QoE-Optimal Scheduling for On-Demand Video Streams over Unreliable Wireless Networks

I-Hong Hou
CESG and Department of ECE
Texas A&M University
College Station, TX 77843, USA
ihou@tamu.edu

Ping-Chun Hsieh
CESG and Department of ECE
Texas A&M University
College Station, TX 77843, USA
lleyfede@tamu.edu

ABSTRACT

Video streaming is anticipated to dominate wireless traffic in the near future. We study wireless systems where an access point delivers video streams to multiple clients over unreliable wireless channels. The performance of each client is measured by the amount of time that its video playback halts due to buffer underflow, which has been shown to have the most impact on client's perceived quality of experience (QoE).

We develop an analytical framework that jointly captures the video playback process and the unreliable and heterogeneous wireless channels. We derive the capacity region for QoE by establishing a lower bound of a weighted sum of video halt durations over all clients. We then propose a QoE-optimal policy that can achieve every point within the capacity region. Finally, we compare our policy against two commonly used policies. Both theoretical analysis and simulation results show that our policy greatly outperforms other policies.

Categories and Subject Descriptors

C.2.1 [**COMPUTER-COMMUNICATION NETWORKS**]: Network Architecture and Design—*Wireless communication*

Keywords

Wireless networks; scheduling; QoE; video streaming

1. INTRODUCTION

Many studies have predicted that on-demand video streaming will dominate wireless traffic in the near future. In on-demand video streaming, content providers, such as *Netflix* and *Youtube*, deliver packets to clients. Clients puts received packets in their buffers, and consume packets from the buffer during the video playback. When the buffer becomes empty, the video playback halts. It has been shown that video halts have the most significant impact on the quality of experience (QoE) of clients [16, 19].

In this paper, we aim to find the capacity of wireless networks for providing QoE, as well as scheduling policies that achieves the capacity. We propose an analytical model for wireless systems where an access point (AP) transmits video streams to a number of wireless clients over unreliable wireless channels. Different clients may have different channel qualities and consume packets at different rates. The performance of a client is then measured by the amount of time that it suffers from video halts.

We characterize the achievable duration of video halts for each client. Apparently, the duration of video halts is a random variable whose distribution is determined by the AP's scheduling policy. Using martingale theory, we first show that a weighted sum of the halt durations of all clients is stochastically lower-bounded by a random variable generated from a Brownian motion. This result provides a lower-bound of the capacity.

Next, we propose a simple on-line scheduling policy for providing the desirable QoE to each client. We prove that this policy is QoE-optimal in the sense that, given any vector of random variables whose weighted sum is larger than the lower-bound, the policy ensures that the halt durations of clients are no larger than the vector of random variables. In addition to being QoE-optimal, this policy also demonstrates that the lower-bound generated from the Brownian motion is indeed tight, and we are able to sharply characterize the capacity of QoE for wireless networks.

Further, we analytically study two widely-used scheduling policies, namely, the weighted random policy and the weighted round-robin policy. We derive the closed-form expression of the distribution of video halt duration of each client for both policies. We show that the performance of these policies can be much worse than our proposed policy. In particular, we show that when the total traffic load is fixed, the performance of the weighted random policy and the weighted round-robin policy deteriorates as the number of clients increases, while the performance of our policy remains the same for any number of clients.

Finally, we conduct simulations to verify our theoretical results. We implement our policy, the weighted random policy, and the weighted round-robin policy in ns-2. Simulation results suggest that our theoretical analysis is very accurate in characterizing the performance of each client. They also show that our proposed policy outperform other policies greatly under various scenarios. In particular, even when the weighted random policy and the weighted round-robin policy spend 5 times as much time to buffer incoming packets before playing videos, their video halt durations are still more than 5 times higher than the video halt duration of our policy.

MobiHoc'15, June 22–25, 2015, Hangzhou, China.

http://dx.doi.org/10.1145/2746285.2746288.

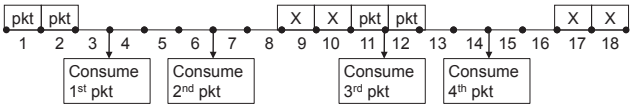


Figure 1: An example of video playback.

The rest of the paper is organized as follows. Section 2 describes our analytical model. Section 3 establishes some basic system properties, including a lower-bound for the achievable video halt durations. In Section 4, we propose a simple online scheduling policy and demonstrate that it is QoE-optimal. Section 5 shows that our analysis provides an accurate approximation for the actual system performance. In Section 6, we derive the closed-form expression of the performance under two popular policies. Section 7 demonstrates the simulation results. Section 8 summarizes some related work. Finally, Section 9 concludes the paper.

2. SYSTEM MODEL

Consider a wireless system where an access point (AP) serves a number of wireless clients, which are labeled as $n = 1, 2, \ldots, N$. We assume that time is slotted and numbered as $t = 1, 2, \ldots$. In each time slot, the AP can transmit to exactly one client.

Each wireless client is associated with an on-demand video stream. We assume that the bottleneck of a video stream is at the "last hop" between the AP and the client, which is a practical assumption for most scenarios and is used in many existing studies [14, 24]. Hence, the AP always has a packet ready for transmission to each client in each time slot. Each client has an infinite buffer so that it can buffer packets that are yet to be played. The playback process of a client is described as follows: Each client n *consumes* one packet in its buffer and plays its content every $\frac{1}{q_n}$ time slots. This also means that client n consumes packets at a rate of q_n packets per time slot. If a client is to consume a packet at time t, but its buffer is empty at that time, then we say that *outage* occurs at time t. Outage continues to occur in all subsequent time slots until the client receives a packet, at which time the client consumes the packet immediately. Outage corresponds to video halts, and the total number of slots that outage occurs is the total duration that the playback halts. We use $D_n(t)$ to denote the total number of slots that outage occurs for client n until time t.

Fig. 1 illustrates an example of the video playback process. In this example, we assume that $q_n = \frac{1}{3}$, and the client receives packets on time slots 1, 2, 11, and 12. As the client consumes one packet every three time slots, the first two packets that it receives are consumed at time 3 and 6, respectively. The client is to consume another packet at time 9. However, it does not have any packet in its buffer at that time, and hence outage occurs at time 9. Outage also occurs at time 10. At time 11, the client receives another packet, and it consumes the packet immediately and the video playback resumes. The client consumes another packet at time $14 = 11 + 3$, and is to consume one packet at time $17 = 14 + 3$. However, the client's buffer is empty at time 17, and hence outage occurs.

We now provide a more mathematical description on $D_n(t)$. Let $A_n^*(t)$ be the total number of packets that client n receives until time t, $C_n^*(t)$ be the total number of packets consumed by n until time t, and $B_n^*(t)$ be the number of packets in the buffer of n at time t. We set $A_n^*(0) = B_n^*(0) = C_n^*(0) = D_n(0) = 0$. We then have

$$C_n^*(t) = \lfloor q_n(t - D_n(t)) \rfloor, \tag{1}$$

because n consumes one packet every $1/q_n$ slots without outage. We also have

$$D_n(t) = \begin{cases} D_n(t-1) + 1, & \text{if } \lfloor q_n(t - D_n(t-1)) \rfloor > A_n^*(t) \\ D_n(t-1), & \text{otherwise,} \end{cases} \tag{2}$$

since $\lfloor q_n(t - D_n(t-1)) \rfloor$ is the number of packets that n would consume if outage does not happen at slot t. Finally, we apparently have

$$B_n^*(t) = A_n^*(t) - C_n^*(t). \tag{3}$$

The definition of $C_n^*(\cdot)$ involves floor function and is difficult to operate with. We can simplify the expression of $D_n(t)$ by defining

$$A_n(t) := A_n^*(t) + (1 - q_n), \tag{4}$$

$$C_n(t) := q_n(t - D_n(t)), \tag{5}$$

$$B_n(t) := [A_n(t) - C_n(t)]^+. \tag{6}$$

We now have $D_n(t) = D_n(t-1) + 1$ if and only if

$$\lfloor q_n(t - D_n(t-1)) \rfloor = A_n^*(t) + 1 \tag{7}$$

$$\Leftrightarrow q_n(t - D_n(t-1) - 1) \geq A_n(t) \Leftrightarrow B_n(t) = 0, \tag{8}$$

where the last step of the derivation follows from $D_n(t) \leq D_n(t-1) + 1$. In other words, $D_n(t)$ increases when $B_n(t) = 0$.

