EURASIP Journal on
Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

Content aware optimization for video delivery over WCDMA

Kartik Pandit[1], Amitabha Ghosh[2*], Dipak Ghosal[1] and Mung Chiang[2]

Abstract

A content-aware networking framework for transmitting video on the uplink of a Wideband Code Division Multiple Access (WCDMA) cellular network is studied in this article. We consider a multi-user scenario and exploit the underlying structure of the video content to optimally schedule frames in order to minimize the total distortion of video quality across all users. In particular, we propose a novel *Content-Aware Distortion Fair* (CADF) scheme that determines, between multiple flows, the best set of frames to schedule using optimal transmit power, while meeting interference and delay constraints. The optimization problem is considered on the time scale of a single Group of Pictures, and is formulated as a restricted Multiple Knapsack Problem. A key contribution of this study is evaluating the CADF scheme on a Qualcomm 3G emulator, called "High data rate System Emulator 2" (HSE 2), and conducting rate control experiments with different types of emulated channel conditions. Our experimental results show that the CADF scheme significantly reduces video distortion, compared to the existing Foschini-Miljanic closed-loop distributed power control algorithm implemented on the WCDMA uplink. In the Appendix section, we also use the emulator to analyze the Traffic-to-Pilot resource allocation algorithm implemented on the reverse link of Evolution-Data Optimized (EV-DO) Revision-A by profiling over different traffic classes.

Keywords: Content-aware networking, Rate distortion fair, Qualcomm 3G emulator, Video delivery, Power control, Scheduling, WCDMA

Introduction

Motivation for content-aware networking

It is projected that mobile video traffic will comprise more than 70% of all mobile data traffic by 2016 [1]. This challenges many of our basic assumptions about designing future computer communications networks. Traditionally, network protocols have been designed to be content-oblivious, i.e., each bit is assumed to be equally important when it is transported over the network. This has given rise to the problem of *content-pipe divide* [2]. On the one hand, pipe-owners such as Internet Service Providers (ISPs), network infrastructure vendors, and municipalities treat all content equally as simply bits of information to be transported between given nodes. On the other hand, content-providers such as media companies, end-users who post videos online, and operators of both peer-to-peer (P2P) systems and content distribution networks (CDNs) generate content treating the network as simply a means of transportation. This gap between the pipe-owners and content-providers is exacerbated especially for video content, where different frames can have different degrees of importance and can contribute differently to the distortion (i.e., quality degradation) perceived by the human eye.

Content-aware networking refers to the methodology of utilizing the rate-distortion (RD) characteristics [3] of the content to design more adaptive and efficient network protocols. It advocates a cross-layer design philosophy where network resources are allocated and provisioned with optimality criteria that are reflective of the content itself [2,4]. Content-aware techniques provide an alternative for optimizing video quality in the presence of network resource constraints.

There exists a rich literature on content-aware networking, contributed by video processing, wireless, networking, and information theory communities. However, there exists a wide gap between the theory of content-aware

*Correspondence: amitabhg@princeton.edu
[2]Department of Electrical Engineering, Princeton University, Princeton, NJ, USA
Full list of author information is available at the end of the article

networking and its practical adoption in 3G cellular systems. This article focuses on connecting the *analytics* of content-aware optimization with the *specifics* of engineering *hooks* and *knobs* in 3G cellular standards, as well as on implementing them in an industry-grade system emulator that offers a configurable and realistic test-bed to validate the theory.

We study the uplink of the Wideband Code Division Multiple Access (WCDMA) [5] standard, and design techniques to make the physical layer and medium access control (MAC) layer content aware. Our proposed architecture is wrapper-based, i.e., it maintains the overall structure of the physical and MAC layer protocols in WCDMA, but provides hooks that can be implemented as *wrappers*. Such a wrapper-based approach can make content-aware features incrementally available in WCDMA networks without causing major disruption in services. The reason we choose to focus on the uplink is its increasing bandwidth demand. With the recent surge in data-hungry, delay-sensitive applications, such as capturing videos on mobile devices and posting them on social networking sites, online gaming, video conferencing, and Voice-over-IP (VoIP), the asymmetry in data rates between the downlink and uplink is closing.

Another focus of this study is to validate our proposed algorithms using a packet-level Qualcomm 3G emulator, called "High data rate System Emulator 2" (HSE 2) [6]. HSE 2 acts as a bridge between terminal equipments by dropping and delaying IP packets to emulate 3G channels. The delay experienced by IP packets going through the emulator is faithful to that in an actual 3G system under various channel conditions. Unlike Wi-Fi networks in the unlicensed band, where software-defined radios like Wireless Open-Access Research Platform (WARP) [7] readily provide a realistic and configurable platform, experimenting with 3G cellular platforms is much more difficult. Testing the content-aware algorithms on an industry-grade emulator is much more realistic than simulation-based evaluations. Since the internal workings of most public cellular technologies are closed from public access, experimental evaluation on an emulator is a key step to faithfully capture the reality.

Overview

We summarize our key contributions below:

1. *Content-Aware Distortion Fair (CADF) Scheme:* We propose a novel *content-aware distortion fair* scheme for efficiently delivering video traffic on the uplink of a WCDMA network. We formulate the problem of video delivery as a constrained optimization problem of minimizing the sum of distortion of all video streams from all users, and map it to a restricted version of the Multiple Knapsack Problem (MKP) [8]. We then modify a known [9] polynomial time approximation scheme (PTAS) to find a solution to the restricted MKP, and propose two heuristics to further improve the solution by incorporating channel variations.
2. *Joint Scheduling and Power Control:* Our proposed CADF scheme utilizes two degrees of freedom for RD-fair optimization: scheduling and transmit power control. This extends previous work which considers either power control or distortion-fair scheduling without interference management [10,11], but not both [12]. The CADF scheme maintains the overall structure of the existing WCDMA algorithms, but provides hooks that can be implemented as wrappers.
3. *Evaluation on Qualcomm 3G Emulator:* We evaluate the content awareness of WCDMA uplink by implementing the CADF scheme on the HSE 2 emulator under realistic conditions. We emulate multiple mobile stations (MS) in the emulator, where each MS transmits video traffic to a single radio base station (RBS). Our results indicate that the CADF scheme outperforms the Foschini-Miljanic closed-loop power control [13] algorithm currently implemented in WCDMA by achieving lower distortion in video quality.

We also conduct an analysis of the Traffic-to-Pilot (T2P) reverse traffic channel (RTC) MAC algorithm implemented in EV-DO Revision-A to study resource allocation for different traffic classes.

The rest of the article is organized as follows. In the following section, we discuss the related work. In Section "Content-aware distortion fair optimization", we describe the network and video model and formulate the CADF optimization problem. In Section "Solving GOP level CADF problem", we present an approach to solve the CADF problem for WCDMA uplink. In Section "Performance evaluation", we present detailed experimental results conducted on the HSE 2 emulator, and in Section "Conclusions", we make some concluding remarks. A background on the HSE 2 emulator as well as an experimental analysis of the EV-DO T2P RTC MAC algorithm is presented in the Appendix.

Related study

In this section, we present some relevant literature on content-aware networking and resource allocation in WCDMA networks. Joint source-channel coding, which is related to content-aware networking, has widely been studied, however, in the review below, we focus more on its wireless communication issues. We also briefly describe some recent work on EV-DO Revision-A.

A joint source adaptation and resource allocation technique by doing discrete time frame selection is proposed in [14] for a single-hop network with both voice and video users. It does not, however, explicitly optimize the distortion by scheduling frames and determining the power vector to achieve the necessary rates. In a follow-up work [15], several parameters, such as the utility of a frame and fairness among competing users are considered, but without any power control. The RD-fairness of uplink scheduling is formulated as a Multi-User Markov Decision Process in [11], but again without considering power control and interference management.

In the context of 3G cellular networks, the authors of [10,16] consider the video frame scheduling problem on the downlink. A content-aware incremental redundancy error-correction scheme is introduced in [17] for combating both packet loss and bit errors. The scheme is employed within an optimization framework that enables a sender to compute which packets to transmit in order to meet an average transmission rate constraint while minimizing the average end-to-end distortion.

