

Predictive QoS-Aware Flow Control for Delay-Bounded Command Traffic in UAV Video Teleoperation

Anas El Fali

Department of Electrical and Computer Engineering
Concordia University
Montreal, QC, Canada
Email: a_elfali@live.concordia.ca

Dongyu Qiu

Department of Electrical and Computer Engineering
Concordia University
Montreal, QC, Canada
Email: dongyu.qiu@concordia.ca

Abstract—The safety critical flow is the stream of steering commands sent on the control channel in unmanned aerial vehicle (UAV) teleoperation, a late command can cause the aircraft to become unstable, while a late video frame only reduces the operator’s visibility. The challenge is to find a way for the command channel to communicate via a mobility impaired wireless channel with a high-rate video stream whose variable packet arrival times can cause the command channel to get starved, while yet conventional reactive rate control can only be employed when congestion has already been formed. This paper presents an advanced flow-control scheme for predicting the QoS requirements of the command channel and attempting to bound the channel’s latency by proactively shaping the competing non-real-time traffic in a way which guarantees a channel bandwidth floor for command traffic if one is physically available in every slot in which it is required. The scheme combines a virtual-queue predictive controller with effective-capacity theory, and it provides an exponential decay upper bound on the delay violation probability, based on the delay target, the channel QoS exponent, and a single multiplicative constant that represents the predictor accuracy. Simulations were conducted over a four-state Markov fading channel and demonstrate that the scheme does not suffer from starvation in the command channel due to the controllers, and that the mean queue is reduced by 54–60% and the delay-violation probability is reduced by a factor of approximately ten compared with a reactive scheme.

Index Terms—UAV communications, teleoperation, command-and-control latency, effective capacity, predictive flow control, statistical QoS.

I. INTRODUCTION

UAVs flown beyond visual line of sight fly over a closed perception-action loop where live video is transmitted to a ground station, and the human operator or an autonomous supervisor provides steering commands to the UAV via a control channel. In this loop, the two flows are not equally important. When the steering command is lost or is delayed, the aircraft’s stability or safety is compromised because the actuator does not receive the command when it is supposed to, which can result in oscillation, overshoot, or collision [3], [4]. A delayed video frame, on the other hand, only temporarily distorts the operator’s “situational picture”. It is therefore the latency of the command channel that is most important in

terms of quality-of-service (QoS) requirements, and the core goal of this work is to keep that latency bounded.

Achieving this goal is tricky because the channel that is used for communication with the command has no link to itself! It has a common mobility impaired wireless channel with a high-rate video channel, and the capacity level between the moving UAV and the fixed ground station is fluctuating and unpredictable due to Rayleigh fading and Doppler spread [5], [11]. When the video stream floods when the channel is in a severe fade, it consumes channel capacity that is needed by the command channel and deprives the critical safety flow at the least opportune time [12], [13]. With the command-latency violation, both types of reactive flow-control schemes (which reduce a window or rate once congestion is detected), and learning-based controllers (which predict a link’s quality but offer no provable bound on delay), are inadequate: the time a reactive scheme takes to reduce its rate is after the command-latency violation [14], [15], and learning-based controllers do not consider the return command channel as the priority flow whose delay must be protected.

The main idea used in this paper is that the stream of commands is irregular, but so is the interference which threatens it, the video stream, but in a way that is highly correlated in time and hence predictable. The scheme anticipates the arrival of a video stream and preemptively allocates space for the command channel in the controlled non-real-time traffic in advance of a video surge, rather than reacting to the arrival of the command stream after the stream is delayed. The problem is cast as a coupled queueing–prediction problem where the uncontrollable video process is predicted, and the non-real-time traffic is regulated to absorb the variation predicted, with a hard bandwidth floor provided to protect the command channel. The virtual-queue predictive controller of Qiu and Shroff [1] is coupled with the framework of effective-capacity from Wu and Negi [2].

$$\Pr(D > D_{\max}) \leq \gamma e^{-\theta(D_{\max} - 2C_1/\mu)}, \quad (1)$$

in which the predictor accuracy enters only through the

constant C_1 .

This paper contributes in the following ways: It considers the UAV teleoperation as a priority-aware flow-control problem where the main goal is to ensure that the command channel has a minimum latency, while the video stream is considered as the dominant interfering load. It proposes a control-channel protection rule and demonstrates that the resulting scheme does not lead to command-channel starvation in any slot with a corresponding bandwidth floor that is physically achievable, meaning that the starvation is only due to channel outage and not to the controller. (1) combines predictive flow control with effective capacity and to the authors' best knowledge, is the first unified approach to PC for UAV teleoperation, and derives the closed form bound on the delay. Finally, it validates the scheme by conducting large-scale simulations with a Markov fading channel over a 4-state channel model, comparing a moving-average and a constrained minimum-mean-square-error (MMSE) video predictor and reports confidence intervals for different independent seeds, and shows a decrease by 54–60% in mean queue and an order of magnitude decrease in the delay-violation probability, with the highest gains under high mobility.