The definitions of $A_n(t), C_n(t), B_n(t)$, and $D_n(t)$ can be viewed as a system where the client consumes q_n packets in each slot without outage and has an initial buffer size of $1 - q_n$. Let $X_n(t) := A_n(t) - q_n t$. Since client n is supposed to consume $q_n t$ packets by time t, $X_n(t)$ loosely reflects whether the number of packets delivered for n matches its consumption rate. We further note that the following relations hold:

$$B_n(t) = X_n(t) + q_n D_n(t) \geq 0, \tag{9}$$

$$X_n(t) = A_n(t) - q_n t, \tag{10}$$

$$[D_n(t) - D_n(t-1)] \in \{0, 1\}, \quad D_n(0) = 0, \tag{11}$$

$$B_n(t)[D_n(t) - D_n(t-1)] = 0. \tag{12}$$

We consider the unreliable wireless transmissions by assuming that the channel reliability of client n is p_n. When the AP transmits a packet to n, the transmission is successful with probability p_n. The AP has instant knowledge on whether the transmission is successful, and can retransmit the same packet if previous transmissions fail.

We further assume that clients choose their values of q_n so that it is feasible to provide a long-term average throughput of q_n packets per time slot to each client. In other words, there exists a scheduling policy under which $\liminf_{t\to\infty} \frac{A_n(t)}{t} \geq q_n$, or, equivalently, $\limsup_{t\to\infty} \frac{D_n(t)}{t} = 0$ for all n, almost surely.

Let $U_n(t)$ be the total number of times that the AP transmits to n until time t. Hou, Borkar, and Kumar [8] has shown that the long-term average throughput of n is at least q_n if and only if $\liminf_{t\to\infty} \frac{U_n(t)}{t} \geq \frac{q_n}{p_n}$. Since the AP transmits one packet in each time slot, we have:

LEMMA 1. *It is feasible to achieve* $\limsup_{t\to\infty} \frac{D_n(t)}{t} = 0$, *for all* n, *if and only if* $\sum_n \frac{q_n}{p_n} \leq 1$.

While the study of $\limsup_{t\to\infty} \frac{D_n(t)}{t} = 0$ reveals some important insights, it cannot precisely characterize $D_n(t)$ for videos with finite length. In this paper, we study the *diffusion limit* of $D_n(t)$, defined as

$$\hat{D}_n(t) := \lim_{k\to\infty} \frac{D_n(kt)}{\sqrt{k}}, 0 \le t \le 1. \quad (13)$$

We will mainly focus on the extreme case with $\sum_n \frac{q_n}{p_n} = 1$. In Section 4.1, we further demonstrate that some results for the extreme case also characterize the performance for the general case with $\sum_n \frac{q_n}{p_n} < 1$. In Section 7, we demonstrate that $\hat{D}_n(t)$ provides a very accurate approximation of $D_n(t)$ for videos that are at least a couple of seconds long.

$\hat{D}_n(t)$ is a random process whose distribution is determined by the scheduling policy. We compare the distributions of $\hat{D}_n(t)$ under different policies through stochastic ordering as defined as follows:

DEFINITION 1. *Let $\hat{D}_n^{\eta_1}(t)$ and $\hat{D}_n^{\eta_2}(t)$ be the process $\hat{D}_n(t)$ under policies η_1 and η_2, respectively. We say that $\hat{D}_n^{\eta_1}(t) \le_{st} \hat{D}_n^{\eta_2}(t)$ if*

$$Prob(\hat{D}_n^{\eta_1}(t) \ge x) \le Prob(\hat{D}_n^{\eta_2}(t) \ge x),$$

for all x and all $t \in [0,1]$. We say that η_1 minimizes $\hat{D}_n^{\eta_1}(t)$ if $\hat{D}_n^{\eta_1}(t) \le_{st} \hat{D}_n^{\eta}(t)$ for all η. □

We also define

$$\hat{X}_n(t) := \lim_{k\to\infty} \frac{X_n(kt)}{\sqrt{k}}, \quad (14)$$

$$\hat{B}_n(t) := \lim_{k\to\infty} \frac{B_n(kt)}{\sqrt{k}}. \quad (15)$$

The stochastic ordering of $\hat{X}_n(t)$ and $\hat{B}_n(t)$ under different policies are defined similarly.

In view of (9) - (12), we have

$$\hat{B}_n(t) = \hat{X}_n(t) + q_n \hat{D}_n(t) \ge 0, \quad (16)$$

$$\frac{d\hat{D}_n(t)}{dt} \ge 0, \quad \hat{D}_n(0) = 0, \quad (17)$$

$$\hat{B}_n(t)\frac{d\hat{D}_n(t)}{dt} = 0, \quad (18)$$

where we slightly abuse notation and define

$$\frac{d\hat{D}_n(t)}{dt} := \lim_{\delta\to 0^+} \frac{\hat{D}_n(t+\delta)}{\delta}.$$

We then have the following theorem:

THEOREM 1. *[5, Theorem 6.1] Given $\hat{X}_n(t)$, there exists a unique pair of $(\hat{B}_n(t), \hat{D}_n(t))$ that satisfies (16) - (18). Further, $\hat{D}_n(t) = \sup_{0\le\tau\le t}(\max\{0, -\frac{\hat{X}_n(\tau)}{q_n}\})$.*

3. FUNDAMENTAL SYSTEM PROPERTIES

Theorem 1 shows that $\hat{D}_n(t)$ can be characterized by $\hat{X}_n(t)$. Hence, we first study the behavior of $\hat{X}_n(t)$. Let $X(t) := \sum_n \frac{X_n(t)}{p_n} = \sum_n \frac{A_n(t)}{p_n} - t$, as $\sum_n \frac{q_n}{p_n} = 1$. Suppose the AP transmits to client n in time slot $t+1$, we have $X(t+1) = X(t) + \frac{1}{p_n} - 1$ with probability p_n, and $X(t+1) = X(t) - 1$ with probability $(1-p_n)$. Hence, no matter which client is scheduled at time $t+1$, we have $E[X(t+1)|\mathcal{H}_t] = X(t)$, where $\mathcal{H}_t$ is the system history up to time t. $X(t)$ is then said to be a martingale.

Suppose, given $\mathcal{H}_t$, the employed policy schedules client n at time slot $t+1$ with probability $r_n(t+1)$. Then, $\sigma_t^2 := E[(X(t+1) - X(t))^2|\mathcal{H}_t] = \sum_n r_n(t+1) \times (\frac{1}{p_n} - 1)$. If the employed policy achieves $\limsup_{t\to\infty} \frac{D_n(t)}{t} = 0$, then $\lim_{t\to\infty} \frac{U_n(t)}{t}$ exists and equals $\frac{q_n}{p_n}$, since $\sum_n \frac{q_n}{p_n} = 1$. Hence, we have $\lim_{t\to\infty} \frac{\sum_{\tau=1}^t r_n(\tau)}{t} = \frac{q_n}{p_n}$, and

$$\sigma^2 := \lim_{t\to\infty} \frac{\sum_{\tau=1}^t \sigma_\tau^2}{t} = \sum_n \lim_{t\to\infty} \frac{\sum_{\tau=1}^t r_n(\tau)}{t}(\frac{1}{p_n} - 1)$$
$$= \sum_n \frac{q_n}{p_n}(\frac{1}{p_n} - 1). \quad (19)$$

Further, for a fixed $t \in [0,1]$,

$$\lim_{n\to\infty} \sum_{\tau=1}^{\lfloor nt \rfloor} \sigma_\tau^2 / \sum_{\tau=1}^{n} \sigma_\tau^2 = t.$$

By the functional central limit theorem for martingale [3] , we can establish that $\hat{X}(t) := \lim_{k\to\infty} \frac{X(kt)}{\sqrt{k}}$ is a driftless Brownian motion with variance σ^2. In particular, given $\hat{X}(\tau)$, for any $\tau, t \ge 0$ with $\tau + t \le 1$, $\hat{X}(\tau+t) - \hat{X}(\tau)$ is a Gaussian random variable with mean 0 and variance $\sigma^2 t$. Let $\Phi(x)$ be the cumulative distribution function (CDF) of a standard Gaussian random variable with mean 0 and unit variance. We then have

$$Prob(\hat{X}(\tau+t) - \hat{X}(\tau) \le b|\hat{X}(\tau)) = \Phi(b/\sqrt{\sigma^2 t}). \quad (20)$$