Content-aware networking is also a form of cross-layer optimization, such as the Network Utility Maximization (NUM) framework first proposed by Kelly et al. [18]. An extensive survey of NUM and its applications can be found in [19-22]. However, not all content-aware networking is NUM. For instance, the study in [11] models the problem of achieving a distortion-fair resource allocation as a Markov Decision Process, and solves it using Bellman Equations [23].

An evaluation of the EV-DO Revision-A physical and MAC layers is given in [24]. It is shown that Revision-A approximately doubles both the uplink spectral efficiency and the number of simultaneously active terminals with delay-sensitive applications. A description of the parameters and algorithms used in Revision-A is given in [25]. The work explores the T2P algorithm and analyzes the per flow throughput and delay of IP packets for varying channel conditions, as well the fluctuation in the cumulative sector Rise-over-Thermal (RoT), the ratio of the total received power to thermal noise. An evaluation of rate-fair allocation in a multi-user environment is presented in [26], where the problem is formulated as an NUM, and the utility functions implicit in the rate control algorithms are identified. It is shown that the algorithms are asymptotically stable and converge to a rate-fair allocation, which turns out to have equivalent throughput at equilibrium for all MS.

Content-aware distortion fair optimization

In this section, we first introduce the network and video model, and then formulate the content-aware distortion fair video delivery problem in WCDMA networks. The notations used in this article are listed in Table 1.

Table 1 Notations

Symbol	Description
N	Number of users
n	Number of frames in a GOP
f_{ij}	Video frame j of user i
f'_{ij}	Reconstructed frame if f_{ij} is dropped
$\lambda(f_{ij})$	Length (in bytes) of frame f_{ij}
$t_{\mathrm{DTS}}(f_{ij})$	DTS of frame f_{ij}
$r(f_{ij})$	Rate required to transmit frame f_{ij}
δ_i	Difference in DTSs between consecutive frames of user i
Λ_i	Set of dropped frames by user i
$d(f'_{ij}, f_{ij})$	Frame-level PSNR distortion metric
$D_i(\Lambda_i)$	Distortion per GOP of user i due to frame drop
$R_i(\Lambda_i)$	Required rate per GOP for user i
p_{ij}	Transmit power for frame f_{ij}
θ_{ij}	Indicator variable for scheduling frame f_{ij}
$p_{\max}$	Maximum transmit power
γ_{ij}	Received SINR at base station for frame f_{ij}
Z_{thresh}	Pre-determined RoT threshold

Network and video model

We consider a total of N mobile phone users within a single sector of a WCDMA network. Each user transmits a pre-encoded video stream over a single-hop uplink to the common RBS in that sector, as shown in Figure 1. The individual video frames arrive one-by-one and are stored in the buffer before being transmitted over the air. Each user can decide whether to transmit or drop a frame, and can choose a power level for each transmitted frame.

Each video is encoded using a Group-of-Pictures (GOP) structure, which is typical of codecs such as MPEG-4, H.264, etc. The GOP structure is repeated for the duration of the video. Frames within a single GOP are encoded interdependently using motion estimation, whereas frames in different GOPs are independent. This intra-GOP frame dependency can be represented using a directed acyclic graph, as shown in Figure 2. For this example, the GOP structure is IBBPBBP. Here, every P frame is dependent on the preceding I or P frame, and every B frame is dependent on the preceding I or P frame, as well as on the succeeding I or P frame.

For simplicity of exposition and without loss of generality, we assume that each user transmits a video that has the same GOP structure comprising a total of n frames. We denote this set of frames by $\mathcal{F}_i = \{f_{i1}, \ldots, f_{in}\}$, where f_{ij} is the jth frame of user i. Each frame f_{ij} has a length $\lambda(f_{ij})$, and a decoder time stamp (DTS) $t_{\mathrm{DTS}}(f_{ij})$. The DTS is the time by which the frame should be received to be successfully decoded at the receiver. Each frame also has a *peak signal-to-noise ratio* (PSNR) [27], which is defined

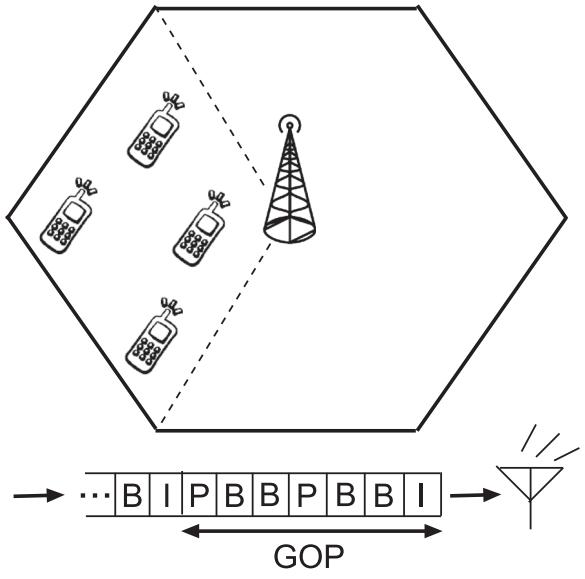

Figure 1 Each user in the same sector transmits a video to a common RBS. Frames arrive one-by-one, and are stored in the buffer before being transmitted in the air.

as a function of the mean square error (MSE) between the original and reconstructed frames. Mathematically, it is expressed as PSNR $= 10\log_{10}(Q^2/D)$, in the logarithmic unit of decibel (dB), where D is the pixel-wise MSE between the original and reconstructed frames and Q is the maximum pixel value (usually 255). We use PSNR as a metric for computing spatial distortion.

Since frames in different GOPs are independent of each other, we consider the CADF problem on the time scale of a single GOP. We assume that the transmission buffer at each MS is large enough to store the frames of at least one GOP before the user can make a scheduling and power control decision. By assigning a transmit power, the user achieves a certain rate that depends on the channel condition as well as the interference caused by other simultaneously transmitting users. Because of interference constraints, the achieved rate can be much lower than the rate required to transmit all the frames in a GOP before their deadlines expire. Therefore, for every GOP, the user selects the *best* subset of frames to transmit and drops the rest.

Let the set of dropped frames for user i be denoted by $\Lambda_i = \{f_{il_1}, \ldots, f_{il_{|\Lambda_i|}}\}$. We assume that a dropped frame is substituted with the nearest frame that was successfully received and decoded. Let the set of reconstructed frames be $\mathcal{F}'_i = \{f'_{i1}, \ldots, f'_{in}\}$, where

$$f'_{ij} = f_{it}, \quad \text{where} \quad t = \arg\max_{f_{im} \notin \Lambda_i} m \leq j. \tag{1}$$

The total distortion $D_i(\Lambda_i)$ of user i as a result of the dropped frames is the sum of the individual frame distortions, i.e.,

$$D_i(\Lambda_i) = \sum_{j=1}^{n} d(f'_{ij}, f_{ij}), \quad \text{for} \quad i = 1, \ldots, N, \tag{2}$$

where the frame distortion metric, denoted by $d(f'_{ij}, f_{ij})$, is the absolute difference in PSNR between the original and the reconstructed frames. The PSNR is typically a non-negative, concave, increasing function of the rate that has diminishing returns for increasing rates. Thus, beyond a certain rate that is required to transmit a frame, the PSNR curve flattens out.

Typically, videos are encoded using a fixed number of frames per second. This makes the difference in DTS values between any two consecutive frames a constant, which we denote by δ_i for user i. Thus, the rate required to transmit frame f_{ij} is $r(f_{ij}) = \lambda(f_{ij})/\delta_i$, and the total rate required to transmit the selected frames of user i within one GOP is

$$R_i(\Lambda_i) = \sum_{j=1}^{n} r(f_{ij}) - \sum_{m=1}^{|\Lambda_i|} r(f_{il_m}). \tag{3}$$

GOP level CADF optimization problem

Each user makes a scheduling and power control decision on a frame-by-frame basis per GOP, i.e., each user periodically considers the frames of the head-of-line GOP stored in the MS buffer, and optimally selects the best subset of frames to transmit as well as their transmit power levels. Making a scheduling and power control decision for every frame is aligned with the physical and MAC layer controls available in HSE 2. The rate and transmit power for a physical layer packet in HSE 2 are chosen based on its payload size and latency requirement. For a video frame, we choose the required rate and transmit power based on its length and DTS. The optimal set of transmitted frames in a GOP depends on solving an optimization problem, which we describe below.

