\left[Delay_{k,m_{h+1}}\right]$ such that the routing policy favors paths leading to significantly lower delays to the destination. $\mathbf{M}_{k,h+1}$ represents a set of nodes m_{h+1} in the h+1-th hop that feedback the information $E\left[Delay_{k,m_{h+1}}\right]$. We set $\beta_{k,h+1,m_{h+1}} = 0$ for the nodes whose information feedback is not received, indicating that node m_{h+1} is not connected to node m_h using the overlay infrastructure [10]. We refer to this relay selecting policy as the *self-learning policy*, since the decision of $\beta_{k,h+1,m_{h+1}}$ will influence the future information feedback. The complete algorithm of the proposed self-learning policy including the information feedback is given in the Appendix. The self-learning policy will dynamically adapt the relay selection to minimize the delay through the network. Finally, the next relay m_{h+1} can be determined for the packet ϑ^* at the node m_h according to the percentage (time fraction) $\beta_{k,h+1,m_{h+1}}$.

This method is inspired by the Bellman-Ford shortest path (delay) routing algorithm [12] that minimizes the end-to-end delay across the network. Our routing algorithm reduces to the well-known Bellman-Ford algorithm when $\beta_{k,h+1,m_{h+1}} = 1$ to the node m_{h+1} that feedbacks the smallest $E\left[Delay_{k,m_{h+1}}\right]$ (which can be implemented using a large φ). Note that our algorithm is prioritized and the delay of class f_k will be influenced by equal or higher priority traffic, which will be discussed in more details in Section IV.

C. Delay-Driven Policy for MAC/PHY

If a node m_{h+1} is selected with probability $\beta_{k,h+1,m_{h+1}}$ for the selected packet ϑ^* at each intermediate node m_h, we

$$E\left[Delay_{k,m_h}\right] = \min_{\beta_{k,h+1,m_{h+1}}} \left(E\left[W_{k,m_h}\left(\beta_{k,h+1,m_{h+1}}, T^{goodput}_{k,m_h,m_{h+1}}\right)\right] + \sum_{m_{h+1}=1}^{M_{h+1}} \beta_{k,h+1,m_{h+1}} E\left[Delay_{k,m_{h+1}}\right] \right) \tag{6}$$

$$E\left[Delay_{k,m_h}\right] = E\left[W_{k,m_h}\right] + \min_{\beta_{k,h+1,m_{h+1}}} \left(\sum_{m_{h+1}=1}^{M_{h+1}} \beta_{k,h+1,m_{h+1}} E\left[Delay_{k,m_{h+1}}\right] \right) \tag{7}$$

can determine the corresponding transmission rate $T_{k,m_h,m_{h+1}}$ and the packet error rate $p_{k,m_h,m_{h+1}}$ for the link by selecting $\theta_{k,m_h,m_{h+1}}$ based on the *link adaptation* scheme presented in [14]. To describe the channel conditions, we assume as in [20] that each wireless link is a memoryless packet erasure channel. The link packet error rate for a fixed packet of length L_k bits is $p_{k,m_h,m_{h+1}}\left(\theta_{k,m_h,m_{h+1}}, L_k\right) = 1 - \left(1 - BER\left(\theta_{k,m_h,m_{h+1}}\right)\right)^{L_k}$, where $BER\left(\theta_{k,m_h,m_{h+1}}\right)$ is the bit error rate when the modulation scheme $\theta_{k,m_h,m_{h+1}}$ is selected. The packet error rate and the effective transmission rate (goodput) can be approximated using the sigmoid function as in [20]:

$$p_{k,m_h,m_{h+1}}\left(\theta_{k,m_h,m_{h+1}}, L_k\right) = \frac{1}{1+e^{\zeta(SINR-\delta)}}, \tag{10}$$

$$T^{goodput}_{k,m_h,m_{h+1}} = \frac{T_{k,m_h,m_{h+1}}\left(\theta_{k,m_h,m_{h+1}}\right)}{1+e^{-\zeta(SINR-\delta)}} \tag{11}$$

where *SINR* is the Signal-to-Interference-Noise-Ratio, and ζ and δ are constants corresponding to the modulation and coding schemes for a given packet length [20]. This method maximizes the goodput given the average packet length L_k of the specific class over a selected link (m_h, m_{h+1}) based on the SINR feedback.

For a fixed $T^{goodput}_{k,m_h,m_{h+1}}$, we choose the retransmission limit $\gamma^{MAX}_{k,m_h,m_{h+1}}$ for the selected packet ϑ^* in the priority class f_k such that the delay constraint is satisfied. Specifically, let $delay^{curr}_{h,m_h}(\vartheta^*)$ represent the current measured delay incurred by the selected packet from the source to a current node m_h. The maximum retransmission limit for the packet of class f_k over the link from m_h to m_{h+1} is determined based on the delay deadline d_k (where $\lfloor\cdot\rfloor$ is the floor operation) [23]:

$$\gamma^{MAX}_{k,m_h,m_{h+1}}(\vartheta^*) = \left\lfloor \frac{T^{goodput}_{k,m_h,m_{h+1}}\left(d_k - delay^{curr}_{h,m_h}(\vartheta^*)\right)}{L_k} \right\rfloor - 1. \tag{12}$$

D. Complexity Analysis in Terms of Route Selection

- **Complexity of a conventional centralized approach (exhaustive search)**

Assume that we have a total of $K = \sum_{v=1}^{V} N_v$ classes across the users in an H-hop network. Let us assume the maximum number of the intermediate nodes that can be selected as a relay for a class f_k packet at the h-th hop is $\mathcal{C}_{k,h}$. The maximum number of possible end-to-end paths is $\prod_{h=1}^{H} \mathcal{C}_{k,h}$ then . Thus, the total complexity (in terms of the number of path combinations) of a centralized exhaustive search can be up to $\prod_{k=1}^{K}\prod_{h=1}^{H} \mathcal{C}_{k,h}$. Due to the informationally decentralized nature of the wireless multi-hop network, the control overhead of the centralized approach can induce a significant amount of delay (inefficient when the number of hops is large) for doing the optimization. Hence, the distributed approach is proposed and its complexity is also investigated.

- **Complexity of the proposed distributed relay selecting algorithm**

In our distributed approach, for a packet (of class f_k) at the node m_h at the h-th hop, the complexity is $\mathcal{C}_{k,h}$ (i.e. $\mathcal{C}_{k,h}$ is the number of relays that can be selected). Thus, the complexity for the packet over the H hops is $\sum_{h=1}^{H} \mathcal{C}_{k,h}$. Then, the total complexity by considering all the different classes equals $\sum_{k=1}^{K}\sum_{h=1}^{H} \mathcal{C}_{k,h}$.

IV. Multi-Hop Priority Queuing Analysis for Multimedia Transmissions

In this section, we present the analysis of the expected queuing delay $E\left[W_{k,m_h}\right]$ (that forms $E\left[Delay_{k,m_h}\right]$) and packet loss probabilities P_{k,h,m_h} using queuing theory. Based on these values, a relay will be dynamically selected. In this section, we do not consider the effect of interference. In the next section, we extend our analysis to a network environment where the interference is considered. Before introducing the queuing model, several assumptions for the priority queuing analysis are made in Section IV-A. Then in Section IV-B, we determine the end-to-end packet loss probability P_k by considering a simple 2-hop network structure (with only one set of intermediate nodes), which we refer to as the "elementary structure." We further extend this result by cascading the elementary structure to create a general H-hop network (with $H-1$ sets of intermediate nodes) in Section IV-C.