Let $\hat{D}(t) := \sup_{0\le\tau\le t}(\max\{0, -\hat{X}(\tau)\})$. Since $\hat{X}(t)$ is a Brownian motion for any policies that achieve $\lim_{t\to\infty} \frac{D_n(t)}{t} = 0$, the distribution of $\hat{D}(t)$ is the same for all such policies. We have the following two lemmas:

LEMMA 2. *The CDF of $\hat{D}(t)$ is given by*

$$Prob(\hat{D}(t) \le x) = \Phi(\frac{x}{\sqrt{\sigma^2 t}}) - \Phi(\frac{-x}{\sqrt{\sigma^2 t}}), \quad (21)$$

for all $x \ge 0, t \ge 0$. The probability density function of $\hat{D}(t)$ is $\sqrt{\frac{2}{\sigma^2 t \pi}} e^{-\frac{x^2}{2\sigma^2 t}}$.

PROOF. This is a direct result of Section 1.6 in [7]. □

LEMMA 3. *For any policy that achieves $\lim_{t\to\infty} \frac{D_n(t)}{t} = 0$, $\hat{D}(t) \le_{st} \sum_n \frac{q_n}{p_n} \hat{D}_n(t)$.*

PROOF. By definition, we have, for every sample path,

$$\hat{D}(t) = \sup_{0\le\tau\le t} (\max\{0, -\hat{X}(\tau)\})$$
$$= \sup_{0\le\tau\le t} (\max\{0, -\sum_n \frac{\hat{X}_n(\tau)}{p_n}\})$$
$$\le \sum_n \sup_{0\le\tau\le t} (\max\{0, -\frac{\hat{X}_n(\tau)}{p_n}\}) = \sum_n \frac{q_n}{p_n} \hat{D}_n(t).$$

□

We can also calculate the expected value of $\hat{D}(t)$ by

$$E[\hat{D}(t)] = \int_0^\infty x\sqrt{\frac{2}{\sigma^2 t\pi}} e^{-\frac{x^2}{2\sigma^2 t}} = \sqrt{\frac{2t}{\pi}\sum_n \frac{q_n}{p_n}(\frac{1}{p_n} - 1)}. \quad (22)$$

Motivated by Lemma 3, we define the *capacity region for QoE* of the system.

DEFINITION 2. *A vector $\delta = [\delta_1, \delta_2, \ldots, \delta_N]$ is said to be* feasible *if there exists a scheduling policy so that $\hat{D}_n(t) \leq_{st} \delta_n \hat{D}(t)/\frac{q_n}{p_n}$, for all n. The* capacity region for QoE *is defined as the set of all feasible δ.* □

The following necessary condition for feasibility is then a direct result of Lemma 3.

THEOREM 2. *A feasible vector δ must have $\delta_n \geq 0, \forall n$, and $\sum_n \delta_n \geq 1$.* □

4. A QOE-OPTIMAL ONLINE POLICY

In this section, we introduce a simple online policy, namely, the *prioritized largest deficit policy*. We prove that the prioritized largest deficit policy is QoE-optimal in the sense that it can achieve every point in the capacity region for QoE.

The prioritized largest deficit policy divides clients into two groups, a *high-priority* group, denoted by S_1, and a *low-priority* group, denoted by S_2, with $S_2 \neq \phi$. In each time slot $t+1$, the AP calculates $\lfloor -X_n(t) \rfloor$ for all $n \in S_1$. The AP schedules an arbitrary client $n \in S_1$ with $\lfloor -X_n(t) \rfloor > 0$. If $\lfloor -X_n(t) \rfloor \leq 0$, for all $n \in S_1$, the AP schedules the client $m \in S_2$ with the largest value of $w_m(q_m t - A_m(t)) = w_m(-X_m(t))$, where w_m is a predetermined value with $w_m > 0$, for all $m \in S_2$.

In practice, the AP may choose the high-priority group and the weights of clients by its own pricing/service policy. For example, the AP may charge clients a premium to be put in the high-priority group, or to have larger w_n. We note that, in the special case where all clients are in the low-priority group, the prioritized largest deficit policy becomes similar to the max weight policy [21, 22] and the largest debt first policy [8], as it schedules the client with the largest value of $w_m(q_m t - A_m(t))$ in each time slot.

4.1 Performance of the High-Priority Group

We first study the performance of clients in the high-priority group. Define

$$Q(t) := \sum_{n \in S_1} \lfloor -X_n(t) \rfloor / p_n = Y(t) + Z(t),$$

where $Y(t)$ and $Z(t)$ are defined by

$$Y(t) := Q(t) + \sum_{n \in S_1} U_n(t) - t,$$
$$Z(t) := t - \sum_{n \in S_1} U_n(t).$$

Since $X_n(t+1) < X_n(t)$ unless n is scheduled in time slot $t+1$, and $|X_n(t+1) - X_n(t)| < 1$, for all t, the design of the prioritized largest deficit policy ensures that $(-X_n(t)) \geq 0$, for all $n \in S_1$ and all t. Hence, $Q(t) = Y(t) + Z(t) \geq 0$. Further, $Z(t)$ is non-decreasing and only increases when the AP schedules a client in S_1, which is equivalent to $Q(t) > 0$. Therefore, we have $Z(t) - Z(t-1) \geq 0$ and $Q(t)[Z(t) - Z(t-1)] = 0$. By a similar argument for Theorem 1, we can establish that

$$Z(t) = \sup_{0 \leq \tau \leq t} \max\{0, -Y(\tau)\}. \quad (23)$$

Next, we study the behavior of $Y(t)$. Let $\epsilon(t) := Q(t) + \sum_{n \in S_1} \frac{X_n(t)}{p_n}$. Then, $|\epsilon(t)| \leq \sum_{n \in S_1} \frac{1}{p_n}$. We can now rewrite $Y(t)$ as

$$Y(t) = \sum_{n \in S_1} (U_n(t) - \frac{A_n(t)}{p_n}) + (\sum_{n \in S_1} \frac{q_n}{p_n} - 1)t + \epsilon(t). \quad (24)$$

If a client n is scheduled at time slot $t+1$, we have,

$$U_n(t+1) - \frac{A_n(t+1)}{p_n}$$
$$= \begin{cases} (U_n(t) - \frac{A_n(t)}{p_n}) + 1 - \frac{1}{p_n}, & \text{with probability } p_n, \\ (U_n(t) - \frac{A_n(t)}{p_n}) + 1, & \text{with probability } 1 - p_n. \end{cases}$$

On the other hand, if n is not scheduled at $t+1$, we have $(U_n(t+1) - \frac{A_n(t+1)}{p_n}) = (U_n(t) - \frac{A_n(t)}{p_n})$. In either case, $E[U_n(t+1) - \frac{A_n(t+1)}{p_n} | \mathcal{H}_t] = U_n(t) - \frac{A_n(t)}{p_n}$, for all n, and $\sum_{n \in S_1} (U_n(t) - \frac{A_n(t)}{p_n})$ is a martingale. By the martingale law of large numbers [6], $\lim_{t \to \infty} \frac{\sum_{n \in S_1} (U_n(t) - A_n(t)/p_n)}{t} = 0$, and hence $\lim_{t \to \infty} \frac{Y(t)}{t} = \sum_{n \in S_1} \frac{q_n}{p_n} - 1$. Since S_2 is not empty, $\sum_{n \in S_1} \frac{q_n}{p_n} < 1$. Hence, by (23), $\lim_{t \to \infty} \frac{Z(t)}{t} = 1 - \sum_{n \in S_1} \frac{q_n}{p_n}$, and $\lim_{t \to \infty} \frac{Q(t)}{t} = 0$. The latter further implies that $\lim_{t \to \infty} \frac{X_n(t)}{t} = 0$ and $\lim_{t \to \infty} \frac{U_n(t)}{t} = \frac{q_n}{p_n}$, for all $n \in S_1$.