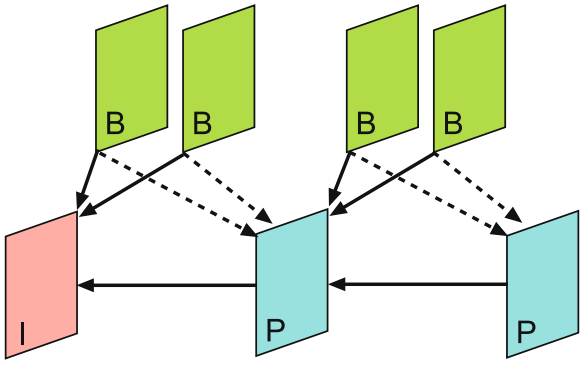

Figure 2 GOP structure IBBPBBP, with arrows indicating frame dependency.

The objective of the GOP level CADF problem is to minimize the sum of distortions over a GOP for all the video streams, subject to the signal-to-interference-plus-noise ratio (SINR) and frame deadline constraints. We denote by θ_{ij} the indicator variable for user i to transmit (set to 0) or drop (set to 1) frame j, and by p_{ij} the transmit power of frame j. Let the maximum transmit power be $p_{\max}$. The rate achieved is a function of the received SINR at the RBS, denoted by γ_{ij} for frame j of user i, which must be at least equal to the frame rate $r(f_{ij})$. However, instead of constraining the achieved rate for *every* selected frame to be at least equal to the frame rate, we only constrain the *total* rate achieved during one GOP to be at least equal to the total rate $R_i(\Lambda_i)$ for transmitting the selected $n - |\Lambda_i|$ frames.

We now formulate the GOP level CADF optimization problem as:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{N} D_i(\Lambda_i) \\ \text{subject to} \quad & R_i(\Lambda_i) \le \sum_{j=1}^{n} d \log\left(1 + c\gamma_{ij}\right), \quad \forall i \\ \text{variables} \quad & 0 \le p_{ij} \le p_{\max}, \quad \forall i,j \\ & \theta_{ij} = \begin{cases} 0, & \text{if frame } f_{ij} \text{ is transmitted} \\ 1, & \text{otherwise} \end{cases} \end{aligned} \tag{4}$$

where

$$\gamma_{ij} = \frac{p_{ij} g_{ii} (1 - \theta_{ij})}{\sum_{k \neq i} p_{kj'} g_{ik} (1 - \theta_{kj'}) + \eta_0}. \tag{5}$$

Here, g_{ii} is the direct channel gain, i.e., the gain on link i for user i, and g_{ik} is the indirect channel gain, i.e., the gain from another user k on link k to the RBS on link i. This indirect channel gain multiplied by the transmit power is the amount of interference caused by another user k when it is transmitting to the RBS as perceived by user i. The transmit power of user k for frame j' is $p_{kj'}$; η_0 is the thermal noise; d is the channel bandwidth; and c is a constant depending on the modulation and coding scheme. The right-hand side of the constraint is the total rate achieved over the duration of a GOP. In our formulation, we ignore the effects of fast fading and assume that the channel gains remain constant over a GOP duration. However, between two consecutive GOPs, the gains are updated using appropriate measurements by the RBS.

The above optimization problem is difficult. It is a (non-convex) mixed-integer nonlinear program, due to the 0–1 integer variables θ_{ij} and the real variables p_{ij}, thus making it NP-Hard [28].

Solving GOP level CADF problem

In this section, we propose an approach to solve the GOP level CADF problem for the WCDMA uplink. We begin by first describing the existing WCDMA power control algorithm to put our solution into context.

Uplink power control in WCDMA

The current WCDMA power control is the well-known Foschini-Miljanic closed-loop distributed power control algorithm [13], which converges to a feasible target SIR vector. The algorithm maintains two loops, as shown in Figure 3. In the outer loop, the RBS estimates the received block error rate (BLER) and selects a target signal-to-interference ratio (SIR) depending on whether the received BLER is above or below a certain target, determined by the service class. In the inner loop, the RBS estimates the received SIR and compares it with a target SIR. Then it sends a transmit power control (TPC) command to the MS to increase or decrease its transmit power based on whether the received SIR is below or above the target SIR, respectively [29].

The outer loop is slower and is executed on the order of 10–100 Hz, whereas the inner loop is faster and is executed in every time slot at 1500 Hz (15 slots per 10 ms frame duration). This implies that the transmit power has a fixed value during a given slot. The power control step size is a parameter of the inner loop and is chosen to be 1 or 2 dB for the uplink, depending on the average mobile speed and other operating parameters. Clearly, this closed-loop power control is not content aware, i.e.,

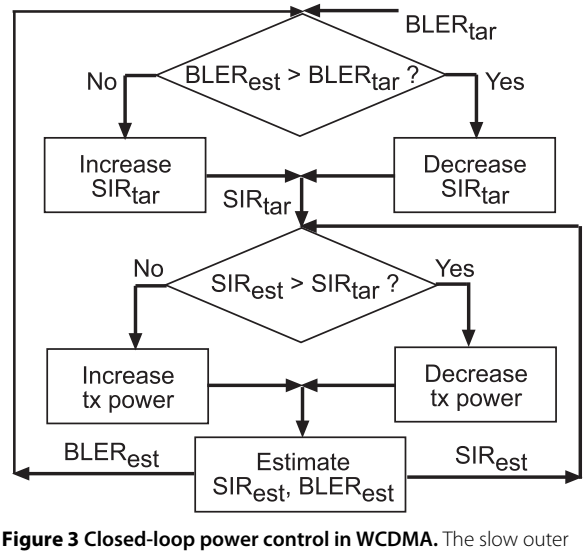

Figure 3 Closed-loop power control in WCDMA. The slow outer loop sets an SIR target by comparing the estimated BLER with a target BLER, whereas the fast inner loop adjusts the transmit power to match the SIR target. The estimated BLER and SIR at the base station are used in the outer loop and inner loop, respectively.

the determination of the target SIR or the transmit power adjustment is done without any knowledge of the video content.

Mapping CADF optimization to restricted MKP

Our approach to solve the GOP level CADF problem is first to map the formulation (4) to a restricted version of MKP [8]. We then modify an existing PTAS [9] to find a solution to the original CADF problem. Finally, we propose two heuristics to modify the PTAS solution for incorporating channel variations that might have occurred during the execution of the PTAS. Our results indicate that the solutions improve when the heuristics are used.

The MKP is a generalization of the single knapsack problem. In MKP, there is a set of items, each with a profit and a weight, and a set of bins (knapsacks), each with a known finite capacity. The goal is to find a subset of items of maximum profit such that they have a feasible packing in the bins. In the following, we show a one-to-one mapping between the CADF problem and a restricted version of the MKP.

In CADF formulation, the individual items are the frames f_{ij}. The profit of an item corresponds to the negative of the PSNR distortion, $-d(f'_{ij},f_{ij})$, and the weight of an item to the frame rate, $r(f_{ij})$. There are a total of $N \times n$ items (n frames per GOP for each user). The capacity of a knapsack corresponds to the total achievable rate, $\sum_{j=1}^{n} d \log(1+c\gamma_{ij})$, for all the selected frames. Thus, there are a total of N knapsacks, one corresponding to each user. In addition, we have the restriction that the total rate $R_i(\Lambda_i)$ required to transmit all the frames selected from the GOP of user i has to fit the achievable rate for user i. In other words, unlike in the MKP, here user i cannot select a frame and schedule it for transmission using the capacity of another user $k \neq i$, i.e., the knapsack capacities are not interchangeable.

An important difference with MKP: We point out an important difference between the GOP level CADF problem and the restricted MKP in terms of knapsack (i.e., channel) capacities. The capacity constraints in the CADF problem are coupled with each other by the SINR constraints. This means allocating a certain transmit power to one user not only consumes its own capacity, but also reduces the capacity of other users due to interference. Thus, it is not until we actually start scheduling the frames and assigning transmit powers that we come to know of the residual channel capacities. Therefore, the achievable rate for a given user is not known prior to actually scheduling the frames, which corresponds to packing the items. This is in contrast to the restricted MKP, where the knapsack capacities are fixed and known *a priori*. In our approach, we overcome this interdependency of channel capacities by iteratively solving the PTAS for the restricted MKP. In the following section, we describe an adaptive solution to the CADF problem using this iterative PTAS, and combine the solution with two heuristics for incorporating channel variations that might have occurred during multiple runs of the PTAS.