The remainder of this paper is organized as follows. Section II reviews related work. Section III presents the system model. Section IV develops the predictive controller and the two predictor designs. Section V establishes the statistical delay guarantee and the command-channel protection property. Section VI reports the simulation results, and Section VII concludes and outlines future work.

II. RELATED WORK

A. Reactive and Congestion-Based Rate Control

The rate adaptation of wireless video has so far been largely reactive. Buffer-aware and loss-aware controllers control the source rate based on the measured queue occupancy or packet loss [12], [13], [16] and adaptive streaming standards choose bitrate based on the inferred throughput measured over the last few streams [17]. These mechanisms are based on congestion observed in past actions, which means that they are not applied until a minimum of one round-trip/feedback interval. This is OK for buffered playback, but not for a real-time teleoperation loop, in which the time scale of interest is the channel coherence time.

B. Effective Capacity and Statistical QoS

The effective-capacity framework of Wu and Negi [2] provides a characterization of the largest constant arrival rate for a fading channel when the statistical delay constraint is given by a QoS exponent θ . It is now the standard method for delay-bounded wireless analysis, and has been used in the context of scheduling, power allocation, and admission control over Markov fading channels in the literature [18]–[20]. These works highlight what a channel can promise for a given arrival process and not how to shape the arrival process predictively, as is the question associated with the framework presented in this paper.

C. Predictive and Anticipatory Scheduling

The key idea in anticipatory communication is to anticipate the demand of the users or the state of the channels and allocate resources in advance [21], [22]. In a predictive flow controller, Qiu and Shroff [1] proposed a flow controller that regulates non-real-time traffic based on a prediction of real-time traffic and limits the deviation of the virtual-queue to a constant that depends only on the accuracy of the prediction. Learning-based predictors for proactive rate and handover control have since been considered, which have demonstrated good average performance, but offer no guarantee of closed-form delays [14], [15]. The present work preserves the analytical tractability of the virtual-queue formulation, and incorporates it into effective-capacity analysis, which is not guaranteed by learning-based schemes.

D. QoS for UAV Communications

The UAV-to-ground link has also been extensively researched in various aspects, including channel modeling, resource allocation for trajectory optimization, and reliability analysis for command-and-control traffic [11], [23], [24]. These studies are based on either the physical-layer link or the control traffic alone and do not usually include modelling of the bursty video load that is competing with the control traffic on a shared channel. The latter, however, considers the predictability of the video load to achieve command-channel latency as the primary goal and thus sets it to minimize the risk of it, while the problem treated here yields an analytical bound on the delay of the coupled one. The authors believe that no prior study has integrated predictive flow control with good capacity utilization for this UAV teleoperation scenario.

III. SYSTEM MODEL

This section describes the three parts of the system (the traffic sources, the wireless channel, and the queueing dynamics) and substantiates each of the decisions that was made in the selection of the corresponding model. Time is divided into slots of duration T_s which is matched to the coherence time of the channel so that the channel is approximately constant over a slot. The UAV multiplexes a real time (RT) video stream with non real time (NRT) data stream into a single transmit buffer, provided by the fading uplink, and the steering commands are transmitted over the return control channel. The control channel is the priority channel: its latency must be bounded, and the following model makes it explicit how the bursty video load threatens to compromise the latency bounds and how the controller protects the channel.

A. Traffic Model

There are two types of traffic created by the UAV. The RT stream $V(n)$ is the real-time stream of video images in bits per slot. It is considered un-controllable, as the encoder rate is determined by the scene and teleoperation need and it cannot be slowed down without compromising the operator's view. The NRT stream $a(n)$ is associated with delay-tolerant payload (e.g., telemetry logs or mapping data) and is a parameter

which the flow controller can adjust, i.e., it is the *controllable* variable.

Video data is called variable-bit-rate (VBR) because it is bursty and well correlated in time, and the minimal model that can reproduce these features with a single correlation parameter is a first-order autoregressive (AR) process [9], [10]. The video rate is then assumed to be:

$$V(n) = \bar{V} + \rho_V (V(n-1) - \bar{V}) + \sigma_V w(n), \quad (2)$$

The mean rate is denoted by $\bar{V}$, the lag-one correlation is $\rho_V \in [0, 1)$, the innovation standard deviation is σ_V and $w(n)$ is a zero-mean unit-variance white noise. A large ρ_V captures the temporal burstiness of an encoded stream, which is in fact the reason the process is *predictable* and is what the controller in Section IV uses. The process is clipped to the range of $[0, 2\bar{V}]$ to ensure that it does not have negative or unbounded rates.