A. Assumptions for Priority Queuing Analysis

The priority queuing analysis is based on the following assumptions:

1) We assume that the arrival traffic at each intermediate node is from various video sources and is assumed to be a Poisson process. This approximation is reasonable if the number of intermediate nodes is large enough and the selection of paths is relatively balanced. We model the queues in the intermediate nodes as preemptive-repeat priority M/G/1 queues [18]. For our analysis, we do not apply the non-preemptive model because when a packet with higher priority arrives at the queue, it will

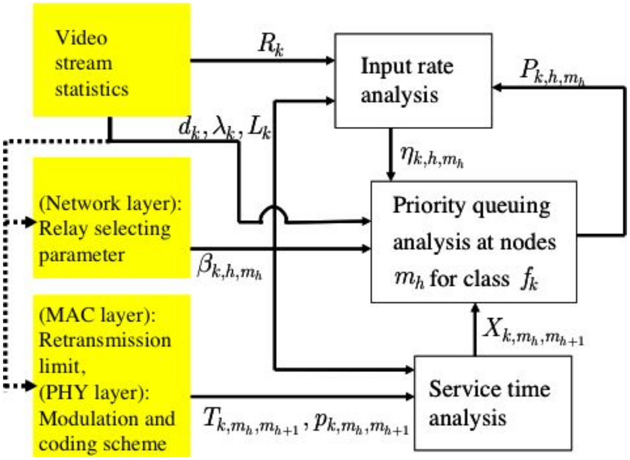

Fig. 3. Priority queuing analysis system map.

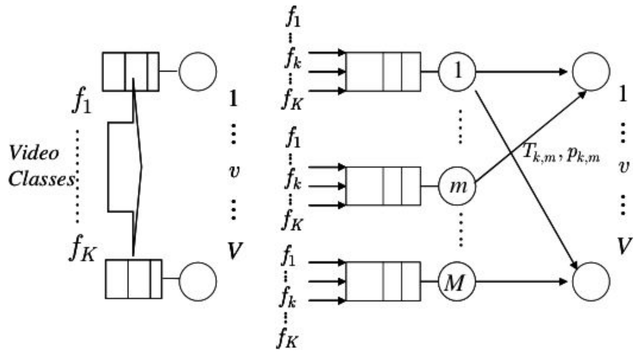

Fig. 4. The elementary structure.

interrupt future transmissions (i.e. the retransmission of the same packet when this is lost or the transmission of a lower priority packet). The preempted packet will be retransmitted later.

2) We assume the transmission rate and the packet error rate for each link are fixed in a SI, as these are determined, as discussed in Section III, by selecting the appropriate modulation and coding scheme using the link adaptation mechanism. As an example, let us consider a link from node m_n to node m_{h+1}. The selected $\theta_{k,m_h,m_{h+1}}$ determines the physical transmission rate $T_{k,m_h,m_{h+1}}$ (Eq. (11)) and packet error rate $p_{k,m_h,m_{h+1}}$ (Eq. (10)) for class f_k over this link. Each packet will be retransmitted until it is either successfully received or discarded because its delay deadline d_k was exceeded. In summary, assuming the packet length of a class f_k is fixed to be L_k, with a header length L_{Header}, we can formulate the service time for a packet as a geometric distribution from these assumptions. If $X_{k,m_h,m_{h+1}}$ is the service time, then the probability of there being exactly i transmissions (including retransmissions) will be Eq. (13), where *Time*$_o$ denotes the time overhead including the time of waiting for the acknowledgement, polling delay, and the expected background traffic in the contention-based period, etc [19].

3) We assume that the queue waiting time dominates the overall delay (i.e. the transmission delay across the various network hops is relatively small).

Fig. 3 illustrates the deployed priority queuing at each intermediate node. Given the application layer video priorities and class characteristics, the relay selecting parameters of the network layer, the retransmission strategy at the MAC layer, and the modulation and coding scheme at the PHY layer, we can determine the average input rate and the service time for the packets in a certain class, thereby obtaining a steady state waiting time distribution for all video priority classes. In the next subsection, we analyze the video quality problem using priority queuing analysis for our elementary structure with only one set of intermediate nodes (a 2-hop structure).

B. Priority Queuing Analysis for an Elementary Structure

We first analyze the priority queuing model for an elementary structure. The elementary structure is an overlay 2-hop network with V video streams and one set of M intermediate nodes (relays) between sources and destinations, as illustrated in Fig. 4. A packet of class f_k will be routed from its source through an intermediate node m with percentage $\beta_{k,m}$ toward its destination. Each intermediate node contains a queue that schedules the waiting packets based on their header information (quality impact parameter λ_k and delay deadline d_k).

From the geometric distribution assumption above, the first and second moment of the service time at queue m are (using the approximation $p_{k,m}^{\gamma_{k,m}^{MAX}+1} \ll 1$):

$$E\left[X_{k,m}\right] = \frac{\widehat{L}_{k,m}\left(1 - p_{k,m}^{\gamma_{k,m}^{MAX}+1}\right)}{T_{k,m}\left(1 - p_{k,m}\right)} \approx \frac{\widehat{L}_{k,m}}{T_{k,m}^{goodput}} \tag{14}$$

$$E\left[X_{k,m}^2\right] = \frac{\widehat{L}_{k,m}^2\left(1 + p_{k,m}\right)}{T_{k,m}^2\left(1 - p_{k,m}\right)^2}. \tag{15}$$

For a class f_k, which is relayed through the intermediate node m, let $\widehat{L}_{k,m}$ be the effective packet "length," which includes both the video packet L_k length and the time overhead *Time*$_o$ (as in Eq. (13)). $T_{k,m}$ and $p_{k,m}$ are the transmission rate and packet error rate for the packets of class f_k that are transmitted through the intermediate node m to the destination. Note that the modulation and coding strategy changes depending on the chosen link status, and this will consequently impact[2] $T_{k,m}$ and $p_{k,m}$ (see Eq. (10), Eq. (11)).

Let $\eta_{k,m}$ be the average arrival rate of the Poisson input traffic of queue m for class f_k. Given the relay selecting parameters $\beta_{k,m}$, we have:

$$\eta_{k,m} = \beta_{k,m} R_k \left(1 - P_{k,0}\right), \tag{16}$$

where $P_{k,0}$ = Prob$(W_{k,0} > d_k)$ is the packet loss probability at the source queue due to packet expiration and can be

[2]To simplify the notation, here as well as in the subsequent part of the paper, we do not explicitly state the dependency of the throughput, goodput, packet error rate, etc. on the optimal modulation strategy chosen for that link, but assume that this is implicitly considered.

$$\text{Prob}\left(X_{k,m_h,m_{h+1}} = i \times \left(\frac{L_k + L_{Header}}{T_{k,m_h,m_{h+1}}} + \mathit{Time}_o\right)\right) = \left(1 - p_{k,m_h,m_{h+1}}\right) p_{k,m_h,m_{h+1}}^{i-1}, \text{ for } i \le \gamma_{k,m_h,m_{h+1}}^{MAX} + 1, \tag{13}$$

calculated from the queue waiting time $W_{k,0}$ tail distribution for each class.

Let $E[W_{k,m}]$ be the average waiting time of class f_k that goes through node m. For a preemptive-priority M/G/1 queue, the priority queuing analysis gives the following result [12]:

$$E[W_{k,m}] = \frac{\sum_{i=1}^{k} \eta_{i,m} E\left[X_{i,m}^2\right]}{2\left(1 - \sum_{i=1}^{k-1} \eta_{i,m} E[X_{i,m}]\right)\left(1 - \sum_{i=1}^{k} \eta_{i,m} E[X_{i,m}]\right)}. \tag{17}$$

Based on this expected average waiting time, the probability of packet loss due to the expiration can be calculated by the tail distribution of the waiting time (see Eq. (18)). In Eq. (18), we adopt the G/G/1 tail distribution approximation based on the work of [16], [17]. Let us now express this probability in terms of the packet delay deadline d_k. This probability of packet loss (at the intermediate node m) is denoted $P_{k,m}$ (recall that the waiting time is assumed to dominate the overall delay):

$$P_{k,m} = \text{Prob}\left(W_{k,m} + E[W_{k,0}] > d_k\right), \tag{19}$$

where $E[W_{k,0}]$ is the expected queuing delay of the packets at the source queue, which depends on the number of packets of a class in one GOP. Then, the end-to-end packet loss probability P_k for class f_k can be calculated as:

$$P_k = 1 - (1 - P_{k,0})\left(1 - \sum_{m=1}^{M} \beta_{k,m} P_{k,m}\right). \tag{20}$$

We can observe from the above derivation that the resulting end-to-end packet loss probability for each class f_k is affected by the various cross-layer parameters (as shown in Eq. (4)): the relay selecting parameters $\beta_{k,m}$, the modulation and coding scheme $\theta_{k,m_h,m_{h+1}}$ that affects the average queue waiting time. Finally, the received video quality can be estimated by substituting Eq. (20) into Eq. (1).