An adaptive solution to CADF problem

The overall algorithm for the CADF problem is described as a flowchart in Figure 4. Similar to the Foschini-Miljanic algorithm in WCDMA [13], our algorithm also implements a closed-loop power control, however, unlike the Foschini-Miljanic case, we use coordinated power control that has three closed loops: (i) a content-aware outermost loop, (ii) a channel-aware outer loop, and (iii) a power control inner loop. Below, we describe the functionality of each of the loops.

Content-aware outermost loop

The content-aware outermost loop is executed for every GOP. Its purpose is to adjust a target SINR for each user by determining an optimal set of frames to be scheduled and their corresponding transmit powers.

The RBS first estimates channel capacities and receives information about the rates and distortions for the frames in one GOP from every user. It then finds an approximate solution to the GOP level CADF problem by iteratively running the PTAS and solving the restricted MKP multiple times. Since the effective channel capacities are mutually coupled with each other through the SINR constraints, and are not known *prior* to actually scheduling the frames, we bootstrap the algorithm by using a small initial guess of the transmit power for all users. This gives fixed channel capacities, which we use to run the PTAS for the first time. For every subsequent iteration, we increase the transmit power for one of the users that has the lowest capacity so far. We continue this iterative process of solving the PTAS until convergence, i.e., until none of the users can increase their transmit power without violating the SINR constraints. In our experiments, we ran the PTAS 25 iterations on average.

A solution to the CADF problem comprises an optimal subset of frames that can be scheduled by the user, as well as an optimal transmit power for each of the selected frames. The iterative execution of the PTAS is represented by the box labeled "Solve GOP level CADF problem" in the flowchart of Figure 4. Using this optimal subset of frames and the transmit powers, the RBS then updates the target SINRs for all users from their previous values used during the last execution of the outermost loop.

Channel-aware outer loop

The channel-aware outer loop is executed once every GOP, right after the content-aware loop has finished execution. Its purpose is to detect any variation in channel capacities that might have occurred during the execution

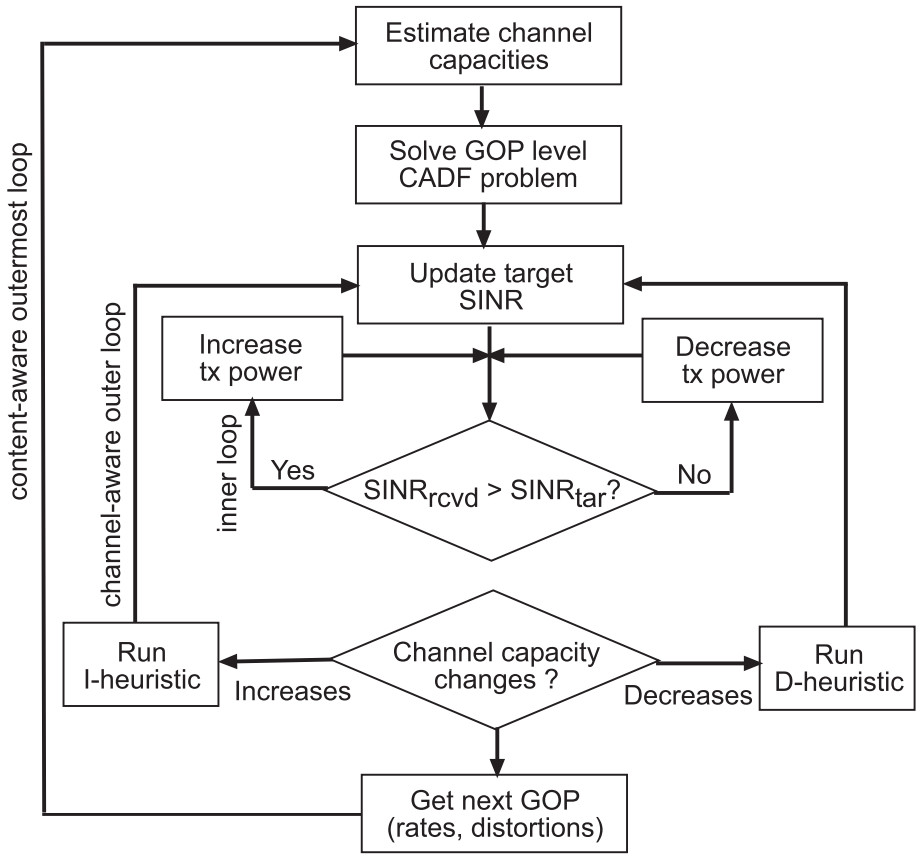

Figure 4 A content-aware adaptive closed-loop power control to the CADF problem for the WCDMA uplink.

of the outermost loop. The loop uses a heuristic depending on whether the channel capacities have increased or decreased.

When the channel capacity increases, implying that some more frames can be selected for transmission, we use a heuristic, called the *I-heuristic*, to perform the frame selection process. In the outermost loop, the RBS keeps track for every user the set of frames not selected (i.e., the frames selected for dropping) for transmission. When the channel capacity increases for a particular user, the I-heuristic first ranks all these frames in increasing order of their required transmission rates, and then breaks ties by a second ranking based on their increasing utility (negative distortion) to rate ratios. The frame selected for transmission is the one with the smallest required rate that fits into the extra capacity and has the best utility to rate ratio, without violating the SINR constraints. An optimal transmit power is also selected for the frame, and the target SINRs are updated to accommodate this newly selected frame. This conservative strategy of packing the smallest unpacked item is proposed in [30].

When the channel capacity decreases, implying that some of the already selected frames need to be dropped, we use another heuristic, called the *D-heuristic*. In the outermost loop, the RBS keeps track for every user the set of frames selected for transmission. When the channel capacity decreases for a particular user, the D-heuristic first ranks the set of selected frames for that user in decreasing order of their required transmission rates, and then ties are broken by a second ranking based on their decreasing utility to rate ratios. The frame selected for dropping is the one with the largest rate and the least utility to rate ratio. The target SINRs are also updated accordingly. Lastly, when the channel capacities remain the same, no frames are dropped and the target SINRs remain the same. At the end of execution of this loop, the set of scheduled frames along with their transmit powers are sent to the respective user.

Power control inner loop

The inner loop power control in our algorithm is the same as the Foschini-Miljanic inner loop. In this loop, the RBS compares for every user the received SINR with the target SINR, and indicates the user using TPC commands whether to increase or decrease its transmit power. In particular, if the received SINR is more than the target SINR,

the RBS indicates the user to decrease its transmit power, otherwise it indicates the user to increase its transmit power.

Performance evaluation

In this section, we describe our experimental testbed and evaluate the performance of the CADF scheme, both with and without heuristics, for the WCDMA uplink.

Testbed setup: Qualcomm 3G emulator HSE 2

The Qualcomm 3G emulator HSE 2 provides several scenarios for choosing different channel conditions that determine the uplink rates. The WCDMA uplink in the emulator implements the Foschini-Miljanic power control algorithm that does rate allocation based on the received SIRs from the users. The emulator does not emulate antenna transmit power, instead, the transmit powers are mapped to the rates, which are then used to transmit the video frames. The channel gain matrices are stored as text files in the emulator. These matrices depend on the channel gains as well as on the type of antennas, such as pedestrian, vehicular, single, and dual. In addition, the background interference or noise can be controlled by setting a reverse activity bit (RAB). The emulator supports ten different channel profiles depending on the antenna, channel gains, and RAB parameters. Each channel profile consists of a text file with 20,000 slots that emulate channel conditions. Each slot represents the channel for a 2-ms interval.

Our testbed is shown in Figure 5, where the laptops emulate the users. HSE 2 emulates the radio link between the RBS and the user by injecting appropriate packet delays, transmission rates, and success or failure probabilities. In all our experiments, we use five laptops transmitting the same video. We set high RAB activity to simulate heavy interference from non-video transmitters. The video we experiment with is the well known "foreman" video sequence [31], which is 12-s long and contains 250 frames across 6 GOPs with an MPEG-2 encoding. The difference in DTS values between two successive frames is 0.04 s. One of our objectives is to verify whether or not the MKP algorithm keeps up with the video playback.