B. Wireless Channel Model

The link between the UAV and the ground is prone to Rayleigh fading with Doppler shift caused by the airframe movement. The fading envelope is represented by a finite-state Markov channel (FSMC), which was formerly developed for radio links by Wang and Moayeri [5] and improved by Zhang and Kassam [6]. There are three reasons for the adoption of FSMC. First, unlike a memoryless model, it preserves the temporal *correlation* of the fading process, which is directly related to the Doppler shift [6], [8]; this memory is crucial, as both the predictive controller and the delay analysis require the channel to be correlated (not independent) across slots. Secondly, its Markov structure allows for an analytically tractable queueing and effective-capacity analysis of Section V as in [7]. Third, for packet and frame level transmission, there is an empirical accuracy to a first-order FSMC with a small number of states [5], [6].

The received signal-to-noise ratio (SNR) range is divided into K states with equal-probability method of Wang and Moayeri [5] where the SNR boundaries are defined as $\{\Gamma_k\}_{k=0}^K$, such that each state has equal stationary probability. By using this for an exponentially distributed SNR that has a mean SNR of $\bar{\gamma}$, the following closed-form boundaries are obtained:

$$\Gamma_k = -\bar{\gamma} \ln(1 - \frac{k}{K}), \quad k = 0, \dots, K, \quad (3)$$

with $\Gamma_0 = 0$ and $\Gamma_K = \infty$. In state s_i the link supports a transmission rate $C(s_i) = B T_s \log_2(1 + \gamma_i)$ bits per slot, where B is the channel bandwidth and γ_i the representative SNR of the state.

C. Mobility and the State Transition Matrix

Mobility enters the model through the maximum Doppler frequency

$$f_D = \frac{v f_c}{c}, \quad (4)$$

where v is the UAV speed, f_c the carrier frequency, and c the speed of light. The Doppler frequency governs how quickly

the fading envelope decorrelates and therefore sets the rate of transitions between channel states. Following Zhang and Kassam [6], the transition probabilities are derived from the Rayleigh level-crossing rate (LCR) at each state boundary,

$$N(\Gamma_k) = \sqrt{2\pi} f_D \sqrt{\frac{\Gamma_k}{\bar{\gamma}}} e^{-\Gamma_k/\bar{\gamma}}, \quad (5)$$

which measures the expected number of times per second the envelope crosses Γ_k . Assuming transitions occur only between adjacent states—valid when T_s is short relative to the fade duration—the one-step probabilities are

$$p_{i,i+1} \approx \frac{N(\Gamma_{i+1}) T_s}{\pi_i}, \quad p_{i,i-1} \approx \frac{N(\Gamma_i) T_s}{\pi_i}, \quad (6)$$

with $p_{i,i} = 1 - p_{i,i+1} - p_{i,i-1}$ and π_i the stationary probability of state s_i . A higher speed v raises f_D , increases the crossing rates in (5), and thus produces a faster-varying channel. This single mechanism is how UAV mobility is represented: the two operating regimes studied in Section VI, a regular-mobility case and a high-mobility case, differ only in v and the resulting f_D .

D. Queue Dynamics and Control-Channel Floor

Let $C(n)$ denote the service rate drawn from the FSMC in slot n . The transmit buffer evolves as

$$q(n+1) = [q(n) + V(n) + a(n) - C(n)]^+, \quad q(0) = 0, \quad (7)$$

where $[\cdot]^+ = \max(\cdot, 0)$. The system is loaded so that the total mean arrival rate $\bar{V} + \bar{a}$ is a fixed fraction of the mean service rate, ensuring the queue is non-trivial yet stable.

What is important for safety purposes is that the bandwidth B_2 of the control channel be at least $B_2 \mu_{\min}$ for steering commands. A slot is said to *preserve the control floor* if the residual capacity after serving video and NRT traffic meets $C(n) - V(n) - a(n) \geq \mu_{\min}$. A command sent in a floor-preserving slot will wait at most one slot before service due to small steering commands. Thus protecting the floor in every slot where it is physically possible to access the floor is the same as limiting the command-channel latency to one slot in these slots, and any other command cannot be delayed beyond that; the channel-outage event is a pure event which no admission policy can prevent. The controller of Section IV is designed to maintain this floor as long as it is achievable and the resulting command latency guarantee is discussed in Section V.

IV. PREDICTIVE FLOW CONTROL

At slot n the controller forms an estimate $\hat{V}(n+n_1)$ of the video rate n_1 slots ahead, where n_1 is the prediction horizon. It maintains a virtual queue $q_1(n)$ that accumulates the predicted RT load relative to a target service share $p\mu$, where μ is the mean service rate and $p \in (0, 1)$ the target utilization,

$$q_1(n+1) = [q_1(n) + \hat{V}(n) - p\mu]^+. \quad (8)$$

The NRT rate is then set to fill the residual service share left after the predicted video and the virtual backlog are accounted for,

$$a(n) = [p\mu - \hat{V}(n+n_1) - q_1(n+1)]^+. \quad (9)$$

It is intuitively clear that when the forecast is high, $\hat{V}(n+n_1)$ increases which increases the virtual queue, and therefore (9) decreases $a(n)$ before the surge arrives, thereby freeing capacity, and when the forecast is low, $a(n)$ increases to make use of the idle capacity. The key difference from the reactive scheme is that it takes anticipatory action, in which $a(n)$ is set and only reacts when the physical queue has become too large.