C. Generalization to the Multi-Hop Case

We now extend our analysis to a general directed acyclic multi-hop overlay network (as shown in Fig. 1) by cascading the elementary structure. Importantly, note that the deployed structure is very general and any multi-hop network that can be modeled as a directed acyclic graph can be modified to fit into this overlay structure by introducing virtual nodes [11]. We introduce virtual nodes with zero service time for users that have a smaller number of hops, and fix the path for particular classes to pass through the virtual node (by setting $\beta_{k,h,m_h} = 1$). Methods to construct such overlay structures given a specific multi-hop network and a set of transmitting-receiving pairs can be found in [21], [22].

The network is assumed to have H hops from sources to destinations. All the queues in the intermediate nodes perform a preemptive-repeat priority M/G/1 model as mentioned in the previous subsection. For the queue at node m_h, let η_{k,h,m_h} be the average arrival rate between the h-th hop and (h+1)-th hop ($1 \le h \le H-1$), and $P_{k,h-1}$ be the packet loss due to delay expiration from the previous hop. $R_{k,h}$ is the updated arrival rate of class f_k for all the intermediate nodes between the h-th hop and (h+1)-th hop, and we set $R_{k,0} = R_k$ for the source nodes. Then, the average arrival rates η_{k,h,m_h} have the following recursive relationship:

$$R_{k,h} = (1 - P_{k,h-1}) R_{k,h-1}, \tag{21}$$

$$\eta_{k,h,m_h} = \beta_{k,h,m_h} R_{k,h}. \tag{22}$$

Eq. (21) illustrates that the video rate was reduced from hop to hop due to the packet deadline expiration. Eq. (22) shows that the average input rate is distributed based on the relay selecting parameters at the h-th hop.

Recall that X_{k,h,m_h} is the service time of the priority M/G/1 queue at node m_h between the h-th hop and (h+1)-th hop. Given the relay selecting parameters, we can obtain the first two moments of the service time (see Eq. (23)). Similarly, recall W_{k,m_h} is the queue waiting time at node m_h for video class f_k. Then, the expected average value can be calculated similarly to Eq. (17) (see Eq. (24)). Therefore, the expectation of the waiting time $E[W_{k,h}]$ over the h-th hop for packets of class f_k is:

$$E[W_{k,h}] = \sum\nolimits_{m_h=1}^{M_h} \beta_{k,h,m_h} E[W_{k,m_h}]. \tag{25}$$

The probability of packet loss due to the expiration becomes Eq. (26). Similar to Eq. (19), the probability of packet loss at the node m_h is the waiting time tail distribution when the accumulated waiting time exceeds the delay deadline. Then, the expected hop-by-hop packet loss probability of the hop h is:

$$P_{k,h} = \sum\nolimits_{m_h=1}^{M_h} \beta_{k,h,m_h} P_{k,h,m_h}. \tag{27}$$

Recursively, we can write

$$(1 - P_{k,H-1}) \cdot R_{k,H-1} = \prod_{h=0}^{H-1} (1 - P_{k,h}) \cdot R_k. \tag{28}$$

Finally, the received video quality can be estimated by substituting Eq. (27) into Eq. (5) and Eq. (1). Note that the model can be applied even for the 1-hop case, the average waiting time at the source $E[W_{k,0}]$, and the packet loss probability $P_{k,0}$ = Prob$(W_{k,0} > d_k)$ can be obtained using the above equations.

V. Priority Queuing Analysis Considering Interference of Wireless Networks

In Section IV, we determined the priority queuing analysis without considering the interference of other simultaneous transmissions. This can be considered as being the case in a network with multiple orthogonal channels for transmission. However, for regular wireless networks, the interference is severely rooted in the broadcasting nature of the medium. Hence, it is important to include the performance degradation

$$\mathrm{Prob}\left(W_{k,m} > t\right) \approx \left(\sum_{i=1}^{k} \eta_{i,m} E\left[X_{i,m}\right]\right) \exp\left(-\frac{t\sum_{i=1}^{k} \eta_{i,m} E\left[X_{i,m}\right]}{E\left[W_{k,m}\right]}\right). \tag{18}$$

$$\begin{aligned} E\left[X_{k,h,m_h}\right] &\approx \sum_{m_{h+1}=1}^{M_{h+1}} \beta_{k,h+1,m_{h+1}} \frac{\widehat{L}_k}{T_{k,m_h,m_{h+1}}\left(1-p_{k,m_h,m_{h+1}}\right)}, \\ E\left[X_{k,h,m_h}^2\right] &\approx \sum_{m_{h+1}=1}^{M_{h+1}} \beta_{k,h+1,m_{h+1}} \frac{\widehat{L}_k^2\left(1+p_{k,m_h,m_{h+1}}\right)}{T_{k,m_h,m_{h+1}}^2\left(1-p_{k,m_h,m_{h+1}}\right)^2}, \end{aligned} \tag{23}$$

$$E\left[W_{k,m_h}\right] = \frac{\sum_{i=1}^{k} \eta_{i,h,m_h} E\left[X_{i,h,m_h}^2\right]}{2\left(1-\sum_{i=1}^{k-1} \eta_{i,h,m_h} E\left[X_{i,h,m_h}\right]\right)\left(1-\sum_{i=1}^{k} \eta_{i,h,m_h} E\left[X_{i,h,m_h}\right]\right)}. \tag{24}$$

$$\begin{aligned} P_{k,h,m_h} &= \mathrm{Prob}\left(W_{k,m_h} > d_k - \sum_{j=0}^{h-1} E\left[W_{k,j}\right]\right) \\ &\approx \left(\sum_{i=1}^{k} \eta_{i,h,m_h} E\left[X_{i,h,m_h}\right]\right) \exp\left(-\frac{\left(d_k - \sum_{j=0}^{h-1} E\left[W_{k,j}\right]\right)\left(\sum_{i=1}^{k} \eta_{i,h,m_h} E\left[X_{i,h,m_h}\right]\right)}{E\left[W_{k,m_h}\right]}\right) \end{aligned} \tag{26}$$

due to the interference effect. First, we introduce two matrices to describe the interference in Section V-A. Then, in Section V-B, we present the priority queuing analysis with the virtual-queue service time modification.

A. Incidence Matrix and Interference Matrix

In [15], a rate matrix was introduced to describe the state of the network at a given time. In [6], an elementary capacity graph was used to represent the physical layer state of the various links. In [26], a node-link incidence matrix was used. Here, we assume a similar incidence matrix to describe a network with n nodes and l links. This matrix is defined as $\mathbf{A} = \left[A_{ij}\right]_{n\times l}$, where i is the nodes' index, and j is the index of directional links:

$$A_{ij} = \begin{cases} 1, & \text{if link } j \text{ flows into node } i \\ -1, & \text{if link } j \text{ flows out of node } i \\ 0, & \text{otherwise} \end{cases}. \tag{29}$$

The existence of links is determined by the SINR value, i.e. links having a SINR below a predetermined value are not considered viable [20].