Next, we study the diffusion limit of $Q(t)$, defined as

$$\hat{Q}(t) := \lim_{k \to \infty} \frac{Q(kt)}{\sqrt{k}}$$
$$= \lim_{k \to \infty} \frac{\sum_{n \in S_1} (U_n(kt) - A_n(kt)/p_n)}{\sqrt{k}} + \frac{(\sum_{n \in S_1} \frac{q_n}{p_n} - 1)kt + \epsilon(kt)}{\sqrt{k}} + \frac{Z(kt)}{\sqrt{k}}, 0 \leq t \leq 1.$$

Since $\lim_{t \to \infty} \frac{U_n(t)}{t} = \frac{q_n}{p_n}$, for all $n \in S_1$, by the functional central limit theorem for martingale,

$$\lim_{k \to \infty} \frac{\sum_{n \in S_1} (U_n(kt) - A_n(kt)/p_n)}{\sqrt{k}}$$

is a Brownian motion. Further, since $\sum_{n \in S_1} \frac{q_n}{p_n} < 1$,

$$\frac{(\sum_{n \in S_1} \frac{q_n}{p_n} - 1)kt + \epsilon(kt)}{\sqrt{k}} \to -\infty,$$

as $k \to \infty$. Therefore, the following lemma shows that $\hat{Q}(t) = 0$. Since $X_n(t) \geq 0$, for all $n \in S_1$ by the design of the prioritized largest deficit policy, we also have $\hat{X}_n(t) = 0$, for all $n \in S_1$.

LEMMA 4. *[5, Lemma 6.4] Let $\{\theta_k\}$ be a sequence of real numbers and $\{x_k(t)\}$ be a sequence of functions defined on $t \in [0, 1]$ with $x_k(0) \geq 0$. Let $z_k(t) = x_k(t) + \theta_k + y_k(t)$, and $y_k(t) = \sup_{0 \leq \tau \leq t} \max\{0, -x_k(\tau) - \theta_k \tau\}$. Suppose that as $k \to \infty$, x_k converges to a continuous function x and $\theta_k \to -\infty$. Then $z_k(t) \to 0$.* □

We summarize the performance of clients in the high-priority set in the following theorem:

THEOREM 3. *Under the prioritized largest deficit policy, for all $n \in S_1$, $\lim_{t \to \infty} \frac{U_n(t)}{t} = \frac{q_n}{p_n}$, $\hat{X}_n(t) = 0$, and $\hat{D}_n(t) = 0$.* □

We also note that the primary condition needed to establish Theorem 3 is $\sum_{n\in S_1}\frac{q_n}{p_n}<1$. Theorem 3 can therefore be used to establish the performance of systems with $\sum_n \frac{q_n}{p_n}<1$.

COROLLARY 1. *For any system with $\sum_{n=1}^N \frac{q_n}{p_n}<1$, it is feasible to achieve $\hat{D}_n(t)=0$, for all n.*

PROOF. For a system with $\sum_{n=1}^N \frac{q_n}{p_n}<1$, we can create a pseudo client $N+1$ and choose p_{N+1}, q_{N+1} so that $\sum_{n=1}^{N+1}\frac{q_n}{p_n}=1$. We then employ the prioritized largest deficit policy by choosing $S_1=\{1,2,\ldots,N\}$ and $S_2=\{N+1\}$. By Theorem 3, we have $\hat{D}_n(t)=0$, for all $1\le n\le N$. □

4.2 Performance of the Low-Priority Group

Next, we study the performance of clients in the low-priority group S_2. We first demonstrate that $[\hat{X}_n(t)|n\in S_2]$ has the property of *state-space collapse* as described in the following:

THEOREM 4. *Under the prioritized largest deficit policy, for all $n,m\in S_2$, we have $w_n\hat{X}_n(t)=w_m\hat{X}_m(t)$, for all $t\in[0,1]$.*

PROOF. Let

$$p'_n := p_n(1-\sum_{m\in S_1}\frac{q_m}{p_m}), \tag{25}$$

$$V_n(t) := (-w_nX_n(t)) + (\sum_{m\in S_2}\frac{X_m(t)}{p'_m})/\sum_{m\in S_2}\frac{1}{w_mp'_m}. \tag{26}$$

The client with the smallest $w_nX_n(t)$ is then the one with the largest $V_n(t)$. We note that $(\sum_{m\in S_2}\frac{X_m(t)}{p'_m})/\sum_{m\in S_2}\frac{1}{w_mp'_m}$ is a weighted average of $w_nX_n(t)$. Therefore, $\max_{n\in S_2}V_n(t)\ge 0$, with equality holds if and only if $w_nX_n(t)=w_mX_m(t)$, for all $n,m\in S_2$.

We consider the fluid limit of the system and define $\bar{X}_n(t)=\lim_{k\to\infty}\frac{X_n(kt)}{k}=\lim_{k\to\infty}\frac{A_n(kt)}{k}-q_nt$, $\bar{V}_n(t)=\lim_{k\to\infty}\frac{V_n(kt)}{k}$. Without loss of generality, we assume that client 1 has the largest $\bar{V}_n(t)$ at time t in the fluid system. By Theorem 3, the fluid system spends a fraction $\sum_{m\in S_1}\frac{q_m}{p_m}$ of time serving clients in S_1, and therefore only spends a fraction $1-\sum_{m\in S_1}\frac{q_m}{p_m}$ of time serving client 1. We now have $\frac{d\bar{X}_1(t)}{dt}=p_1(1-\sum_{m\in S_1}\frac{q_m}{p_m})-q_1=p'_1-q_1$, and $\frac{d\bar{X}_n(t)}{dt}=-q_n, \forall n\neq 1, n\in S_2$. Therefore,

$$\begin{aligned}\frac{d\bar{V}_1(t)}{dt}&=-w_1p'_1+w_1q_1+\frac{\sum_{m\in S_2}\frac{q_m}{p'_m}-1}{\sum_{m\in S_2}1/w_mp'_m}\\&=-w_1p'_1+w_1q_1,\\\frac{d\bar{V}_n(t)}{dt}&=w_nq_n+\frac{\sum_{m\in S_2}\frac{q_m}{p'_m}-1}{\sum_{m\in S_2}1/w_mp'_m}\\&=w_nq_n.\end{aligned}$$

Define a Lyapunov function $L(t)=\sum_{n\in S_2}\frac{1}{2w_np'_n}\bar{V}_n(t)^2$. The Lyapunov drift is then

$$\begin{aligned}\frac{dL(t)}{dt}&=\sum_{n\in S_2}\frac{1}{w_np'_n}\bar{V}_n(t)\frac{d\bar{V}_n(t)}{dt}\\&=\sum_{n\in S_2}\frac{q_n}{p'_n}\bar{V}_n(t)-\bar{V}_1(t)\le 0,\end{aligned}$$

where the last inequality follows because $\sum_{n\in S_2}\frac{q_n}{p'_n}=1$, and $\bar{V}_1(t)\ge\bar{V}_n(t)$, for all $n\in S_2$.

Now, recall that $(\sum_{m\in S_2}\frac{X_m(t)}{p'_m})/\sum_{m\in S_2}\frac{1}{w_mp'_m}$ is a weighted average of $w_nX_n(t)$. Therefore, if $L(t)>0$, $w_1\bar{X}_1(t)<w_n\bar{X}_n(t)$, and $\bar{V}_1(t)>\bar{V}_n(t)$ for some $n\in S_2$. In this case, we have $\frac{dL(t)}{dt}<0$, and the fluid system is strongly stable. Therefore, the process $[V_n(t)]$ is positive recurrent, and we have $\lim_{k\to\infty}\frac{V_n(kt)}{\sqrt{k}}\to 0$. This establishes that $w_n\hat{X}_n(t)=(\sum_{m\in S_2}\frac{\hat{X}_m(t)}{p'_m})/\sum_{m\in S_2}\frac{1}{w_mp'_m}$, for all $n\in S_2$. □

Theorem 3 has already established that $\hat{X}(t)=\sum_n\frac{\hat{X}_n(t)}{p_n}=\sum_{n\in S_2}\frac{\hat{X}_n(t)}{p_n}$. By Theorem 4, we further have

$$\hat{X}_n(t)=\frac{1/w_n}{\sum_{m\in S_2}1/(p_mw_m)}\hat{X}(t).$$

Therefore,

$$\begin{aligned}\hat{D}_n(t)&=\sup_{0\le\tau\le t}\max\{0,-\frac{\hat{X}_n(\tau)}{q_n}\}\\&=\frac{1/(p_nw_n)}{\sum_{m\in S_2}1/(p_mw_m)}\hat{D}(t)/\frac{q_n}{p_n}.\end{aligned}\tag{27}$$

The performance of the prioritized largest deficit policy can then be summarized as below:

THEOREM 5. *Under the prioritized largest deficit policy, $\hat{D}_n(t)=0$ for all $n\in S_1$, and $\hat{D}_n(t)$ is given by (27), for all $n\in S_2$.* □

By (22), we can derive the mean of $\hat{D}_n(t)$ as

$$E[\hat{D}_n(t)]=\sqrt{\frac{2t}{\pi}\sum_{m=1}^N\frac{q_m}{p_m}(\frac{1}{p_m}-1)}\frac{1/q_nw_n}{\sum_{m\in S_2}1/(p_mw_m)},\tag{28}$$

for all $n\in S_2$.