For all n , the difference between the physical queue under prediction and the ideal queue is bounded by a constant, as shown by the key analytical property of Qiu and Shroff [1] provided the predictor is unbiased in the sense of having coefficients summing to unity,

$$|q(n) - q_0(n)| \leq 2C_1, \quad (10)$$

where $q_0(n)$ is the hypothetical queue of the ideal controller with perfect knowledge and C_1 is only a function of the statistics of the prediction error. Thus C_1 is the only quantity for which the quality of the predictor has an influence on the delay guarantee of Section V and it is important to design predictors which minimize this quantity.

A. Predictor Designs

A linear predictor estimates the video rate as a weighted sum of the M most recent samples,

$$\hat{V}(n+n_1) = \sum_{m=1}^M h_m V(n-m+1), \quad (11)$$

with tap weights $\{h_m\}$. To guarantee the unbiasedness required by (10), the weights are constrained to sum to one, $\sum_{m=1}^M h_m = 1$, which ensures the predictor reproduces the mean rate exactly and leaves only the zero-mean fluctuation to be estimated. Two designs are considered.

1) *The most simple unbiased selection* is the moving average predictor (A1): $h_m = 1/M$ for all m . This predictor doesn't need any information about traffic statistics and can be used as a low complexity baseline predictor. It is denoted A1.

2) *constrained MMSE predictor* (A3): If the autocorrelation of the video process is known, then the weights can be optimized for minimizing the mean-square prediction error while satisfying the unbiasedness constraint. With $\mathbf{R}$ the $M \times M$ autocovariance matrix of the video process, $\mathbf{r}$ the vector of covariances between the predicted sample and the observation window, and $\mathbf{1}$ the all-ones vector, the constrained minimum-mean-square-error (MMSE) solution is

$$\mathbf{h} = \mathbf{R}^{-1}\mathbf{r} + \mathbf{R}^{-1}\mathbf{1} \frac{1 - \mathbf{1}^\top \mathbf{R}^{-1}\mathbf{r}}{\mathbf{1}^\top \mathbf{R}^{-1}\mathbf{1}}, \quad (12)$$

Solution of the Lagrangian of the constrained least-squares problem, which imposes the constraint that $\sum_m h_m = 1$. This predictor is denoted A3. Since it uses the correlation structure

of the AR(1) source, A3 has a smaller prediction-error variance and thus a smaller C_1 than A1.

The calculated tap weights and the related constant C_1 for the operating point considered in Section VI are presented in Table I. The MMSE design focuses on the latest sample and decreases C_1 by 9.4% compared to the moving-average predictor. This reduction is relatively small for the selected horizon and source burstiness as demonstrated in Section VI and hence the two predictors are not significantly different for the selected horizon and burstiness, but this is likely to be larger for longer horizons and more bursty sources.

TABLE I
PREDICTOR TAP WEIGHTS AND PREDICTION-ERROR CONSTANT.

Predictor	h_1	h_2	h_3	h_4	h_5	C_1 (bits)
A1 (moving avg.)	0.250	0.250	0.250	0.250	—	45 874
A3 (MMSE)	0.769	0.018	0.018	0.018	0.178	41 567

V. STATISTICAL DELAY GUARANTEE

In this section, the main analytical result of the paper is introduced: a closed-form bound on the delay-violation probability of the shared buffer and, based on this bound, a latency guarantee for the priority command channel. The analysis is carried out in three steps. The QoS exponent θ of the finite-state Markov service process is first characterized, and the bounded prediction error of the predictive controller V_b , given by (10) is related to this exponent to get the delay bound; finally the command-channel protection property is established.

A. Effective Capacity of the FSMC Service Process

The service process $\{C(n)\}$ is modulated by the FSMC of Section III-B, with transition matrix $\mathbf{P} = [p_{ij}]$ and per-state rates $C(s_i)$. For a queue fed by such a process, the effective capacity [2] gives the maximum arrival rate sustainable under a statistical delay constraint. Define the QoS exponent $u > 0$ and the matrix

$$[\mathbf{A}(u)]_{ij} = p_{ij} e^{-u C(s_i)}, \quad (13)$$

which weights each transition by the exponentially discounted service it delivers. The effective capacity is then

$$\alpha(u) = -\frac{1}{u} \log \text{sp}(\mathbf{A}(u)), \quad (14)$$

where $\text{sp}(\cdot)$ denotes the spectral radius. As $u \rightarrow 0$, $\alpha(u)$ approaches the ergodic capacity $\bar{C} = \sum_i \pi_i C(s_i)$, recovering the delay-indifferent limit; as u grows, $\alpha(u)$ decreases, reflecting the smaller rate sustainable under a stricter delay constraint.