Performance of CADF power control algorithm

We evaluate the GOP level CADF power control algorithm both with and without heuristics on a high-gain, high-noise environment with all users transmitting using vehicular antennae. The performance results for both cases, as well as a baseline algorithm running a variant of the Foschini-Miljanic power control loop, are shown in Figure 6. The baseline algorithm is called the EV-DO power control algorithm and we refer readers to see the section "Uplink power control in EV-DO Revision-A" in Appendix for details.

Evaluating CADF scheme without heuristics

The CADF scheme comprising just the solution of the restricted MKP can be thought of as a content-aware outer loop that runs every second. This outer loop specifies a target rate for the user, which then tries to achieve that target rate in the Foschini-Miljanic inner loop.

The results shown in Figure 6 indicate that for GOPs 1 and 4, the CADF scheme without heuristics incurs higher distortion than the baseline Foschini-Miljanic algorithm.

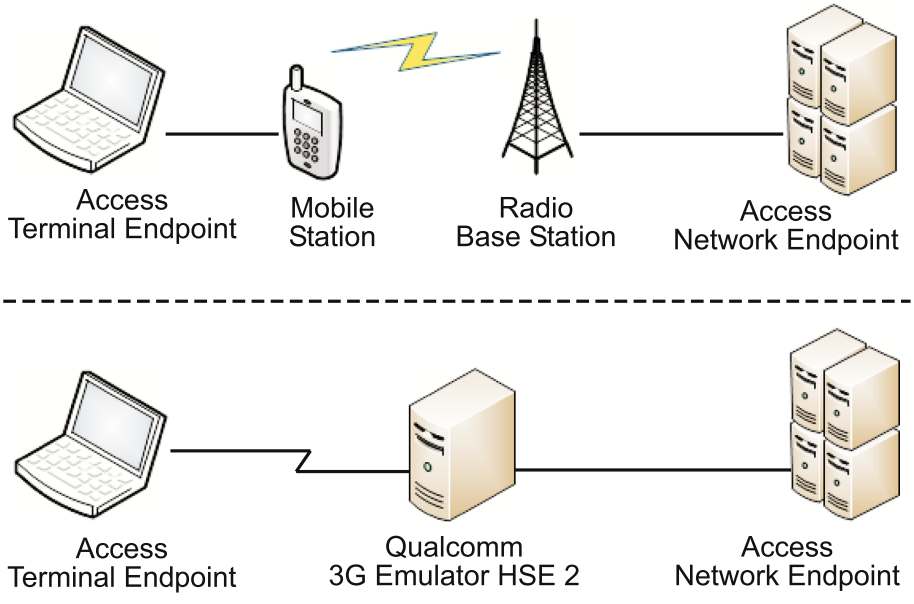

Figure 5 Experimental testbed setup for WCDMA uplink power control using the Qualcomm 3G emulator HSE 2.

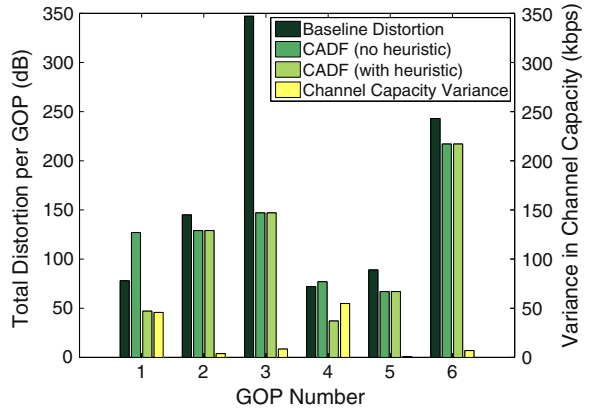

Figure 6 Comparison of GOP level distortion of the baseline Foschini-Miljanic algorithm with the CADF scheme, both with and without heuristics, for different channel capacity variances. The CADF scheme without heuristics has higher distortion for GOPs 1 and 4 than the baseline Foschini-Miljanic, whereas with heuristics it performs better for all GOPs because it incorporates channel variations.

This is because, without the heuristics, the CADF scheme measures channel conditions only once per second. In particular, the channel parameters are taken from the first 2 ms slot during which the GOP is transmitted. Since one GOP takes about 1000 time slots for the chosen video, as we go forward in time, our optimization continues to use the old channel gains. Thus, if there is a substantial change in channel gains from the original values, the optimization performs sub-optimally. This is indicated by the measurements of the channel capacity variance, also shown in the same figure plotted with reference to the right-hand y-axis. For GOPs 1 and 4, we see a high variance in channel capacity. The Foschini-Miljanic algorithm, however, uses current channel conditions while determining the target SIR in the outer loop, and therefore performs better in some cases.

Evaluating CADF scheme with heuristics

To incorporate channel variations, we augment the restricted MKP algorithm with two adaptive heuristics that modify the optimization decisions. We evaluate the CADF scheme with heuristics on all ten scenarios, but due to space limitations, we will present the worst-case scenario to showcase the results. This scenario is the same high-gain, high-load, vehicular antenna scenario shown in Figure 6. The results show that the CADF scheme with heuristics performs better than the baseline Foschini-Miljanic for all GOPs, even in the segments of the video that experience a high variance in channel capacity.

In Figure 7, for a single user we show the performance of the CADF scheme with heuristics on a frame-by-frame basis, as compared to the Foschini-Miljanic algorithm. We observe that in the CADF scheme with heuristics, frames 13 and 14 are not selected for transmission, as seen by the spikes in distortion (triangular markers). This frees up channel capacity for the CADF scheme to schedule the higher distortion frames 17–19. On the other hand, although the Foschini-Miljanic algorithm transmits frames 13 and 14, the higher distortion frames 17–19 are either dropped or could not be decoded. As expected, the CADF scheme with heuristics minimizes the overall distortion for this particular user.

For simplicity of exposition, our formulation of the GOP level CADF optimization problem, as well as the experiments we conducted, assume the same GOP structure for all users. However, the framework and algorithms we presented also work when different videos with different GOP structures. Since our optimization time scale is the duration of one GOP, we then need to consider the respective number of frames in each GOP for different users, and choose the best subset for each user. Since our framework is centralized, so long as the RBS knows about the different GOPs, the algorithms will work.

Conclusions

In this study, we proposed a content-aware optimization framework for efficient video delivery in WCDMA cellular networks. We focused on how to connect optimization theory results with practical systems in 3G standards. We presented a content-aware distortion fair scheme to optimally choose transmit powers and schedule frames across multiple users, each transmitting a video on the

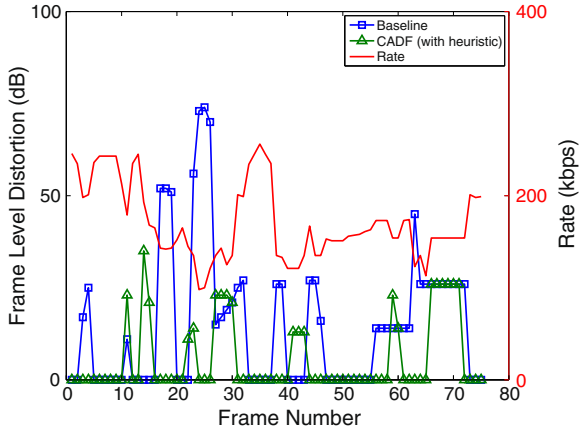

Figure 7 Comparison of frame-level distortion of the baseline Foschini-Miljanic algorithm with that of the CADF algorithm with heuristics for given rates. The CADF algorithm with heuristics sacrifices scheduling frames 13 and 14 of low distortion to free up the channel for future high distortion frames 17–19, whereas the baseline Foschini-Miljanic algorithm myopically schedules low distortion frames and suffers later for dropping high distortion frames.

WCDMA uplink. Our proposed CADF scheme modifies a known PTAS, and combines it with two adaptive heuristics to find a solution to the CADF problem. Moreover, the proposed scheme maintains the overall structure of the existing Foschini-Miljanic power control algorithm in WCDMA systems, and only provide hooks so it can be implemented as a wrapper in existing standards. We evaluated the CADF scheme on the Qualcomm 3G emulator HSE 2 by conducting detailed experiments. Lastly, we also conducted an analysis using HSE 2 to profile the behavior of the T2P resource allocation algorithm in EV-DO Revision-A by experimenting with multiple traffic classes. Our future work lies in investigating content aware techniques for 4G Long-Term Evolution.