4.3 Achieving Every Point in the Capacity Region

Based on Theorem 5, we can design an algorithm for the prioritized largest deficit policy to achieve every point in the capacity region for QoE. Assume that a vector $\delta=[\delta_n]$ with $\delta_n\ge 0$, for all n, and $\sum_n\delta_n\ge 1$ is given. We choose $S_1=\{n|\delta_n=0\}$, and $w_n=\frac{1}{p_n\delta_n}$ for all $n\notin S_1$. With this setting, we have $\hat{D}_n(t)=0=\delta_n\hat{D}(t)$, for all $n\in S_1$, and $\hat{D}_n(t)=\frac{\delta_n}{\sum_{m\in S_2}\delta_m}\hat{D}(t)/\frac{q_n}{p_n}\le_{st}\delta_n\hat{D}(t)/\frac{q_n}{p_n}$, for all $n\in S_2$. In addition, Theorem 5 and Theorem 2 provide a sharp description of the capacity region for QoE.

THEOREM 6. *The capacity region for QoE is the set of vectors δ with $\delta_n\ge 0$, for all n, and $\sum_n\delta_n\ge 1$.* □

5. STRONG APPROXIMATION OF THE DIFFUSION LIMIT

We focus on diffusion limit in this paper because it provides a much more precise characterization of $D_n(t)$ than the study of fluid limit, i.e., $\lim_{t\to\infty}\frac{D_n(t)}{t}$. Indeed, we can simply approximate $D_n(t)$ by $\hat{D}_n(t)$. In this section, we establish a strong approximation theorem for $\hat{D}(t)$. Our analysis is based on the following theorem:

THEOREM 7. *[20, Theorem 4.4] Let $W(t)$ be a martingale with $E[(W(t+1)-W(t))^2|\mathcal{H}_t]<\infty$ for all t. Define $G(t)=$*

$\sum_{i\le t} E[(W(i+1)-W(i))^2|\mathcal{H}_i]$. *If* $G(t)\to\infty$ *almost surely as* $t\to\infty$. *Also, there is a nondecreasing function* $f(t)$ *such that* $\frac{f(t)}{t}$ *is nonincreasing and*

$$\sum_{t\ge 0} f(G(t))^{-1}\int_{\mathcal{B}} x^2 dProb(W(t+1)-W(t)\le x|\mathcal{H}_t)<\infty, \tag{29}$$

where $\mathcal{B}=\{x: x^2 > f(G(t))\}$, *then there exits a Brownian motion* $\hat{W}(t)$ *such that*

$$\sup_{i\le t}|W(i)-\hat{W}(i)| = o((tf(t))^{1/4}\log t), \tag{30}$$

almost surely. □

We now state and prove our strong approximation theorem for $\hat{D}(t)$:

THEOREM 8. *For any* $0<\varepsilon<\frac{1}{4}$, *there exists a random process* $D^*(t)$ *that has the same distribution as* $\hat{D}(t)$, *and*

$$|D(t)-D^*(t)| = o(t^{\frac{1}{4}+\varepsilon}), \tag{31}$$

almost surely.

PROOF. We first find a Brownian motion $X^*(t)$ such that $\sup_{i\le t}|X(i)-X^*(i)| = o(t^{\frac{1}{4}+\varepsilon})$. We choose

$$f(t) = [\frac{t^\varepsilon}{\log(\max\{t, e^{\frac{1}{\varepsilon}}\})}]^4,$$

which is a nondecreasing function, and $\frac{f(t)}{t}$ is nonincreasing. In Section 3 we already show that

$$\min_n \frac{1}{p_n}-1 \le E[(X(t+1)-X(t))^2|\mathcal{H}_t] \le \max_n \frac{1}{p_n}-1, \tag{32}$$

and $X(t+1)-X(t)\le \max_n \frac{1}{p_n}-1$, for all t and on all sample paths. Therefore, there exists a sufficiently large T such that $Prob(X(t+1)-X(t)\le x)=1$, for all x such that $x^2 > f(G(t))$, for all $t>T$ and on every sample path. The condition (29) then holds. By Theorem 7, there exists a Brownian motion $X^*(t)$ such that $\sup_{i\le t}|X(i)-X^*(i)| = o(t^{\frac{1}{4}+\varepsilon})$, almost surely. Further, since $\hat{X}(t) := \lim_{k\to\infty}\frac{X(kt)}{\sqrt{t}}$ is also a Brownian motion, $X^*(t)$ must have the same distribution as $\hat{X}(t)$, and $D^*(t) := \sup_{0\le\tau\le t}(\max\{0,-X^*(\tau)\})$ has the same distribution as $\hat{D}(t)$.

Finally, we have

$$\begin{aligned}&|D(t)-D^*(t)|\\ =&|\sup_{0\le\tau\le t}(\max\{0,-X(\tau)\}) - \sup_{0\le\tau\le t}(\max\{0,-X^*(\tau)\})|\\ \le&\sup_{0\le\tau\le t}|X(\tau)-X^*(\tau)| = o(t^{\frac{1}{4}+\varepsilon}),\end{aligned}$$

almost surely. □

In Section 7, we will demonstrate that $\hat{D}_n(t)$ provides a very accurate estimate for videos that are at least a couple of seconds long through simulations.

6. PERFORMANCE OF POPULAR POLICIES

In this section, we discuss the performance of two popular policies, namely, the weighted random policy and the weighted round robin policy.

In the weighted random policy, the AP randomly picks a client and transmits a packet to it. The probability that a client n is chosen by the AP is $\frac{q_n}{p_n}$. Therefore, for a fixed client n, it receives a packet with probability q_n in each time slot. We have

$$X_n(t+1)-X_n(t) = \begin{cases} 1-q_n, & \text{with probability } q_n,\\ -q_n, & \text{with probability } 1-q_n.\end{cases}$$

$(X_n(t+1)-X_n(t))$ then has mean 0 and variance $q_n(1-q_n)$. By the functional central limit theorem, $\hat{X}_n(t) = \lim_{k\to\infty}\frac{X(kt)}{\sqrt{k}}$ is a driftless Brownian motion with variance $q_n(1-q_n)$. Since $\hat{D}_n(t) = \sup_{0\le\tau\le t}(\max\{0, \frac{-\hat{X}_n(\tau)}{q_n}\})$, we can derive the CDF of $\hat{D}_n(t)$ as follows:

THEOREM 9. *Under the weighted random policy, the CDF of* $\hat{D}_n(t)$ *is given by*

$$Prob(\hat{D}_n(t)\le x) = \Phi(\frac{xq_n}{\sqrt{q_n(1-q_n)t}}) - \Phi(\frac{-xq_n}{\sqrt{q_n(1-q_n)t}}), \tag{33}$$

for all $x\ge 0$, $t\ge 0$. *The expected value is*

$$E[\hat{D}_n(t)] = \sqrt{\frac{2t(1-q_n)}{\pi q_n}}. \tag{34}$$

□

Next, we study the weighted round robin policy. In the weighted round robin policy, the AP group time slots into *frames*, where each frame consists of the T time slots $\{jT+1, jT+2, \ldots, (j+1)T\}$, for some positive integer j. The value of T is chosen as the least number such that $T\frac{q_n}{p_n}$ is an integer for each n. In each frame, the AP transmits to client n exactly $T\frac{q_n}{p_n}$ times.