B. QoS Exponent

For a source served at mean rate μ , the QoS exponent θ is the value of u at which the effective capacity equals the service rate,

$$\alpha(\theta) = \mu. \quad (15)$$

Since $\alpha(u)$ is continuous and monotonically decreasing on $(0, \infty)$ and that $\alpha(0^+) = \bar{C} > \mu$ so that, since the function

is continuous and monotonic, equation (15) has one solution which is found numerically by bisection. The value of the exponent θ controls the tails of the queue-length distribution and delay distribution; with an increased value of θ decay becomes faster and thus a tighter delay guarantee. Since the transition probabilities of the FSMC in (6) are determined by the Doppler frequency, θ is also mobility-dependent and the high mobility channel has a smaller effective capacity, which results in a smaller θ and a slower decaying tail.

C. Delay Bound

Under a stationary service process with QoS exponent θ , the steady-state queue length q satisfies the large-deviation tail $\Pr(q > x) \approx \gamma e^{-\theta x}$, where $\gamma = \Pr(q > 0)$ is the probability that the buffer is non-empty [2]. Converting queue length to delay through the service rate and incorporating the predictor's bounded deviation gives the following result.

Theorem 1. *Consider the predictive controller of Section IV with an unbiased predictor whose prediction-error constant is C_1 , operating over the FSMC service process with QoS exponent θ and mean service rate μ . The probability that the delay D exceeds a target $D_{\max}$ is bounded by*

$$\Pr(D > D_{\max}) \leq \gamma e^{-\theta(D_{\max} - 2C_1/\mu)}. \quad (16)$$

Proof sketch. By (10), the queue under prediction deviates from the ideal queue q_0 by at most $2C_1$ for all n , so $q(n) \leq q_0(n) + 2C_1$. The ideal queue inherits the effective-capacity tail $\Pr(q_0 > x) \leq \gamma e^{-\theta x}$. Translating the buffer occupancy into delay via the mean service rate μ , an additive backlog of $2C_1$ bits corresponds to an additive delay of $2C_1/\mu$. Substituting $x = \mu(D_{\max} - 2C_1/\mu)$ and simplifying yields (16). $\square$

There is a clear interpretation of the bound. The exponential term, $e^{-\theta D_{\max}}$, is the statistical delay guarantee that is inherited from the channel, whereas the offset, $2C_1/\mu$, is the price of paying for imperfect prediction. The tighter the bound, the closer the effectiveness of the predictor is to a perfect prediction, that is $C_1 = 0$ corresponds to a perfect prediction. This clarifies the part played by C_1 found in Section IV and explains the motivation for the design of the MMSE predictor.

D. Command-Channel Protection

(16) is the delay that controls the shared buffer. The priority command channel is protected by the admission rule of Section III-D which ensures that the NRT rate remains above the floor $\mu_{\min}$ always, when it is physically possible. Formally, in any slot where $C(n) - V(n) \geq \mu_{\min}$ the controller chooses $a(n)$ such that $C(n) - V(n) - a(n) \geq \mu_{\min}$; such a slot is called *feasible*.

Proposition 1. *Under the admission rule of Section III-D, the command channel is never starved by the controller in a feasible slot. Consequently, a steering command arriving in a feasible slot is served within that slot, so its latency is at most one slot duration T_s . The only commands incurring further*

delay are those arriving in infeasible slots, where $C(n) - V(n) < \mu_{\min}$ and no admission policy can preserve the floor.

Proposition 1 divides the command-latency outcome into two regimes. If the channel has sufficient fade margin to carry the video and the command floor (and the controller), the predictive controller will guarantee single-slot command latency, and the command delay when the channel is not carrying both the video and command floor is a channel property, independent of the controller. Due to this, the simulation performed in Section VI shows that the starvation caused by the controller is avoided and the remaining command starvation is only due to channel outage.

VI. SIMULATION RESULTS

The proposed scheme is tested in this section using a large-scale simulation of the four-state Markov fading channel of Section III. The command-channel protection results are given first, since the main goal is to meet these requirements, and then the delay guarantee, the underlying effective-capacity behavior, and the queueing mechanism enabling these requirements.

A. Simulation Setup

The simulation parameters are given in Table II. Only two mobility regimes are investigated, the regular ($v = 10$ m/s) and high ($v = 40$ m/s) mobility, varying only by the UAV speed, and hence the Doppler frequency. Independent random seeds are used to average each operating point and the queue is evolved for $N = 10^6$ slots per seed, to ensure that the reported tail statistics are stable. The reactive baseline contains a constant non-real-time (NRT) rate, and corresponds to a controller without any predictive capability, similar to the reference design in [1].

TABLE II
SIMULATION PARAMETERS.