Additionally, we introduce here a matrix $\mathbf{C}$ to characterize the interference in the multi-hop network. Two types of interference are considered in this paper. One type of interference is the transmission rate decrease due to the SINR degradation. The other type of interference, which is referred as the feasibility of simultaneous transmission links, is from the fact that in a regular wireless network environment, a node cannot transmit and receive data at the same time, and it cannot transmit two flows and receive two flows at the same time due to the wireless radio limitation. First, let $\mathbf{B} = \left[B_{jk}\right]_{l\times l} = \mathbf{A}^T\mathbf{A}$. If $B_{jk} > 0$, there exists transmitter-receiver interference between link j and k. If $B_{jk} < 0$, there exists transmitter-transmitter or receiver-receiver interference between link j and k. If $B_{jk} = 0$, there exists no second type of interference between link j and k. The interference matrix $\mathbf{C} = \left[C_{jk}\right]_{l\times l}$ is defined as:

$$C_{jk} = \begin{cases} 1, & \text{if } B_{jk} = 0 \\ 0, & \text{if } B_{jk} \neq 0 \end{cases}. \tag{30}$$

Note that the interference matrix $\mathbf{C}$ is defined to observe the feasibility of simultaneous transmission links. Link j and link k could transmit simultaneously if and only if $C_{jk} = 1$.

Given the interference matrix $\mathbf{C}$, the set $\Phi = \{\Phi_z\}$ represents all the combinations of transmission links that can transmit simultaneously. A combination Φ_z must satisfy the following condition:

$$\prod_{j,k\in\Phi_z} C_{jk} = 1. \tag{31}$$

We denote link $l_h = (m_h, m_{h+1})$ to be the link connecting node m_h with node m_{h+1}. Denote the air-time fractions r_{Φ_z} as the average time portion (a probability) for the link combination Φ_z to happen in a SI [19]. Note that $\sum_{\Phi} r_{\Phi_z} = 1$.

In general, the decision on the routing as well as the nodes participating in the video streaming session depend largely on a number of system-related factors that transcend the video streaming problem [9] (e.g. node cooperation strategies/incentives and network coordination and routing policies imposed by the utilized protocols). Hence, such information can be provided by the negotiation and arbitration of the polling-based contention-free MAC protocol statistically. We define $PR^{(I)}_{\Phi_z,l_h}$ as the probability that a particular combination of links that simultaneously transmit (i.e. Φ_z) occurs, given that the link l_h is transmitting:

$$PR^{(I)}_{\Phi_z,l_h} = \begin{cases} 0, & \text{if } l_h \notin \Phi_z \\ \frac{r_{\Phi_z}}{\sum_{l_h\in\Phi_i} r_{\Phi_i}}, & \text{if } l_h \notin \Phi_z \end{cases} . \quad (32)$$

B. Virtual-Queue Service Time Modification

Since our model has only one server per queue at each intermediate node, only one transmission can take place at a time from the same queue. However, we still have to avoid the case that a receiver simultaneously receives more than one packet from distinct nodes. In fact, for a regular polling-based wireless network with a single channel, the packets are kept in the servers while waiting for the interfering transmission to finish the service. Hence, we assume that the servers at each intermediate node form a "virtual queue" to the same destination [31]. In a virtual queue, packets of different queues wait in turns at the servers to be transmitted to the same destination. The concept is similar to the "service on vacation" [12] in queuing theory, and the waiting time of the virtual queue can be regarded as the "vacation time." The total sojourn time (queue waiting time plus the transmission service time) of the virtual queue now becomes the actual service time at each of the intermediate nodes. As the packet in the server is waiting in the virtual queue, the node is able to receive packets from the previous hop. For simplicity, we assume that the receiving process can still be approximated as a regular Poisson process. In addition, the arrival process of the virtual queue is also assumed to be an M/G/1 priority queue.

Let $\widetilde{\eta}_{k,m_{h+1}}$ be the average arrival rate of class f_k to the virtual M/G/1 queue that has node m_{h+1} as its destination. Next, we denote all random variables for the virtual queues with a tail on it.

$$\widetilde{\eta}_{k,m_{h+1}} = \beta_{k,h+1,m_{h+1}} R_{k,h}, \quad (33)$$

where $R_{k,h}$ is the updated input rate after the h-th hop defined in Eq.(21).

Denote X_{k,l_h,Φ_z} as the service time of the priority M/G/1 queue in the node m_h, when the transmission is on the link $l_h = (m_h, m_{h+1})$ in the combination Φ_z. Both the first moment and the second moment need to be modified, since the channel is different due to the SINR degradation from simultaneous transmissions. $T_{k,m_h,m_{h+1}}$ is changed into T_{k,l_h,Φ_z}, and $p_{k,m_h,m_{h+1}}$ is changed into p_{k,l_h,Φ_z}. Let $\widehat{L}_{k,l_h,\Phi_z}$ represent the new effective packet length including the time overhead $Time_o$ for MAC operations similar to Eq.(14). The first three moments of X_{k,l_h,Φ_z} become (assuming $p_{k,l_h,\Phi_z}^{\gamma^{MAX}_{k,l_h}+1} \ll 1$):

$$\begin{aligned} E\left[X_{k,l_h,\Phi_z}\right] &\approx \frac{\widehat{L}_{k,l_h,\Phi_z}}{T_{k,l_h,\Phi_z}\left(1-p_{k,l_h,\Phi_z}\right)}, \\ E\left[X^2_{k,l_h,\Phi_z}\right] &\approx \frac{\widehat{L}^2_{k,l_h,\Phi_z}\left(1+p_{k,l_h,\Phi_z}\right)}{T^2_{k,l_h,\Phi_z}\left(1-p_{k,l_h,\Phi_z}\right)^2}, \\ E\left[X^3_{k,l_h,\Phi_z}\right] &\approx \frac{\widehat{L}^3_{k,l_h,\Phi_z}\left(1+4p_{k,l_h,\Phi_z}+p^2_{k,l_h,\Phi_z}\right)}{T^3_{k,l_h,\Phi_z}\left(1-p_{k,l_h,\Phi_z}\right)^3}. \end{aligned} \quad (34)$$

Let $\widetilde{S}_{k,m_{h+1}}$ be the service time of the virtual queue having destination node m_{h+1}. The first moment of service time for class f_k of this virtual M/G/1 queue can be obtained as:

$$\begin{aligned} E\left[\widetilde{S}_{k,m_{h+1}}\right] &= \sum_{m_h=1}^{M_h} \beta_{k,h,m_h} E\left[\widetilde{S}_{k,m_h,m_{h+1}}\right] \\ &= \sum_{m_h=1}^{M_h} \beta_{k,h,m_h} \sum_z PR^{(I)}_{\Phi_z,l_h} E\left[X_{k,l_h,\Phi_z}\right], \end{aligned} \quad (35)$$

where $E\left[\widetilde{S}_{k,m_h,m_{h+1}}\right]$ is the statistical average service time from intermediate node m_h to node m_{h+1} through all the possible transmission combinations Φ_z. The second and the third moment are similarly:

$$\begin{aligned} E\left[\widetilde{S}^2_{k,m_{h+1}}\right] &= \sum_{m_h=1}^{M_h} \beta_{k,h,m_h} \sum_z PR^{(I)}_{\Phi_z,l_h} E\left[X^2_{k,l_h,\Phi_z}\right], \\ E\left[\widetilde{S}^3_{k,m_{h+1}}\right] &= \sum_{m_h=1}^{M_h} \beta_{k,h,m_h} \sum_z PR^{(I)}_{\Phi_z,l_h} E\left[X^3_{k,l_h,\Phi_z}\right] \end{aligned} \quad (36)$$

Let random variable $\widetilde{W}_{k,m_{h+1}}$ be the waiting time of the virtual queue with node m_{h+1} as its destination. From the Pollaczek-Khinchin formula, the first moment of $\widetilde{W}_{k,m_{h+1}}$ for the virtual queue [12] is:

$$E\left[\widetilde{W}_{k,m_{h+1}}\right] = \frac{\sum_{i=1}^{k} \widetilde{\eta}_{i,m_{h+1}} E\left[\widetilde{S}^2_{i,m_{h+1}}\right]}{2\left(1-\sum_{i=1}^{k} \widetilde{\eta}_{i,m_{h+1}} E\left[\widetilde{S}^2_{i,m_{h+1}}\right]\right)} \quad (37)$$