Under the weighted round robin policy, the number of packets that client n receives in a frame is a binomial random variable. In particular, we have $X_n((j+1)T)-X_n(jT) = i-Tq_n$ with probability $\binom{Tq_n/p_n}{i}p_n^i(1-p_n)^{Tq_n/p_n-i}$. $X_n((j+1)T)-X_n(jT)$ then has mean 0 and variance $Tq_n(1-p_n)$. By the functional central limit theorem, for a fixed integers i and j, $lim_{k\to\infty}\frac{X_n(k(j+i)T)-X_n(kiT)}{\sqrt{k}}$ is a Gaussian random variable with mean zero and variance $Tq_n(1-p_n)j$. Further, for all $t\in[jT,(j+1)T]$, $|X_n(t)-X_n(jT)|\le T$. Therefore, $\hat{X}_n(t)$ is a Brownian motion with mean 0 and variance $q_n(1-p_n)$. We then have

THEOREM 10. *Under the weighted round robin policy, the CDF of* $\hat{D}_n(t)$ *is given by*

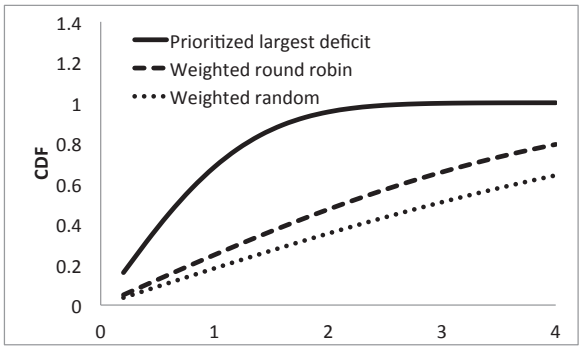
$$Prob(\hat{D}_n(t)\le x) = \Phi(\frac{xq_n}{\sqrt{q_n(1-p_n)t}}) - \Phi(\frac{-xq_n}{\sqrt{q_n(1-p_n)t}}), \tag{35}$$

for all $x\ge 0$, $t\ge 0$. *The expected value is*

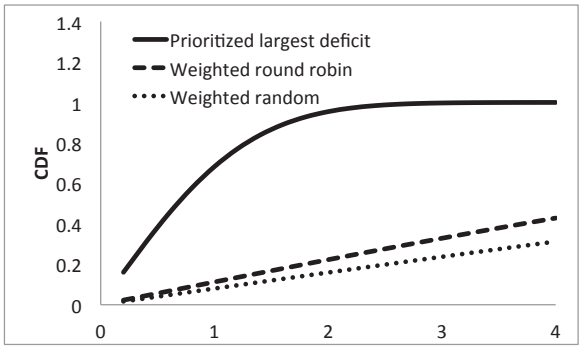
$$E[\hat{D}_n(t)] = \sqrt{\frac{2t(1-p_n)}{\pi q_n}}. \tag{36}$$

□

In order to compare the performance of the weighted random policy and the weighted round robin policy against that of the prioritized largest deficit policy, we consider the symmetric case that all clients have the same channel reliability and they consume packets at the same rate. In particular, we



(a) $N = 10$



(b) $N = 50$

Figure 2: Comparisons of the three policies in a symmetric system.

set $p_n \equiv p$ and $q_n \equiv \frac{p}{N}$. We also set $S_1 = \phi$ and $\delta_n \equiv \frac{1}{N}$ for the prioritized largest deficit policy. By (19), (21), and (22), we have, under the prioritized largest deficit policy,

$$Prob(\hat{D}_n(t) \leq x) = \Phi(\frac{xp}{\sqrt{p(1-p)t}}) - \Phi(\frac{-xp}{\sqrt{p(1-p)t}}). \tag{37}$$

and

$$E[\hat{D}_n(t)] = \sqrt{\frac{2t(1-p)}{\pi p}}. \tag{38}$$

Since $q_n < p$, it is easy to see that

$$\hat{D}_n^{PLD}(t) \leq_{st} \hat{D}_n^{WRR}(t) \leq_{st} \hat{D}^{WR}(t),$$

where $\hat{D}_n^{PLD}(t), \hat{D}_n^{WRR}(t)$, and $\hat{D}^{WR}(t)$ denotes $\hat{D}_n(t)$ under the prioritized largest deficit policy, the weighted round robin policy, and the weighted random policy, respectively.

To illustrate the difference between the three policies, we plot the CDF of $\hat{D}_n(1)$ under the three policies for systems with 10 clients and 50 clients in Figure 2. We set $p = 0.5$ in both cases. Clearly, the prioritized largest deficit policy achieves much smaller $\hat{D}_n(1)$ than the other policies. In the system with 10 clients, $Prob(\hat{D}_n^{PLD}(1) \leq 2) = 0.95$, which is more than twice as large as $Prob(\hat{D}_n^{WRR}(1) \leq 2) = 0.47$ and $Prob(\hat{D}_n^{WR}(1) \leq 2) = 0.35$. The difference becomes larger as N increases. With 50 clients, $Prob(\hat{D}_n^{WRR}(1) \leq 2) = 0.22$, $Prob(\hat{D}_n^{WR}(1) \leq 2) = 0.16$, while the distribution of $\hat{D}_n^{PLD}(1)$ remains unchanged.

7. SIMULATION RESULTS

We have implemented our prioritized largest deficit policy, as well as the weighted round-robin and the weighted random policies in ns-2. In this section, we present our simulation results. In all simulations, we consider an IEEE 802.11a system where the AP transmits data at 54 Mbps. The time needed to transmit a packet and to receive an ACK is about 660 μs. All results presented in this section are the average of 50 simulation runs.

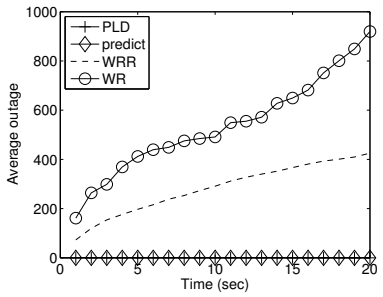
We first consider a system with 30 clients and $\sum_n \frac{q_n}{p_n} = 1$. We divide the clients into three groups: 10 clients are in the high-priority group, 10 clients are in the low-priority group with $w_n = 2$, and the other 10 clients are in the low-priority group with $w_n = 1$. Each group is further divided into two subgroups where clients in the first subgroup have $p_n = 0.75$ and $q_n = \frac{1}{40}$, while clients in the other subgroup have $p_n = 0.5$ and $q_n = \frac{1}{60}$. We simulate this system under different policies and compute the average outage of each client in each of the six subgroups.

Fig. 3 demonstrates the simulation results, where we use "PLD," "WRR," and "WR" to denote the prioritized largest deficit policy, the weighted round robin policy, and the weighted random policy, respectively. We also plot the estimate of $D_n(t)$ using $D_n(t) \approx \hat{D}_n(t)$, which is denoted by "predict." As can be seen in the figure, diffusion limit provides a very accurate estimate of $D_n(t)$ for each subgroup even for a small duration of time. This result suggests that $\hat{D}_n(t)$ is suitable for estimating the performance of on-demand video streams. In contrast, since $\lim_{t\to\infty} \frac{D_n(t)}{t} = 0$, for all n and for all simulated policies, it is clear that $\lim_{t\to\infty} \frac{D_n(t)}{t} = 0$ fails to capture important characteristics of videos with finite lengths. Further, we can also observe that the prioritized largest deficit policy outperforms the other two policies by large margins in all subgroups. Even for the low priority group with $w_n = 1$, which has the highest outage under our policy, the outage under our policy is still less than half of that under the other policies, as shown in Figs. 3(e) and 3(f).