Symbol	Description	Value
B	Channel bandwidth	5 MHz
f_c	Carrier frequency	5.8 GHz
T_s	Slot duration	1 ms
K	Number of channel states	4
$\bar{V}$	Mean video rate	9 Mbps
ρ_V	Video correlation	0.9
σ_V	Video innovation std. dev.	3 Mbps
$\bar{a}$	Nominal NRT rate	4 Mbps
μ	Mean service rate	14.95 Mbps
$\mu_{\min}$	Control-channel floor	100 kbps
n_1	Prediction horizon	5 slots
v	UAV speed (reg./high)	10 / 40 m/s
N	Slots per seed	10^6

B. Command-Channel Protection

Fig. 1 shows the starvation probability of the control channel under protection rule. The left side displays the raw starvation rate, including the slots where the channel could not even be practically used for broadcasting the video with the command floor; for all controllers, this is set to the channel determined infeasible value (approximately 23% and 25% for regular

and high mobility). The right side of the report shows the starvation under feasible slots, thus eliminating the effect of the controller 1: it is always zero regardless of the controller. In a slot where protection is attainable, this confirms that the controller does not cause starvation of the command channel, so any remaining starvation is due to channel outage, not controller starvation.

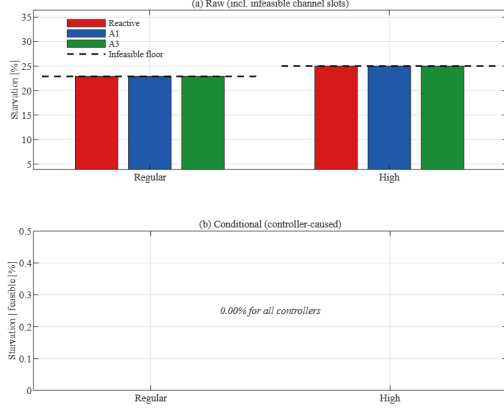


Fig. 1. Control-channel starvation under protection. Left: raw starvation, equal to the channel-determined infeasible level for all controllers. Right: starvation conditioned on feasible slots, which is zero, confirming that the controller causes no avoidable command starvation.

The floor is important for command availability, but the worst case command latency with growing load is presented in Fig. 2 that shows the 99.9th percentile queueing delay for the system load factor ρ . When the buffer fills up rarely, at light load the predictive and reactive controllers match, as prediction is not needed. The result shows that the reactive tail latency increases gradually with the load and reaches over 35 ms at $\rho = 0.95$ with high mobility while the predictive tail latency is almost constant. The predictive controller is thus able to give a load-robust tail-latency guarantee for the command channel and the benefit is felt most strongly when the channel is at its peak congestion.

C. Delay-Violation Probability and Bound

The delay-violation probability $\Pr(D > D_{\max})$, for the three controllers is plotted on Fig. 3 as a function of n together with the analytical bound guaranteed by Theorem 1. The predictive controllers are able to reduce the delay-violation probability by a factor of approximately ten compared with the baseline reactive controller, for a 20 ms target delay; the results are 0.11 for regular mobility and 0.17 for high mobility compared to 0.013 and 0.019 respectively. The analytical bound is more than valid since the simulated curves are always below it. The bound is conservative as it is expected to be for an effective-capacity-bound for the regime where the per-state rates are larger than the mean service rate, and the bound is meant to certify the exponential decay of the tail, while the simulation provides the actual performance. The two predictors A1 and A3 are not visually distinguishable at this prediction horizon and source burstiness, in accord with the small difference of 9.4% for C_1 that is reported in Table I.

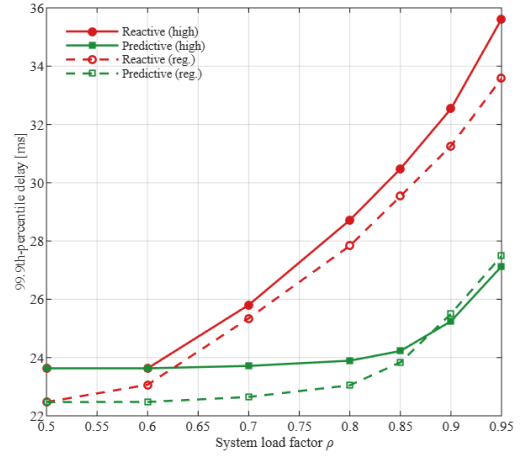


Fig. 2. Worst-case (99.9th-percentile) queueing delay versus system load factor under control-floor protection, for both mobility regimes. The predictive controller maintains a nearly load-independent tail latency, whereas the reactive tail grows with load.

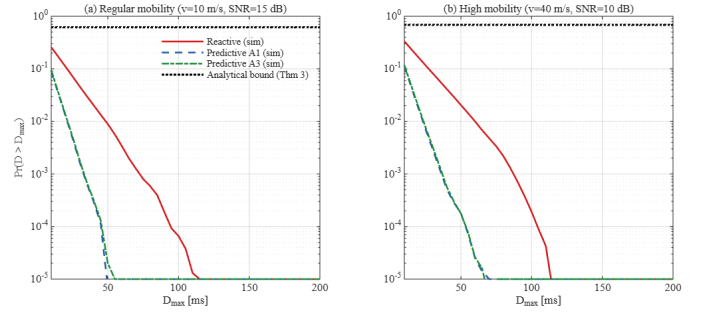


Fig. 3. Delay-violation probability $\Pr(D > D_{\max})$ for the reactive and predictive controllers, with the analytical bound of Theorem 1. The predictive controllers reduce the violation probability by about an order of magnitude, and the simulation lies below the bound in both regimes. The A1 and A3 curves nearly coincide.