Using the Takacs recurrence formula [18], we have the second moment:

$$\begin{aligned} E\left[\widetilde{W}^2_{k,m_{h+1}}\right] &= 2E\left[\widetilde{W}_{k,m_{h+1}}\right]^2 \\ &+ \frac{\sum_{i=1}^{k} \widetilde{\eta}_{i,m_{h+1}} E\left[\widetilde{S}^3_{i,m_{h+1}}\right]}{3\left(1-\sum_{i=1}^{k} \widetilde{\eta}_{i,m_{h+1}} E\left[\widetilde{S}^2_{i,m_{h+1}}\right]\right)} \end{aligned} \quad (38)$$

The expected virtual queue waiting time $E\left[\widetilde{W}_{k,m_{h+1}}\right]$ are the same through all the intermediate nodesm m_h, since the packets eventually join the same virtual queue (to node m_{h+1}). However, the sojourn time $\widetilde{D}_{k,m_h,m_{h+1}}$ of the virtual queue will be different, since the transmission time from various intermediate nodes m_h to the same m_{h+1} are different. The first moment and the second moment of the sojourn time are:

$$E\left[\widetilde{D}_{k,m_h,m_{h+1}}\right] = E\left[\widetilde{W}_{k,m_{h+1}}\right] + E\left[\widetilde{S}_{k,m_h,m_{h+1}}\right] \quad (39)$$

$$\begin{aligned} E\left[\widetilde{D}^2_{k,m_h,m_{h+1}}\right] \approx\, & E\left[\widetilde{W}^2_{k,m_{h+1}}\right] \\ & + 2E\left[\widetilde{W}_{k,m_{h+1}}\right] E\left[\widetilde{S}_{k,m_h,m_{h+1}}\right] \\ & + E\left[\widetilde{S}^2_{k,m_h,m_{h+1}}\right] \end{aligned} \quad (40)$$

Note that Eq. (40) is obtained by ignoring the correlation of the waiting and service time. Finally, the service time of the priority M/G/1 queue at the intermediate node can be modified as (similar to Eq. (23)):

$$\begin{aligned} E\left[\widetilde{X}_{k,m_h}\right] &= \sum_{m_{h+1}=1}^{M_{h+1}} \beta_{k,h+1,m_{h+1}} E\left[\widetilde{D}_{k,m_h,m_{h+1}}\right], \\ E\left[\widetilde{X}^2_{k,m_h}\right] &= \sum_{m_{h+1}=1}^{M_{h+1}} \beta_{k,h+1,m_{h+1}} E\left[\widetilde{D}^2_{k,m_h,m_{h+1}}\right] \end{aligned} \quad (41)$$

Let $W^{(I)}_{k,m_h}$ be the waiting time for a packet of class f_k that goes through an intermediate node m_h when the interference effect is considered (see Eq. (42)). The expectation of the waiting time over the h-th hop for packets of class f_k is (as in Eq.(25)):

$$E\left[W^{(I)}_{k,h}\right] = \sum\nolimits_{m_h=1}^{M_h} \beta_{k,h,m_h} E\left[W^{(I)}_{k,m_h}\right] \quad (43)$$

The probability of packet loss of class f_k at intermediate node m_h due to the expiration now becomes Eq. (44).

VI. Convergence Discussion of the Self-Learning Algorithm

Next, we show that the self-learning routing algorithm will converge to a steady-state under certain assumptions:

Lemma: Given a set of fixed (pre-determined) outgoing relay selecting parameters $\{\beta_{k,h+1,m_{h+1}} | m_{h+1} = 1, \ldots M_{h+1}, k = 1, \ldots, K\}$, the incoming relay selecting parameters $\{\beta_{k,h,m_h} | m_h = 1, \ldots M_h, k = 1, \ldots, K\}$ will converge to a steady-state, under the assumption that the network condition is not changing over time, and given stationary statistics for the video sources.

Proof: Since all the $\beta_{k,h+1,m_{h+1}}$ are fixed and the network condition is not changing, the first two moments of the service time $E[X_{k,h,m_h}]$ and $E\left[X^2_{k,h,m_h}\right]$ remain constant over time (see Eq. (23)). Thus, the balking arrival queues converge to a steady state (see [18] for more details) having the average queue waiting times $E[W_{k,m_h}]$. In addition, the fixed $\beta_{k,h+1,m_{h+1}}$ also implies that the expected delays $E\left[Delay_{k,m_{h+1}}\right]$ from the relay m_{h+1} (in the next hop) are fixed over time for every class of traffic. Consequently, from Eq.(7), $E[Delay_{k,m_h}]$ will also converge to a steady state for every node m_h (at the current hop). This ensures that the incoming relay selecting parameters β_{k,h,m_h} will also have a steady-state, because they only depend on these $E[Delay_{k,m_h}]$ (see Eq. (8)). ■

TABLE I
The Characteristic Parameters of the Video Classes of the Two Video Sequences

Video Classes	Video 1 "Mobile" *1668 Kbps*				Video 2 "Coastguard" *1500 Kbps*			
f_k	f_1	f_4	f_6	f_8	f_2	f_3	f_5	f_7
λ_k (dB/Kbps)	0.0170	0.0064	0.0042	0.0031	0.0105	0.0064	0.0048	0.0042
R_k (Kbps)	556	333	334	445	500	300	300	400

Theorem: The self-learning policy over an H-hop directed acyclic overlay network will converge to a steady-state solution for the relay selecting parameters.

Proof: Since the relay selecting parameters β_{k,H,m_H} at the last hop are fixed according to the pre-determined destination node of each traffic class, the relay selecting parameters $\beta_{k,H-1,m_{H-1}}$ will converge in time to a steady-state according to the above *Lemma*. Then, starting from the last hop, the relay selecting parameters of the entire multi-hop infrastructure will converge sequentially to a steady-state. ■

VII. Simulation Results

In this section, two video sequences "Mobile," and "Coastguard" (16 frames per GOP at a frame rate of 30 Hz) are compressed using an embedded scalable video codec [25]. Each scalable bitstream is separated into 4 classes ($N_v = 4, K = 8$). The characteristic parameters of the video classes of the two video streams are given in Table I (see [13], [23] for more details on how to determine these parameters). In the simulation, the packet length L_k is up to 1000 bytes. No further fragmentation is performed at the lower layers (network or MAC layer). The application playback delay deadline is set to 0.533 seconds. We analyze the performance of our algorithms in terms of the received video quality (PSNR) of the various users. We compare our analytical results based on a steady-state analysis of the proposed distributed solution with the simulation results obtained using a multi-hop overlay network test-bed [10].

In our simulation, we captured the packet-loss pattern under different channel conditions (described in the paper by the link SINR) using our wireless streaming test-bed [10]. In this way, we can assess the efficiency of our system under real wireless channel conditions and link adaptation mechanisms currently deployed in state-of-the-art 802.11a/g wireless cards with 802.11e extension. Link adaptation selects one appropriate physical-layer mode (modulation and channel coding) depending on the link condition, in order to continuously maximize the experienced goodput [10]. The various efficiency levels are represented by varying the available time fraction for the contention-free period in the polling-based MAC protocol, which induces the various available transmission rates for the video packets over the links. In our elementary structure, these network efficiency levels are represented by the transmission rate multiplier T_m ranging from 0.3 Mbps to 0.6 Mbps. A larger transmission rate multiplier gives a higher network efficiency.