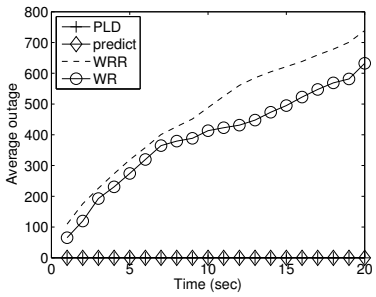
Next, we consider a system that is slightly under-loaded, that is, $\sum_n \frac{q_n}{p_n} < 1$. The system setting is very similar to that in the previous paragraphs. The only difference is that we now set $p_n = 0.75$, and $q_n = \frac{1}{40}$ for half of the clients, and $p_n = 0.5$, $q_n = \frac{1}{65}$ for the other half. Under this setting, we now have $\sum_n \frac{q_n}{p_n} = \frac{25}{26}$.

Simulation results are shown in Fig. 4. Clearly, the prioritized largest deficit policy still outperforms the other two policies. Even for the two subgroups in the low priority group and have $w_n = 1$, their average amounts of outage within 20 seconds are only 25 and 23 time slots. Since wireless systems rarely operate on the boundary of the capacity regions, the simulation results show that our policy is still preferable for practical systems. We notice that the weighted round robin policy is in general better than the weighted random policy. This is consistent with the analysis in Section 6.

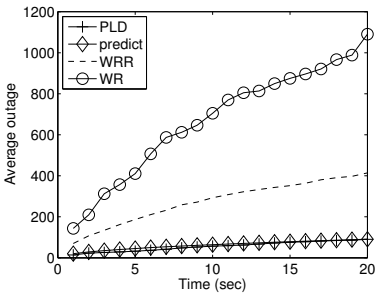
Finally, we evaluate the scenario where clients may buffer incoming packets and only start the playback when the number of packets in the buffer exceed a certain threshold. This is a popular approach to mitigate uncertainties of networks



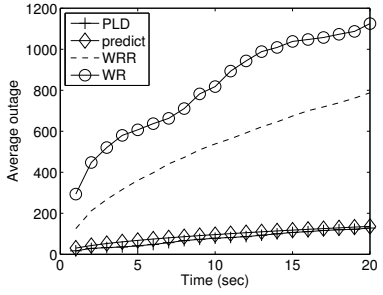
(a) High priority, $p_n = 0.75$, $q_n = 1/40$



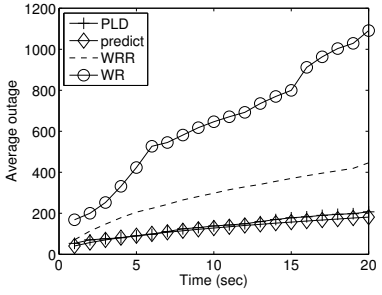
(b) High priority, $p_n = 0.5$, $q_n = 1/60$



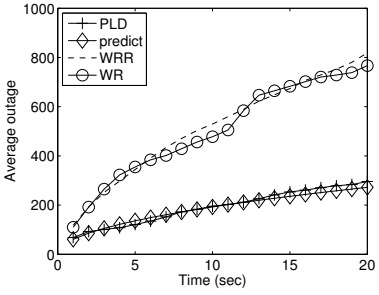
(c) Low priority, $p_n = 0.75$, $q_n = 1/40$, $w_n = 2$



(d) Low priority, $p_n = 0.5$, $q_n = 1/60$, $w_n = 2$



(e) Low priority, $p_n = 0.75$, $q_n = 1/40$, $w_n = 1$



(f) Low priority, $p_n = 0.5$, $q_n = 1/60$, $w_n = 1$

Figure 3: Average outage for a system with $\sum_n \frac{q_n}{p_n} = 1$.

[15, 17, 23]. We consider the same system as that for Fig. 3. However, we assume that, when the system starts, each client has a certain number of packets in its buffer. We then simulate the per-client average outage in five minutes for different initial buffer sizes. To make the comparison fair, the prioritized largest deficit policy sets $S_1 = \phi$ and $\delta_n \equiv \frac{1}{N}$. In addition, we also compute the portion of clients that do not suffer from any outage in five minutes time.

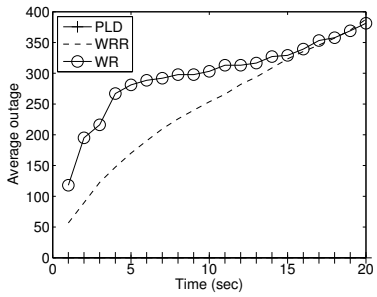
Simulation results are shown in Fig 5. It is shown that our policy outperforms the other two greatly. With an initial buffer size as small as 25 packets, which corresponds to less than one second's video content, 96% of the clients do not encounter any outage through the playback under our policy. On the other hand, even when the weighted round-robin policy starts with an initial buffer size of 125 packets, its average video halt duration is still more than 5 times larger than that under our policy with an initial buffer size of 25 packets.

8. RELATED WORK

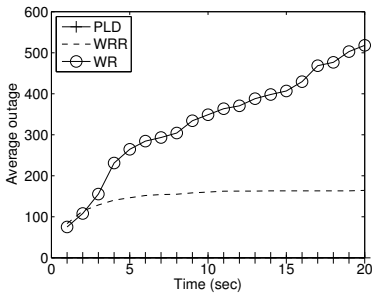
Many recent studies on video streaming have shifted emphasis from traditional quality of service (QoS) metrics, such as average delay and delay jitter, to QoE metrics that provide more direct measure on the perceived quality of end users. Video streaming needs to meet the strict deadline of each packet. Some studies model QoE by assuming that deadline misses cause packet losses and glitches in playback. Hou et al. [8], Jaramillo et al. [9], Li and Eryilmaz [13], Kim et al. [12] and Kang et al. [11] study the scheduling problem when each user has its requirements of per-packet deadline and deadline miss ratios. Li et al. [14] consider broadcasting an on-demand video stream where all packets are present at the beginning, but different packets have different deadlines. They then propose a network coding scheme to minimize the amount of packet losses.

In this paper, we focus on the scenario where deadline misses do not cause packet losses. Rather, they make video playback halt. Singh et al. [18] consider the radio resource management problem to reduce video halt. Chandur and Sivalingam [4] propose a scheduling heuristic to improve QoE. Bhatia et al. [2] propose a scheduling policy that exploits slow fading for on-demand video streaming. Joseph and de Veciana [10] propose a joint rate control and scheduling algorithm for optimizing QoE. These studies only focus on long-term average performance, and may be inadequate to capture important characteristics for videos with finite lengths.

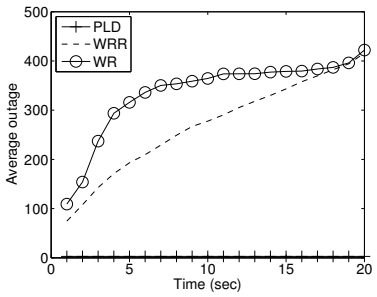
Xu et al. [23] study the probability that video playback halts under a fixed scheduling discipline. ParandehGheibi et al. [17] study the tradeoff between rate, initial buffer size, and the probability of video halt. Anttonen and Mammelaa [1] derive an upper-bound of video halt probability for a G/G/1 queue. Yang et al. [25] uses large deviation theory to approximate video halt probability. These studies focus on the performance of a single user and does not consider the scheduling problem in a multi-user network.



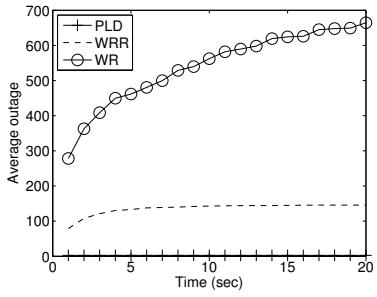
(a) High priority, $p_n = 0.75$, $q_n = 1/40$



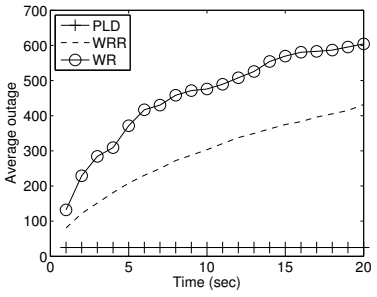
(b) High priority, $p_n = 0.5$, $q_n = 1/65$



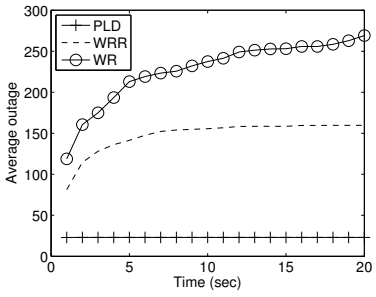
(c) Low priority, $p_n = 0.75$, $q_n = 1/40$, $w_n = 2$



(d) Low priority, $p_n = 0.5$, $q_n = 1/65$, $w_n = 2$



(e) Low priority, $p_n = 0.75$, $q_n = 1/40$, $w_n = 1$



(f) Low priority, $p_n = 0.5$, $q_n = 1/65$, $w_n = 1$

Figure 4: Average outage for an under-loaded system.