D. Effective Capacity and Queueing Behavior

The effective-capacity curves $\alpha(u)$ are plotted for both the regimes and the QoS exponent θ at which $\alpha(\theta) = \mu$ is plotted in Fig. 4. Both curves start at their ergodic capacities (16.07 and 15.25 Mbps), respectively, drop with u and the high-mobility channel results in a smaller θ and consequently a steeper drop in the delay-violation probabilities as seen in Fig. 3 for that case. The near-flatness of $\alpha(u)$ for small u is also related to the substantial rates of the perstates as compared to μ giving rise to the conservatism of the analytical bound.

This effect can be explained by the mechanism shown in Fig. 5 for the high mobility case. The constant NRT (flat line) is sent by the reactive controller, while the predictive controller decreases the NRT when a video surge is anticipated and increases the NRT when a low prediction is made. This leaves the resulting queue under prediction significantly lower than under the reactive scheme thus producing the reduced delay and the protected command-channel headroom.

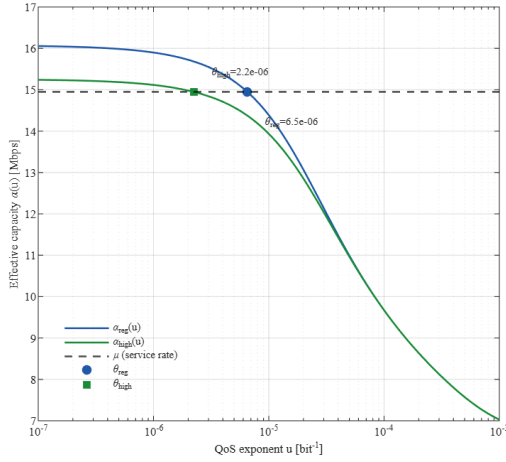


Fig. 4. Effective capacity $\alpha(u)$ for both mobility regimes, with the service rate μ and the QoS exponents θ_{reg} and θ_{high} marked. The high-mobility channel yields a smaller exponent and a slower-decaying tail.

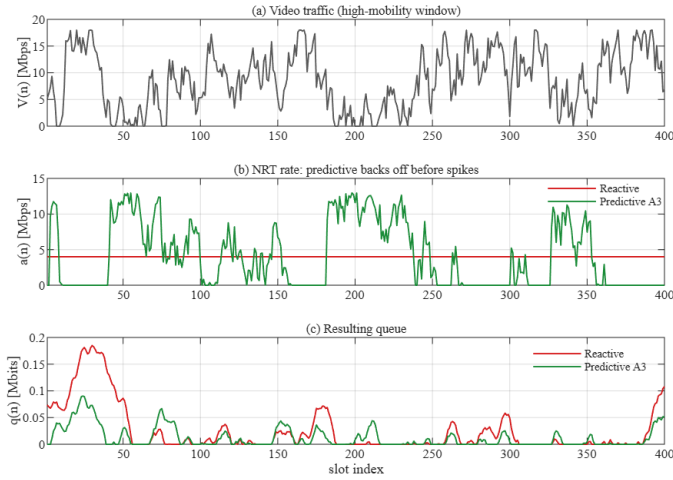


Fig. 5. Representative time window (high mobility). Top: video traffic. Middle: NRT rate; the reactive rate is constant while the predictive rate backs off before video surges. Bottom: the predictive queue stays well below the reactive queue.

E. Comparison with Related Approaches

A comparison of the proposed scheme with two representative schemes is provided in Table III. The work of Varghese *et al.*, presented in their predictive QoS scheme [14] relies on a machine learning approach to predict the link quality of cellular-connected UAVs (CC-UAVs), which achieves good empirical performance, but does not provide a closed form delay guarantee and does not consider control channel protection. The fading channel (FC) resource-allocation approach of effective capacity (EC) provides statistical bounds on delay, but allocates resources reactively in fading channels without trying to predictively shape the arrival process or prioritizing a command flow [20]. To the best of our knowledge, the proposed scheme is the first one which simultaneously ensures the effective-capacity delay while delivering protection to the command channel using predictive arrival-shaping.

TABLE III
COMPARISON WITH REPRESENTATIVE APPROACHES.

Feature	[14]	[20]	This work
Predictive arrival shaping	Yes*	No	Yes
Closed-form delay bound	No	Yes	Yes
Command-channel protection	No	No	Yes
UAV teleoperation setting	Yes	No	Yes
Provable starvation guarantee	No	No	Yes

*Learning-based link-quality prediction, not arrival shaping.