In the analytical results, we determine the end-to-end packet loss rate based on the average measured SINR and the average

$$E\left[W_{k,m_h}^{(I)}\right] = \frac{\sum\limits_{i=1}^{k} \eta_{i,m_h} E\left[\widetilde{X}_{i,m_h}^2\right]}{2\left(1-\sum\limits_{i=1}^{k-1}\eta_{i,m_h} E\left[\widetilde{X}_{i,m_h}\right]\right)\left(1-\sum\limits_{i=1}^{k}\eta_{i,m} E\left[\widetilde{X}_{i,m_h}\right]\right)} \tag{42}$$

$$\begin{aligned} P_{k,m_h}^{(I)} &= \text{Prob}\left(W_{k,m_h}^{(I)} > d_k - \sum_{j=0}^{h-1} E\left[W_{k,j}^{(I)}\right]\right) \\ &\approx \left(\sum_{i=1}^{k}\eta_{i,m_h} E\left[\widetilde{X}_{i,m_h}\right]\right)\exp\left(-\frac{\left(d_k - \sum\limits_{j=0}^{h-1} E\left[W_{k,j}^{(I)}\right]\right)\sum\limits_{i=1}^{k}\eta_{i,m_h} E\left[\widetilde{X}_{i,m_h}^2\right]}{E\left[W_{k,m_h}^{(I)}\right]}\right) \end{aligned} \tag{44}$$

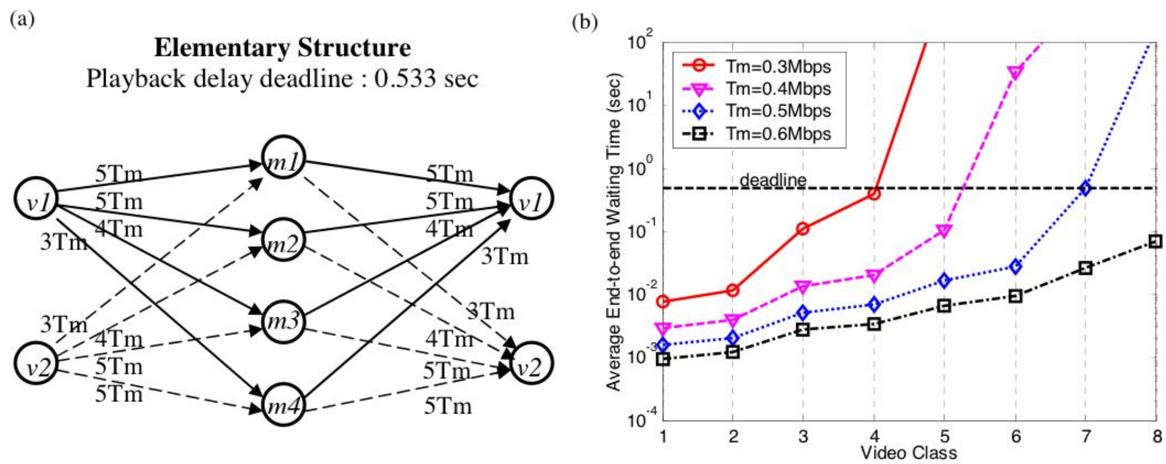

Fig. 5. (a) Network settings of the elementary structure. (b) Analytical average end-to-end waiting time of the 8 video classes.

TABLE II
ANALYTICAL AND SIMULATION RESULTS FOR UNIFORM RELAY SELECTING PARAMETERS OVER THE ELEMENTARY STRUCTURE

Elementary Structure		Video 1 "Mobile"					Video 2 "Coastguard"				
Relay selecting policy: uniform	**Tm** (Mbps)	Packet loss rate P_k f_1	f_4	f_6	f_8	Y-PSNR (dB)	Packet loss rate P_k f_2	f_3	f_5	f_7	Y-PSNR (dB)
Analytical result	**0.3**	0	8.2%	100%	100%	30.15	0	0	100%	100%	32.49
	0.4	0	0	100%	100%	30.34	0	0	0	100%	33.93
	0.5	0	0	0	100%	31.74	0	0	0	15%	35.34
	0.6	0	0	0	0	33.12	0	0	0	0	35.61
Simulation result	**0.29**	0	39%	78%	99%	29.34	4.5%	23%	69%	98%	32.26
	0.41	0	6.5%	32%	95%	31.41	0.4%	2.5%	19%	71%	34.29
	0.50	0	0.5%	9.8%	77%	32.00	0	0	3.1%	30%	35.10
	0.61	0	0.3%	2.1%	10%	33.05	0	0	0.8%	2.2%	35.59

T_m obtained for each link from the test-bed over the duration of the simulation experiments. Fig. 5 shows the elementary structure with the two video streams and four intermediate nodes. The analytical expected end-to-end delays $E[D_k]$ of the packets in the eight classes are also shown for different network efficiency levels. The dashed line represents the delay deadline. Once the end-to-end delay exceeds the delay deadline, the packets in that class are dropped. Table II shows the results of the end-to-end packet loss probability for each video class using our priority queuing approach. The almost-binary results (0 or 100%) obtained by our packet loss analysis are due to the fact that in Eq.(44), we approximate $delay_{k,m_h}^{curr}$ (current delay, see Eq. (12)) using $\sum_{j=0}^{h-1} E\left[W_{k,j}^{(I)}\right]$ instead of

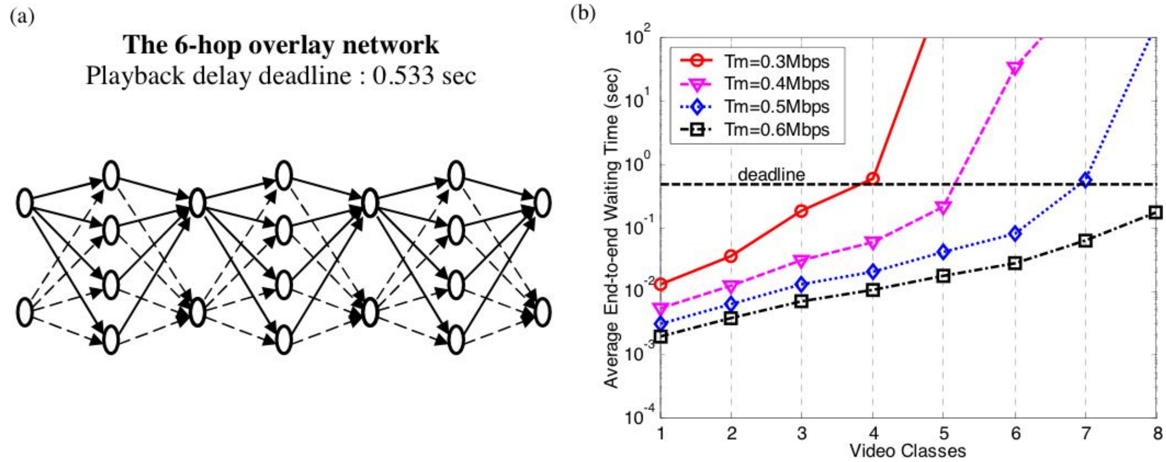

Fig. 6. (a) Network settings of the 6 hop overlay network (by cascading the elementary structure). (b) Analytical average end-to-end waiting time of the 8 video classes.