Appendix

Background on 3G emulator and EV-DO

In this appendix, we briefly describe the Qualcomm 3G emulator HSE 2, which we use extensively to validate the algorithms proposed in this study, as well as give a brief overview of the resource allocation techniques implemented in EV-DO Revision-A uplink. This background is helpful in understanding how the baseline algorithm does power control in our experiments.

Qualcomm 3G Emulator HSE 2

The HSE 2 is an industry-grade packet level software emulator that emulates the behavior of a 3G system at the IP layer. Among its many features, the HSE 2 supports the following for uplink emulation:

- Emulates the air interface and backhaul delays for EV-DO Revision-A.
- Determines the uplink rates from the T2P algorithm implemented in ED-VO Revision-A.
- Supports five different intra- as well as inter-MS QoS classes on the uplink.

The emulation behavior is controlled by a set of configuration items, which can be adjusted as necessary to emulate a specific scenario. Each scenario is described by its own configuration file, and is referred to as a *canned scenario*. The configuration files are:

- Channel files, specifying channel gains, and antenna profiles.
- Three QRAB files (light, medium, and heavy), specifying the uplink sector loading.
- MS scenario setup files, which configure the number of MS and the type of uplink traffic.
- Uplink QOS-rules file, which for every MS flow matches a packet to an appropriate QoS rule. There are a total of five rules for the five QoS classes.

The emulator also allows event logging and creates a log file during the course of emulation. The log file logs a number of different events, but for this study, we focus only on those related to the T2P algorithm, such as T2PInflowMin, T2PInflowMax, DeltaT2PInflow, T2PInflow, T2POutflow, BucketFactor, BucketLevelSat, BucketLevel, QRAB, and FRAB.

Uplink power control in EV-DO Revision-A

The EV-DO Revision-A system was designed primarily to improve the uplink capacity by increasing each user's spectral efficiency, and to support low-latency delay-sensitive applications, such as VoIP, online gaming, multimedia streaming, etc. High spectral efficiency in EV-DO Revision-A is achieved by a closed loop two-sided power control. The RBS determines a long-term average transmission resource, called the T2PInflow, for each flow in the network, while the MS controls the time-critical allocation, called the TxT2P (Transmit T2P), for each physical layer packet depending on a T2P profile (see Figure 8). An adaptive token bucket mechanism is used at the MS to convert the T2PInflow into a packet level resource. Since each MS perceives other-MS power as interference, this comprehensive control limits the overall sector received

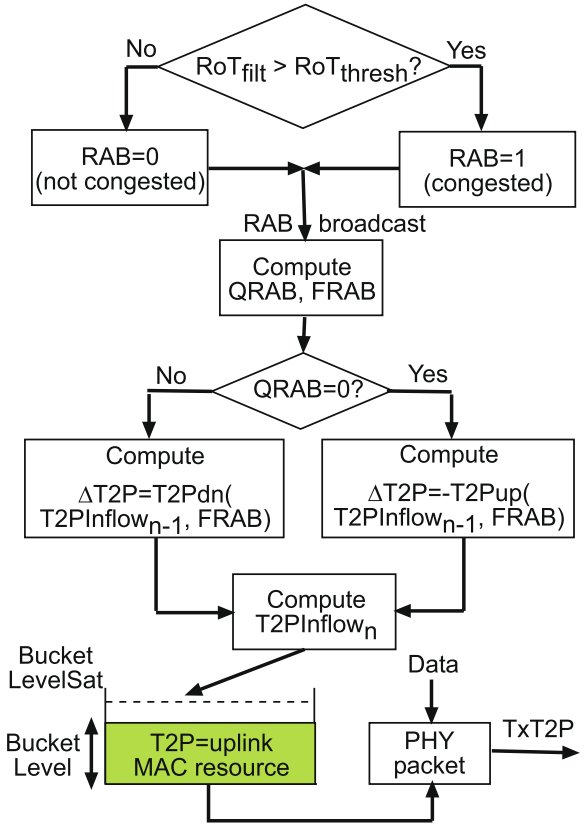

Figure 8 Closed loop power control in EV-DO Revision-A.

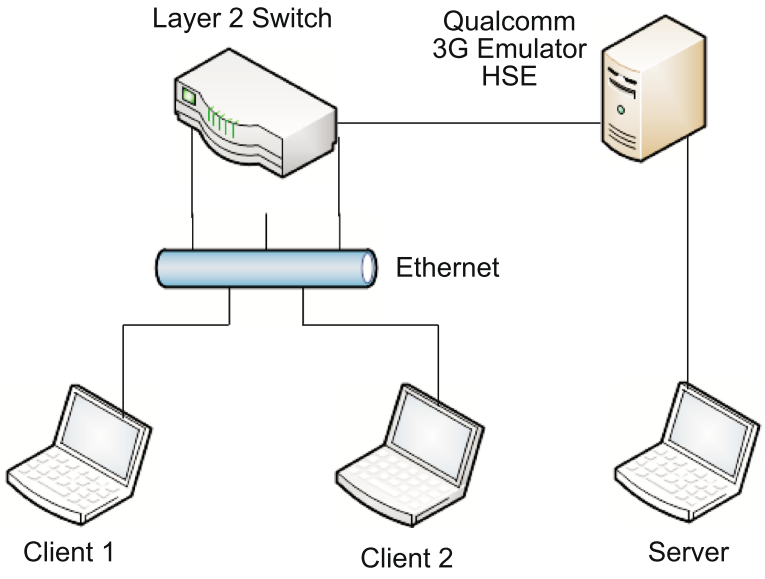

Figure 9 Experimental testbed setup for EV-DO Revision-A uplink T2PInflow analysis using the 3G emulator HSE 2.

power in such a way that the RoT is maintained under a certain threshold.

The T2PInflow is determined based on a single control bit, called the RAB, and two load-adaptive ramping functions, T2Pup and T2Pdn. In each sector, the RBS measures the RoT in each slot and passes it through a filter to compute a filtered RoT. This filtered RoT is then compared with the RoT threshold and the RAB is set accordingly. If it is larger than the threshold, the RAB is set to 1 (congested), otherwise it is set to 0 (not congested). Each flow receives in every slot the RABs from all the sectors in its active set of sectors and computes a Quick-RAB (QRAB) using a logical OR operation. The QRAB is an indication of instantaneous sector loading. The MS also computes a filtered RAB (FRAB) that reflects the long-term sector loading. If the QRAB is 1, the MS flow decreases its T2PInflow level by T2Pdn, otherwise it increases the T2PInflow by T2Pup. Thus, when a sector

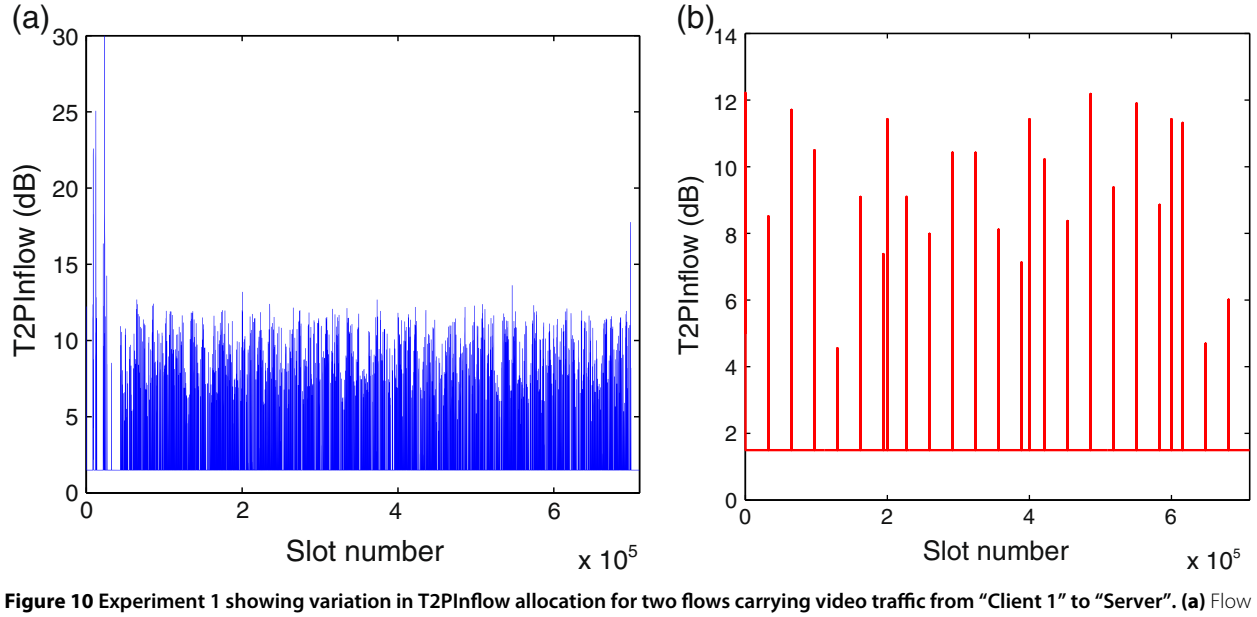

Figure 10 Experiment 1 showing variation in T2PInflow allocation for two flows carrying video traffic from "Client 1" to "Server". (a) Flow f8 shows high T2PInflow values due to its highest priority, and rapid T2P variation due to fast ramping functions. **(b)** Flow f9 shows low T2PInflow values due to its lowest priority, and only occasional T2P variation due to slow ramping functions.