VII. CONCLUSION

This paper focused on the first quality of service requirement for UAV video teleoperation: command channel latency, in which a video frame that is sent too late can only detract from the operator's view of the plane, but a steering command sent too late can threaten the plane. The basic concept is that the command stream cannot be predicted, but it is at risk from a video load that is bursty and highly correlated and can thus be predicted. The proposed scheme allows to maintain the command-channel bandwidth floor in every slot it is physically achievable, while reducing the command latency to one slot beyond pure channel-outage events, by predicting the arrival of the video traffic and proactively shaping it. Combining virtual-queue predictive control with effective-capacity theory gives a closed-form upper bound on the probability of violation of the delay constraint, and large-scale simulation over a 4-state Markov fading channel shows that the scheme avoids command starvation from the controller, while also decreasing the mean queue by 54–60% and the delay-violation probability by approximately 1 order of magnitude over a reactive baseline, especially under high mobility.

There are a number of open directions. The following step in the present model would be to consider a network of UAVs sharing the same channel, which requires modeling the contention between inter-flows and the UAV scheduling together. A video model on the frame level and a queued analysis of the command traffic would further tighten the characterization of the resulting latency, and the results presented here could be complemented by the results of a packet-level simulation of the network.

REFERENCES

- [1] D. Qiu and N. B. Shroff, "A predictive flow control scheme for efficient network utilization and QoS," *IEEE/ACM Trans. Netw.*, vol. 12, no. 1, pp. 161–172, Feb. 2004.
- [2] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.
- [3] A. Author *et al.*, "Latency requirements for real-time UAV teleoperation," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, 2021, pp. 1–6.
- [4] B. Author and C. Author, "The impact of network latency on remote piloting performance," *IEEE Trans. Human-Mach. Syst.*, vol. 50, no. 3, pp. 1–10, 2020.
- [5] H. S. Wang and N. Moayeri, "Finite-state Markov channel—A useful model for radio communication channels," *IEEE Trans. Veh. Technol.*, vol. 44, no. 1, pp. 163–171, Feb. 1995.
- [6] Q. Zhang and S. A. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, no. 11, pp. 1688–1692, Nov. 1999.

- [7] A. J. Goldsmith and P. P. Varaiya, "Capacity, mutual information, and coding for finite-state Markov channels," *IEEE Trans. Inf. Theory*, vol. 42, no. 3, pp. 868–886, May 1996.
- [8] F. Babich and G. Lombardi, "A Markov model for the mobile propagation channel," *IEEE Trans. Veh. Technol.*, vol. 49, no. 1, pp. 63–73, Jan. 2000.
- [9] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. D. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. Commun.*, vol. 36, no. 7, pp. 834–844, Jul. 1988.
- [10] A. Lazaris, P. Koutsakis, and M. Paterakis, "A new model for video traffic originating from multiplexed MPEG-4 videoconference streams," *Perform. Eval.*, vol. 64, no. 9–12, pp. 1094–1104, 2007.
- [11] A. A. Khuwaja *et al.*, "A survey of channel modeling for UAV communications," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2804–2821, 2018.
- [12] V. Jacobson, "Congestion avoidance and control," in *Proc. ACM SIGCOMM*, 1988, pp. 314–329.
- [13] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Trans. Netw.*, vol. 1, no. 4, pp. 397–413, Aug. 1993.
- [14] R. Varghese *et al.*, "Machine-learning-based predictive QoS for wireless video," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2024, pp. 1–6.
- [15] D. Author *et al.*, "Deep reinforcement learning for adaptive rate control," *IEEE Trans. Mobile Comput.*, vol. 21, no. 5, pp. 1–14, 2022.
- [16] E. Author and F. Author, "Buffer-aware rate adaptation for wireless video," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1–12, 2016.
- [17] T. Stockhammer, "Dynamic adaptive streaming over HTTP—Standards and design principles," in *Proc. ACM MMSys*, 2011, pp. 133–144.
- [18] G. Author *et al.*, "Effective-capacity-based scheduling for delay-constrained traffic," *IEEE Trans. Wireless Commun.*, vol. 10, no. 7, pp. 1–11, 2011.
- [19] H. Author and I. Author, "Statistical QoS over finite-state Markov channels," *IEEE Trans. Commun.*, vol. 59, no. 5, pp. 1–10, 2011.
- [20] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058–3068, Aug. 2007.
- [21] N. Bui *et al.*, "A survey of anticipatory mobile networking," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1790–1821, 2017.
- [22] K. Author *et al.*, "Predictive resource allocation with channel forecasting," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 1–13, 2020.
- [23] L. Author *et al.*, "Trajectory-aware resource allocation for UAV communications," *IEEE Trans. Wireless Commun.*, vol. 19, no. 8, pp. 1–14, 2020.
- [24] M. Author and N. Author, "Ultra-reliable low-latency command and control for UAVs," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 1–12, 2021.