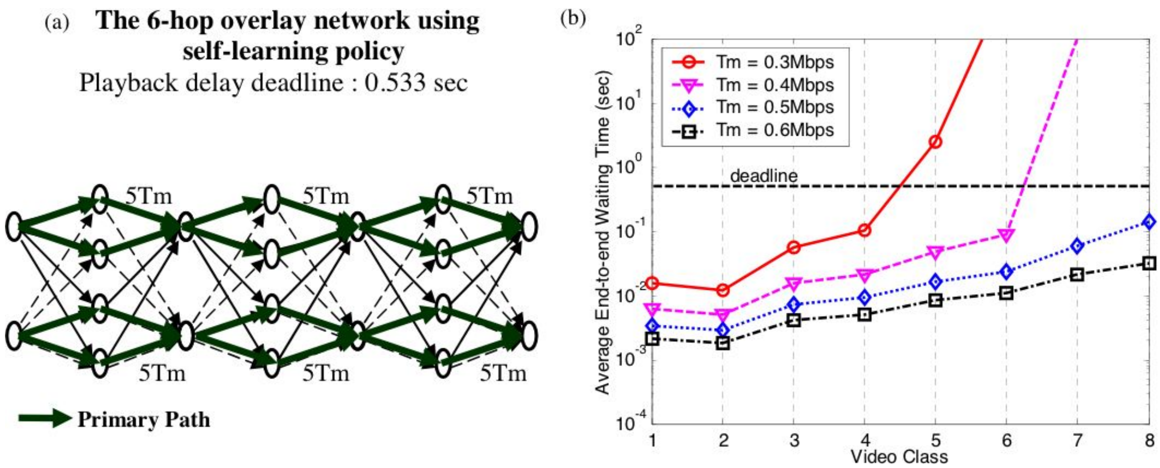

Fig. 7. (a) Primary paths of the 6-hop overlay network using self-learning policy. (b) Analytical average end-to-end waiting time of the 8 video classes.

TABLE III

ANALYTICAL AND SIMULATION RESULTS FOR UNIFORM RELAY SELECTING PARAMETERS OVER THE 6-HOP NETWORK

6-hop network		Video 1 "Mobile"					Video 2 "Coastguard"				
Relay selecting policy: uniform	**Tm** (Mbps)	Packet loss rate P_k				Y-PSNR (dB)	Packet loss rate P_k				Y-PSNR (dB)
		f_1	f_4	f_6	f_8		f_2	f_3	f_5	f_7	
Analytical result	**0.3**	0	100%	100%	100%	28.20	0	0.3%	100%	100%	32.48
	0.4	0	0	100%	100%	30.34	0	0	0.4%	100%	33.92
	0.5	0	0	0.1%	100%	31.74	0	0	0	100%	33.93
	0.6	0	0	0	1%	33.12	0	0	0	0.1%	35.61
Simulation result	**0.30**	0	75%	97%	100%	28.39	7.7%	42%	88%	100%	31.86
	0.39	0	21%	65%	99%	30.21	0.1%	11%	38%	93%	33.56
	0.51	0	3.4%	12%	92%	31.35	0	1.6%	12%	64%	33.88
	0.60	0	0	1.1%	39%	32.85	0	0	0.4%	10%	35.58

$\sum_{j=0}^{h-1} W_{k,j}^{(I)}$, i.e. we use the expected waiting time instead of the exact waiting time, as this is only known instantaneously, at each queue, during the streaming simulation. Note though that the estimations of P_k are accurate enough for the important classes, thereby leading to an accurate video quality estimation.

In Fig. 6, we consider a larger network (the 6-hop network) with the same network settings as in Fig. 5. By increasing the number of hops, both the average queue waiting time and the end-to-end packet error rate increase. Comparing the results in Table III with the results in Table II, the error between the analytical and simulation results decreases, since

TABLE IV
ANALYTICAL AND SIMULATION RESULTS FOR SELF-LEARNING POLICY RELAY SELECTING PARAMETERS (THE ANALYTICAL RESULTS ARE APPROXIMATED ACCORDING TO THE PRIMARY PATH SELECTED BY THE SELF-LEARNING POLICY)

6-hop network		Video 1 "Mobile"					Video 2 "Coastguard"				
Self-learning policy	Tm (Mbps)	Packet loss rate P_k				Y-PSNR (dB)	Packet loss rate P_k				Y-PSNR (dB)
		f_1	f_4	f_6	f_8		f_2	f_3	f_5	f_7	
Analytical result	0.3	0	0	100%	100%	30.34	0	0	100%	100%	32.49
	0.4	0	0	0	100%	31.74	0	0	0	100%	33.93
	0.5	0	0	0	0	33.12	0	0	0	0	35.61
	0.6	0	0	0	0	33.12	0	0	0	0	35.61
Simulation result	0.31	0.4%	21%	53%	83%	30.42	0	8.3%	35%	66%	33.27
	0.42	0	0	7.3%	48%	32.53	0	0	0.5%	14%	35.23
	0.50	0	0	0	3.3%	33.10	0	0	0	1.2%	35.61
	0.60	0	0	0	0	33.10	0	0	0	0	35.61

TABLE V
COMPARISON OF THE DYNAMIC SELF-LEARNING POLICY WITH THE CONVENTIONAL FIXED SINGLE-PATH AND MULTI-PATH ALGORITHMS

Method	Tm = 0.3 (Mbps) Low network efficiency		Tm = 0.6 (Mbps) Medium network efficiency	
	"Mobile" Y-PSNR (dB)	"Coastguard" Y-PSNR (dB)	"Mobile" Y-PSNR (dB)	"Coastguard" Y-PSNR (dB)
Fixed Optimal Path	24.98	30.67	31.37	34.32
Fixed Multi-path	28.39	31.86	32.85	35.58
Self-learning Policy	30.42	33.27	33.10	35.61

the assumption that the waiting time dominates the overall delay is more accurate in a larger network. The accuracy of the analysis could be further improved by separating the video into a larger number of classes.

The results of the proposed self-learning policy are shown in Table IV. Note that in Table II and Table III, we use a uniform relay selection among the intermediate nodes of each hop. The resulting primary paths are marked in bold arrows in the network plot of Fig. 7. We observe significant improvements in terms of end-to-end packet loss and video qualities using the self-learning policy. Interestingly, similarly to the Bellman-Ford algorithm, we found that this policy tries to transmit the two video streams over distinct paths in order to limit the effect of interference and congestion among the flows.

In Table V, we compare the proposed "Self-learning Policy" with a state-of-the-art routing algorithm [27] –"Fixed Optimal Path" and a multi-path routing algorithm [32] – "Fixed Multi-path." In "Fixed Optimal Path," we statically select the links for transmission such that the goodput is maximized (determined a single path per class). In "Fixed Multi-path," besides the optimal path, several loop-free link-disjoint paths are also statically selected per class. As our dynamic "Self-learning policy," the proposed priority queuing framework is also deployed for the other two algorithms using the same network settings. The simulation results show that the proposed dynamic routing approach significantly outperforms the static routing algorithms, since it provides the ability to alleviate congestion and interference.

VIII. CONCLUSIONS

In this paper, we present a novel distributed cross-layer streaming algorithm for the transmission of multiple videos over a multi-hop wireless network. The essential feature behind our approach is the priority queuing, based on which, the most important video packet is selected and transmitted at each intermediate node over the most reliable link, until it is successfully transmitted or its deadline is expired. Besides the application layer scheduling and MAC layer retransmission policy, the transmission strategy over the network includes selecting the optimal modulation and coding scheme. Importantly, our end-to-end cross-layer strategy also includes the selection of the appropriate relay nodes for multi-hop routing. We introduce a self-learning policy for dynamic routing that minimizes the end-to-end packet loss for each class of the video streams. The end-to-end packet loss probabilities are estimated given the information feedback from the nodes of the next hops. The proposed distributed algorithm is fully adaptive to changes in the network, number of users, priorities of the users.