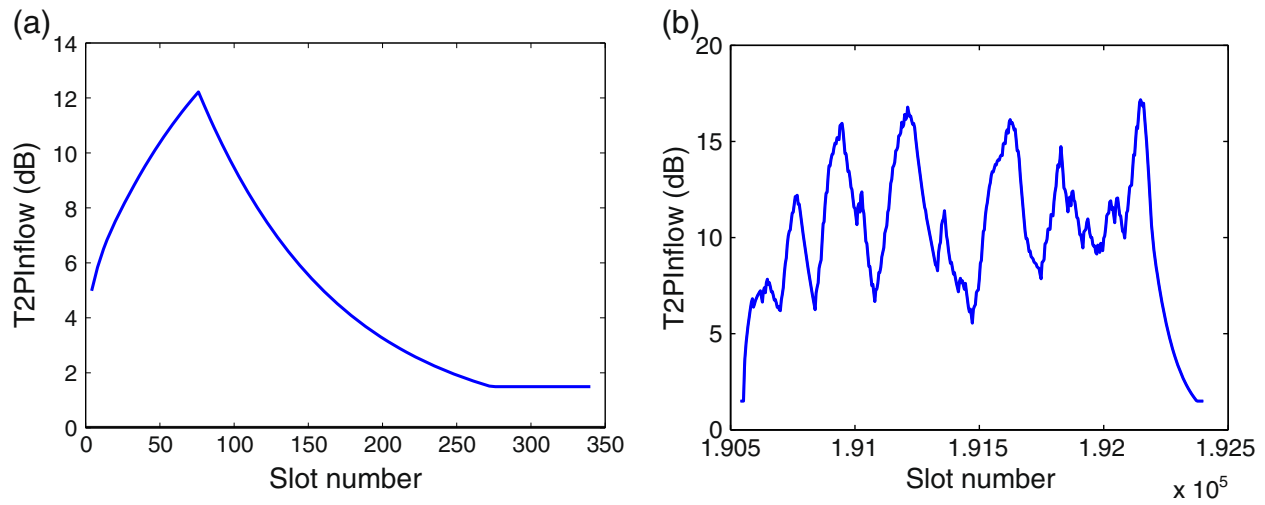

Figure 11 Experiment 2 showing variation in T2PInflow allocation for flow f8 carrying video traffic from "Client 1" to "Server". (a) Initial slots of flow f8 showing a gradual increase in T2PInflow until it reaches a maximum when the sector RoT threshold is exceeded, and then a gradual decrease to a steady value before increasing again. **(b)** Intermediate slots of flow f8 showing peaks and troughs in T2PInflow due to exceeding the RoT threshold when the T2PInflow allocation becomes very high, and then decreasing until the RoT comes down.

is busy, the associated flows ramp down their T2PInflow levels, thereby lowering the sector RoT, and vice verse.

T2PInflow analysis on 3G emulator HSE 2

In this section, we give experimental results that describe the behavior of the T2P RTC MAC algorithm in EV-DO Revision-A.

Testbed setup

We conduct an analysis of the T2PInflow assignment to various flows with different QoS classes for an EV-DO Revision-A uplink. Our experimental testbed is shown in Figure 9. The testbed consists of two laptops that act as two MSs, labeled as "Client 1" and "Client 2". They are connected via a layer 2 switch to the HSE 2, which, in turn, is connected to a laptop labeled as "Server". The path between the client laptops and the server laptop acts as an EV-DO air interface, which emulates packet delays and drops that would take place in an actual EV-DO sector. The Revsion-A standard specifies a per flow T2P allocation that is a function of the QoS class to which the flow belongs. The emulator supports five different QoS classes

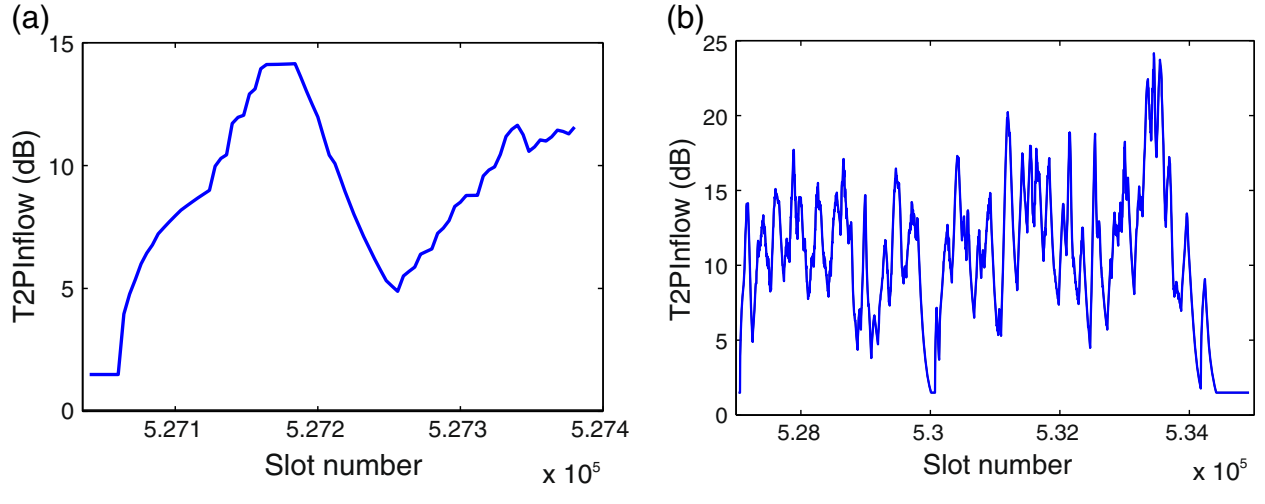

Figure 12 Experiment 2 showing variation in T2PInflow allocation for flow f20 carrying video traffic from "Client 2" to "Server". (a) Initial slots of flow f20 showing a gradual increase in T2PInflow until it reaches a maximum when the sector RoT threshold is exceeded, and then a gradual decrease to a steady value before increasing again. **(b)** Intermediate slots of flow f20 showing peaks and troughs in T2PInflow due to exceeding the RoT threshold when the T2PInflow allocation becomes very high, and then decreasing until the RoT comes down.

on the uplink, namely "flow1", "flow2", "flow3", "flow4", and "flow5", ordered from highest to lowest priority.

Experiment 1: six flows with FTP-only traffic

In our first experiment, we set up six uplink flows divided between the two MSs. "Client 1" has three flows, which we enumerate as f2, f8, and f9. Of these, flow f8 is set to the highest priority QoS class "flow1", and flow f2 is set to class "flow3". Flow f12 is set to the lowest priority class "flow5". We set up FTP traffic on all these three flows. Similarly, for "Client 2" we also set up three flows, enumerated as f10, f11, and f12. All these flows are set to the lowest priority QoS class "flow5", and all of them carry FTP traffic.