9. CONCLUSIONS

In this paper, we study the capacity of wireless networks in providing quality of experience for on-demand video streams. We characterize the performance of a stream by the diffusion limit of its video halt duration. We derive a closed-form expression for the capacity region for QoE. We also propose a QoE-optimal scheduling policy that achieves every point in the capacity region. In addition, we study the performance of two other widely-used scheduling policies. Both theoretical analysis and simulation results demonstrate that our QoE-optimal policy outperforms other policies by large margins.

10. REFERENCES

[1] ANTTONEN, A., AND MAMMELAA, A. Interruption probability of wireless video streaming with limited video lengths. *Multimedia, IEEE Transactions on 16*, 4 (June 2014), 1176–1180.

[2] BHATIA, R., LAKSHMAN, T., NETRAVALI, A., AND SABNANI, K. Improving mobile video streaming with link aware scheduling and client caches. In *INFOCOM, 2014 Proceedings IEEE* (2014), IEEE, pp. 100–108.

[3] BROWN, B. M. Martingale central limit theorems. *The Annals of Mathematical Statistics 42*, 1 (1971), 59–66.

[4] CHANDUR, P., AND SIVALINGAM, K. M. Quality of experience aware video scheduling in lte networks. In *Communications (NCC), 2014 Twentieth National Conference on* (2014), IEEE, pp. 1–6.

[5] CHEN, H., AND YAO, D. D. *Fundamentals of queueing networks: Performance, asymptotics, and optimization*, vol. 46. Springer, 2001.

[6] CSÖRGŐ, M. On the strong law of large numbers and the central limit theorem for martingales. *Transactions of the American Mathematical Society 131*, 1 (1968), 259–275.

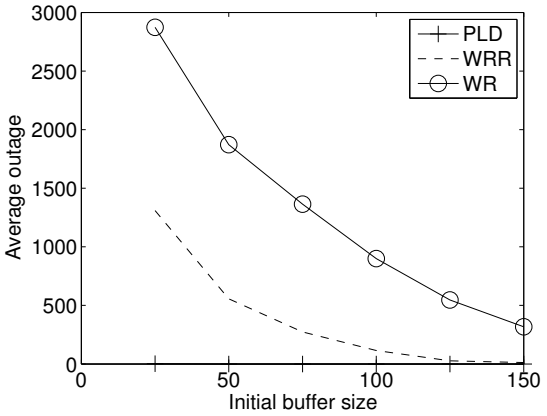
[7] HARRISON, J. M. *Brownian motion and stochastic flow systems*. Wiley New York, 1985.

[8] HOU, I.-H., BORKAR, V., AND KUMAR, P. R. A theory of QoS for wireless. In *Proc. of INFOCOM* (2009).

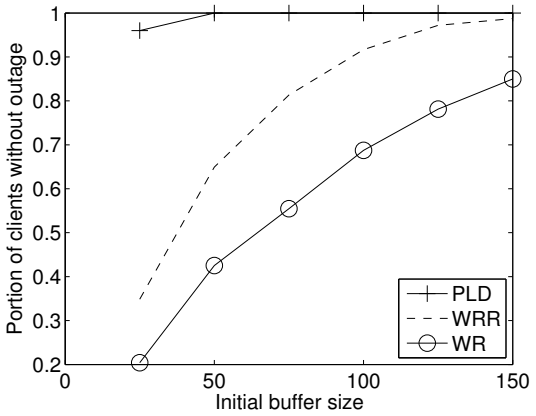
[9] JARAMILLO, J. J., SRIKANT, R., AND YING, L. Scheduling for optimal rate allocation in ad hoc networks with heterogeneous delay constraints. *Selected Areas in Communications, IEEE Journal on 29*, 5 (2011), 979–987.

[10] JOSEPH, V., AND DE VECIANA, G. Nova: Qoe-driven optimization of dash-based video delivery in networks. In *INFOCOM, 2014 Proceedings IEEE* (April 2014), pp. 82–90.

[11] KANG, X., WANG, W., JARAMILLO, J. J., AND YING, L. On the performance of largest-deficit-first for scheduling real-time traffic in wireless networks. In *Proceedings of the fourteenth ACM international symposium on Mobile ad hoc networking and computing* (2013), ACM, pp. 99–108.



(a) Average outage



(b) Portion of clients without outage

Figure 5: Performance for different initial buffer sizes.

[12] KIM, K. S., LI, C.-P., AND MODIANO, E. Scheduling multicast traffic with deadlines in wireless networks. In *INFOCOM, 2014 Proceedings IEEE* (2014), IEEE, pp. 2193–2201.

[13] LI, R., AND ERYILMAZ, A. Scheduling for end-to-end deadline-constrained traffic with reliability requirements in multihop networks. *IEEE/ACM Transactions on Networking (TON) 20*, 5 (2012), 1649–1662.

[14] LI, X., WANG, C.-C., AND LIN, X. On the capacity of immediately-decodable coding schemes for wireless stored-video broadcast with hard deadline constraints. *Selected Areas in Communications, IEEE Journal on 29*, 5 (May 2011), 1094–1105.

[15] LIANG, G. Effect of delay and buffering on jitter-free streaming over random vbr channels. *Multimedia, IEEE Transactions on 10*, 6 (2008), 1128–1141.

[16] MOK, R., CHAN, E., AND CHANG, R. Measuring the quality of experience of http video streaming. In *Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on* (May 2011), pp. 485–492.

[17] PARANDEHGHEIBI, A., MÉDARD, M., OZDAGLAR, A., AND SHAKKOTTAI, S. Avoiding interruptions - a qoe reliability function for streaming media applications. *Selected Areas in Communications, IEEE Journal on 29*, 5 (2011), 1064–1074.

[18] SINGH, S., OYMAN, O., PAPATHANASSIOU, A., CHATTERJEE, D., AND ANDREWS, J. G. Video capacity and qoe enhancements over lte. In *Communications (ICC), 2012 IEEE International Conference on* (2012), IEEE, pp. 7071–7076.

[19] STAELENS, N., MOENS, S., VAN DEN BROECK, W., MARIEÌĹN, I., VERMEULEN, B., LAMBERT, P., VAN DE WALLE, R., AND DEMEESTER, P. Assessing quality of experience of iptv and video on demand services in real-life environments. *Broadcasting, IEEE Transactions on 56*, 4 (Dec 2010), 458–466.

[20] STRASSEN, V. Almost sure behavior of sums of independent random variables and martingales. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Contributions to Probability Theory, Part 1* (1967), University of California Press, pp. 315–343.

[21] TASSIULAS, L., AND EPHREMIDES, A. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *Automatic Control, IEEE Transactions on 37*, 12 (1992), 1936–1948.

[22] TASSIULAS, L., AND EPHREMIDES, A. Dynamic server allocation to parallel queues with randomly varying connectivity. *Information Theory, IEEE Transactions on 39*, 2 (1993), 466–478.

[23] XU, Y., ALTMAN, E., EL-AZOUZI, R., HADDAD, M., ELAYOUBI, S., AND JIMENEZ, T. Analysis of buffer starvation with application to objective qoe optimization of streaming services. *Multimedia, IEEE Transactions on 16*, 3 (April 2014), 813–827.

[24] XU, Y., ELAYOUBI, S., ALTMAN, E., AND EL-AZOUZI, R. Impact of flow-level dynamics on qoe of video streaming in wireless networks. In *INFOCOM, 2013 Proceedings IEEE* (April 2013), pp. 2715–2723.

[25] YANG, J., HU, H., XI, H., AND HANZO, L. Online buffer fullness estimation aided adaptive media playout for video streaming. *Multimedia, IEEE Transactions on 13*, 5 (2011), 1141–1153.