REFERENCES

[1] W. Wei and A. Zakhor, "Multipath unicast and multicast video communication over wireless ad hoc networks," in *Proc. Int. Conf. Broadband Networks*, *Broadnets*, 2002, pp. 496-505.
[2] E. Setton, T. Yoo, X. Zhu, A. Goldsmith, and B. Girod, "Cross-layer design of ad hoc networks for real-time video streaming," *IEEE Wireless Commun. Mag.*, pp. 59-65, Aug. 2005.
[3] B. Awerbuch and T. Leighton, "Improved approximation algorithms for the multi-commodity flow problem and local competitive routing in dynamic networks," in *Proc. 26th ACM Symposium Theory Comput.*, May 1994.
[4] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *IEEE J. Select. Areas Commun.*, vol. 23. no. 1, Jan. 2005.
[5] A. Butala and L. Tong, "Cross-layer design for medium access control in CDMA ad-hoc networks," *EURASIP J. Applied Signal Processing*, Feb. 2005, pp. 129-143.
[6] Y. Wu, P. A. Chou, Q. Zhang, K. Jain, W. Zhu, and S.Y. Kung, "Network planning in wireless ad hoc networks: A cross-layer approach," *IEEE J. Select. Areas Commun.*, vol. 23, no. 1, pp. 136-150, Jan. 2005.
[7] D. S. J. De Couto, D. Aguayo, J. Bicket, and R. Morris, "A high throughput path metric for multi-hop wireless routing," in *Proc. ACM Conf. Mob. Comput. Networking*, 2003, pp. 134-146.
[8] R. Draves, J. Padhey, and B. Zill, "Routing in multi-radio, multi-hop wireless mesh networks," in *Proc. ACM Conf. Mob. Comput. Networking,* 2004, pp. 114-128.

[9] Y. Andreopoulos, N. Mastronarde, and M. van der Schaar, "Cross-layer optimized video streaming over wireless multi-hop mesh networks," *IEEE J. Select. Areas Commun.*, special issue on multi-hop wireless mesh networks, to appear.

[10] D. Krishnaswamy, and J. Vicente, "Scalable adaptive wireless networks for multimedia in the proactive enterprise," *Intel Technology J.* [Online.] Available: `http://developer.intel.com/technology/itj/2004/volume08issue04/art04\_scalingwireless/p01\_abstract.htm`

[11] J. R. Evans and E. Minieka, *Optimization Algorithms for Networks and Graphs*. New York: Marcel Dekker, 1993.

[12] D. Bertsekas and R. Gallager, *Data Networks*. Upper Saddle River, NJ: Prentice Hall, 1987.

[13] M. van der Schaar and D. S. Turaga, "Cross-layer packetization and retransmission strategies for delay-sensitive wireless multimedia transmission," *IEEE Trans. Multimedia*, Jan. 2007.

[14] D. Qiao, S. Choi, and K. G. Shin, "Goodput analysis and link adaptation for IEEE 802.11a wireless LAN," *IEEE Trans. Mobile Comput.*, vol. 1, no. 4, 2002.

[15] S. Toumpis and A. J. Goldsmith, "Capacity regions for wireless ad hoc network," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 736-748, July 2003.

[16] T. Jiang, C. K. Tham, and C. C. Ko, "An approximation for waiting time tail probabilities in multiclass systems," *IEEE Commun. Lett.*, vol. 5, no. 4, pp 175-177, Apr. 2001.

[17] J. Abate, G. L. Choudhury, and W. Whitt, "Exponential approximations for tail probabilities in queues I: Waiting times," *Operations Research*, vol. 43, no. 5, pp 885-901, 1995.

[18] L. Kleinrock, *Queuing Systems Volume I: Theory*. New York: Wiley-Interscience Publication, 1975.

[19] IEEE 802.11e/D5.0, Draft Supplement to Part 11: "Wireless medium access control (MAC) and physical layer (PHY) specifications: Medium access control (MAC) enhancements for quality of service (QoS)," June 2003.

[20] D. Krishnaswamy, "Network-assisted link adaptation with power control and channel reassignment in wireless networks," *3G Wireless Conf.*, 2002, pp. 165-170.

[21] M. Waldvogel and R. Rinaldi. "Efficient topology-aware overlay network," *ACM SIGCOMM Computer Commun. Review*, vol. 33, no. 1, pp. 101-106, Jan. 2003.

[22] J. Jannotti, "Network-layer support for overlay networks," in *Proc. IEEE Conf. Open Architectures Network Programming*, June 2002.

[23] 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. Mobile Comput.*, vol. 5, no. 6, pp. 755-768, June 2006.

[24] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Mag.*, vol. 15, no. 6, pp. 23-50, Nov. 1998.

[25] Y. Andreopoulos, A. Munteanu, J. Barbarien, M. van der Schaar, J. Cornelis, and P. Schelkens, "In-band motion compensated temporal filtering," *Signal Processing: Image Commun.,* special issue on "subband/wavelet interframe video coding," vol. 19, no. 7, pp. 653-673, Aug. 2004.

[26] L. Xiao, M. Johansson, and S. P. Boyd, "Simultaneous routing and resource allocation via dual decomposition," *IEEE Trans. Commun.*, vol. 52, no. 7, pp. 1136-1144, July 2004.

[27] D. B. Johnson, and D. A. Maltz, "Dynamic source routing in ad hoc wireless networks," chapter in *Mobile Computing*. Kluwer Academic Publishers, 1996.

[28] P. A. Chou, and Z. Miao, "Rate-distortion optimized streaming of packetized media," *IEEE Trans. Multimedia*, vol. 8, no. 2, pp. 390-404. Apr. 2006.

[29] M. Wang and M. van der Schaar, "Operational rate-distortion modeling for wavelet video coders," *IEEE Trans. Signal Processing*, vol. 54, no. 9, pp. 3505-3517, Sep. 2006.

[30] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. I. Belmont, MA: Athena Scientific, 2005.

[31] J. R. Moorman and J. W. Lockwood, "Implementation of the multiclass priority fair queuing (MPFQ) algorithm for extending quality of service in existing backbones to wireless endpoints," in *Proc. IEEE Global Telecommun. Conf.*, 1999, vol. 5, pp. 2752-2757.

[32] M. K. Marina and S. R. Das, "Ad hoc on-demand multi-path distance vector routing (AOMDV)," in *Proc. International Conf. Network Protocols*, 2001, pp. 14-23.

[33] D. Jurca and P. Frossard, "Packet selection and scheduling for multipath video streaming," *IEEE Trans. Multimedia*, vol. 9, no. 2, Apr. 2007.

Hsien-Po Shiang received M.S. degree from National Taiwan University, Taiwan in 2002, and is currently a Ph.D. candidate at the Proactive MediaNet Lab under the guidance of Prof. Mihaela van der Schaar at EE department, UCLA. His research is focusing on the cross-layer optimizations/adaptations for video streaming over wireless mesh network, and the dynamic routing based on collaborative information exchange for delay-sensitive applications. In 2006, he was an intern at Intel Corp., Folsom CA, and did research on overlay network infrastructures for wireless mesh networks.

Mihaela van der Schaar received the M.S. and Ph.D. degrees from Eindhoven University of Technology, Eindhoven, The Netherlands, in 1996 and 2001, respectively. She is currently an Assistant Professor in the Electrical Engineering Department at UCLA. She has published extensively on multimedia compression, processing, communications, networking and architectures and holds 28 granted US patents. Since 1999, she was an active participant to the ISO Motion Picture Expert Group (MPEG) standard to which she made more than 50 contributions and for which she received three ISO recognition awards. She was an Associate Editor of IEEE Transactions on Multimedia and SPIE Electronic Imaging Journal and is currently an Associate Editor of IEEE Transactions on Circuits and System for Video Technology and of IEEE Signal Processing Letters. She received the NSF CAREER Award in 2004, IBM Faculty Award in 2005, Okawa Foundation Award in 2006, and the Best Paper Award for her paper published in 2005 in the IEEE Transactions on Circuits and Systems for Video Technology.

Appendix The Self-Learning Algorithm

Initialization: Set β_{k,h,m_h} as uniform distribution at each node of each hop.
For each service interval
  For each priority class
    For hop h+1 ($0 \leq h \leq H-1$) at each node m_h
      Receive the $E\left[Delay_{k,m_{h+1}}\right]$ from all the nodes m_{h+1} at the end of this hop.
      Determine $\beta_{k,h+1,m_{h+1}}$ using Eqs. (8) and (9).
      Estimate the $E[W_{k,m_h}]$ using Eq. (42).
      Feedback to the nodes m_{h-1} of the previous hop h with $E[Delay_{k,m_h}]$ using (7).
      Send packets according to $\beta_{k,h+1,m_{h+1}}$.