In Figure 10a and b, we show the T2PInflow allocation for flows f8 and f9, respectively, over the duration of the experiment. As expected, flow f8 being of the highest priority class "flow1" has a larger T2PInflow allocation than the lowest priority flow f9 of class "flow5". In addition, the rapid variation of T2PInflow for flow f8 indicates that whenever the sector RoT goes below the predetermined threshold, the RBS ramps up the allocation very quickly using the adaptive ramping function T2Pup. However, flow f9 being of lowest priority, the ramping is not as fast as in flow f8, and the allocation thus remains constant at around 1.5 dB over a large number of time slots with only occasional jumps.

Experiment 2: 12 flows with FTP & video traffic

In our second experiment, we set up 12 uplink flows divided between the two MS. "Client 1" has 6 flows, which we enumerate as f2, f8, f9, f10, f11, and f12. Of these, flow f8 is set to the highest priority QoS class "flow1", and flow f2 is set to class "flow3". Flows f9, f10, and f11 are of class "flow4", and flow f12 is set to the lowest priority class "flow5". We set up MPEG-4 video traffic on flows f8 and f2, and FTP traffic on flows f9, f10, f11, and f12. Similarly, for "Client 2" we also set 6 flows, which we enumerate as f14, f20, f21, f22, f23 and f24. Of these 6 flows, flow f20 is set to the highest priority class "flow1", and flow f14 to class "flow3". The rest of the flows are set to class "flow4". We set up MPEG-4 video traffic on flow f20, and FTP traffic on all the other flows.

In Figure 11a, we show the T2PInflow allocation during the first 340 slots for flow f8, which is carrying MPEG-4 video traffic from "Client 1" to "Server" through the emulator. As expected, the T2PInflow gradually increases when the video transmission starts until the sector RoT is exceeded at slot 76. At that point, the RBS sets the RAB for this sector to be busy, and the T2PInflow starts decreasing before reaching a steady value of 1.5 dB.

In Figure 11b, we show the T2PInflow allocation over a large number of intermediate slots (from slot 190,540 to slot 250,000) so we can observe the trend when other flows also exist in the network. As before, the T2PInflow values show a predictable trend of first gradually increasing to reach a maximum, at which point the sector RoT exceeds the threshold, and then slowly decreasing to let the RoT come down, before increasing again.

In Figure 12a and b, we show similar tends in the T2PInflow allocation over a short range and a long range of time slots, respectively, for flow f20 that also carries MPEG-4 video traffic and is of highest priority QoS class "flow1" from "Client 2" to "Server". All other low priority flows which carry FTP traffic have constant T2PInflow allocations and are not shown here.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This study was in part supported by the NSF grant 917315, NSF NeTS grant, DURIP grant, and Qualcomm donation. We would like to thank Prashanth Hande from Qualcomm, and Michael Wang from the EDGE Lab at Princeton University for their insightful discussions during the course of this study.

Author details

[1]Department of Computer Science, University of California, Davis, CA, USA. [2]Department of Electrical Engineering, Princeton University, Princeton, NJ, USA.

Received: 16 February 2012 Accepted: 16 June 2012
Published: 13 July 2012

References

1. Cisco-Systems, Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011–2016. http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf (2011)
2. M Chiang, The content-pipe divide. in *IEEE International Conference on Multimedia and Expo*. (New York, 1550–1551, 2009)
3. GJ Sullivan, T Wiegand, Rate-distortion optimization for video compression. IEEE Signal Process. Mag. **15**(6), 74–90 (1998)
4. PA Chou, M Zhourong, Rate-distortion optimized streaming of packetized media. IEEE Trans. Multimed. **8**(2), 390–404 (2006)
5. 3GPP Wideband Code Division Multiple Access (WCDMA), http://www.3gpp.org/article/w-cdma
6. HDR System Emulator 2, http://hse.qualcomm.com
7. WARP: Wireless Open-Access Research Platform, Rice University, http://warp.rice.edu
8. D Pisinger, An exact algorithm for large multiple knapsack problems. Eur. J. Oper. Res. **114**(3), 528–541 (1999)
9. C Chekuri, S Khanna, A PTAS for the multiple knapsack problem. in *ACM-SIAM Symposium on Discrete Algorithms*. (San Francisco, 213–222, 2000)
10. A Dua, CW Chan, N Bambos, J Apostolopoulos, Channel, deadline, and distortion (CD2) aware scheduling for video streams over wireless. IEEE Trans. Wirel. Commun. **9**(3), 1001–1011 (2010)
11. F Fu, M van der Schaar, A systematic framework for dynamically optimizing multi-user wireless video transmission. IEEE J. Sel. Areas Commun. **28**(3), 308–320 (2009)
12. Y Li, A Markopoulou, N Bambos, J Apostolopoulos, Joint power-playout control for media streaming over wireless links. IEEE Trans. Multimed. **8**(4), 830–843 (2006)
13. GJ Foschini, Z Miljanic, A simple distributed autonomous power control algorithm and its convergence. IEEE Trans. Veh. Technol. **42**(4), 641–646 (1993)
14. J Huang, Z Li, M Chiang, AK Katsaggelos, Joint source adaptation and resource allocation for multi-user wireless video streaming. IEEE Trans. Circuits Syst. Video Technol. **18**(5), 582–595 (2008)

15. Y Li, Z Li, M Chiang, RA Calderbank, Content-aware distortion-fair video streaming in congested networks. IEEE Trans. Multimed. **11**(6), 1182–1193 (2009)
16. X Zhu, B Girod, Distributed media-aware rate allocation for wireless video streaming. IEEE Trans. Circuits Syst. Video Technol. **20**(11), 1462–1474 (2010)
17. J Chakareski, PA Chou, Application layer error-correction coding for rate-distortion optimized streaming to wireless clients. IEEE Trans. Commun. **52**(10), 1675–1687 (2004)
18. FP Kelly, AK Maulloo, DKH Tan, Rate control for communication networks: shadow prices, proportional fairness and stability. J. Oper. Res. Soc. **49**(3), 237–252 (1998)
19. M Chiang, SH Low, RA Calderbank, JC Doyle, Layering as optimization decomposition: a mathematical theory of network architectures. Proc. IEEE. **95**(1), 255–312 (2007)
20. SH Low, A duality model of TCP and queue management algorithms. IEEE/ACM Trans. Netw. **11**(4), 525–536 (2003)
21. SH Low, DE Lapsley, Optimal flow control I: basic algorithm and convergence. IEEE/ACM Trans. Netw. **7**(6), 861–874 (1999)
22. DP Palomar, M Chiang, A tutorial on decomposition methods for network utility maximization. IEEE J. Sel. Areas Commun. **24**(8), 1439–1451 (2006)
23. DP Bertsekas, *Dynamic Programming and, Optimal Control*. 3rd edn., Vol. I. (Belmont, MA: Athena Scientific, 2005)
24. N Bhushan, C Lott, P Black, R Attar, YC Jou, M Fan, D Ghosh, J Au, CDMA2000 1xEV-DO Revision A: a physical layer and mac layer overview. IEEE Commun. Mag. **44**(2), 37–49 (2006)
25. C Lott, N Bhushan, D Ghosh, R Attar, J Au, M Fan, Reverse traffic channel mac design of CDMA2000 1xEV-DO Revision A system. in *IEEE Vehicular Technology Conference,* vol. 3, (Stockholm, Sweden, 1416–1421, 2005)
26. P Tinnakornsrisuphap, C Lott, On the fairness and stability of the reverse link MAC layer in CDMA2000 1xEV-DO. in *IEEE International Conference on Communications,* vol. 1, (Hong Kong, 144–148, 2004)
27. A Bovik, *The Essential Guide to Video Processing*. (New York: Elsevier, 2009)
28. R Karp, Reducibility among combinatorial problems. in *Complexity of Computer Computations*, ed. by Miller RE and Thatcher JW. (New York: Plenum Press, 85–103, 1972)
29. P Hande, S Rangan, M Chiang, X Wu, Distributed uplink power control for optimal SIR assignment in cellular data networks. IEEE/ACM Trans. Netw. **16**(6), 1420–1433 (2008)
30. S Martello, P Toth, Heuristic algorithms for the multiple knapsack problem. Computing. **27**(2), 93–112 (1981)
31. Foreman video, http://media.xiph.org/video/derf/

doi:10.1186/1687-1499-2012-217
Cite this article as: Pandit *et al.*: **Content aware optimization for video delivery over WCDMA.** *EURASIP Journal on Wireless Communications and Networking* 2012 **2012**:217.