Link-Condition Based Proxies for QoS Management in Wireless Networks

Kin K. Leung Zhimei Jiang
AT&T Labs - Research
Red Bank, NJ 07701
{kkleung,jiang}@research.att.com
(732)345-3153,3156

Abstract— **In this paper, we propose and analyze the performance of proxy functions for controlling the quality of service in wireless networks, where the proxy performs aggressive content reduction as a means to throttle traffic when the radio link is congested. A queueing model is constructed and solved to study the performance tradeoffs among system parameters and in particular, how the proxy performance is affected by the feedback control delay between base stations and the proxy.**

Using image compression as an example, we examine response time and fraction of compression as a function of the control delay. Our results reveal that the proxy function can effectively control response time in the case of link congestion, if the delay is reasonably small (e.g., when the proxy is located close to the radio link). We also find that to assess the effectiveness of the control mechanism, it is necessary to examine whether the proxy performs content reduction when and only when link congestion occurs. Towards this end, we study the system-state probabilities and the correlation between response time and compression.

I. Introduction

With the promise of fast and convenient data access, the 3rd generation (3G) wireless networks unveil a bright prospect for ubiquitous computing. Mobile users will soon be able to connect to the Internet and their office computer systems from anywhere at anytime. Many researchers and companies have been exploring and implementing various advanced techniques to improve the capacity and performance of the 3G networks. Despite all these efforts, the scarcity of radio resources will continue to impose significant limitations on the wireless access experience, especially in terms of data rate. Therefore, alternative methods must be pursued to enhance user experience in wireless networks. One promising approach for achieving this goal is to place proxies on the path between the content server and mobile users, as shown in Fig. 1.

Proxies can perform a wide variety of functions in wireless networks. They range from providing better security, caching requested information for future use,

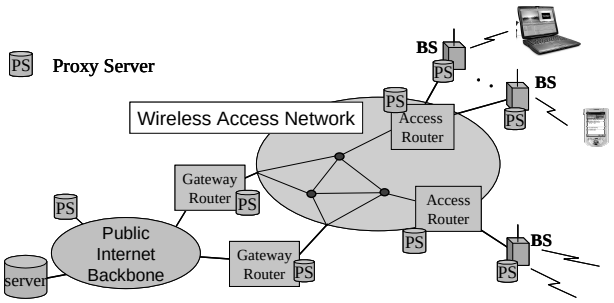

Fig. 1. An illustration of a cellular network with proxies located at various points of the network.

to converting data into a different format based on network and device capabilities, as well as improving performance of protocols such as TCP [1], [2], [8], [10], [14], [18]. Many of these proxy functions are designed to deal with the radio impairment of the wireless networks and the heterogeneous characteristics and capabilities of mobile terminals [4], [11], [13], [21]. In particular, the content reduction proxies enable efficient use of network resources and improving the quality of service (QoS) by reducing the amount of data delivered to the mobile terminals whenever it is needed [7]. Fig. 2 illustrates an example of such proxies, where the proxy converts large high resolution images into small low resolution ones, which are more suitable for transmission over the radio and for display on the mobile device. Since the radio link is often the "weakest" section (in terms of bandwidth availability and reliability of data delivery) of the communication path, we focus on the issue of using proxies to manage the QoS over the radio link in this paper.

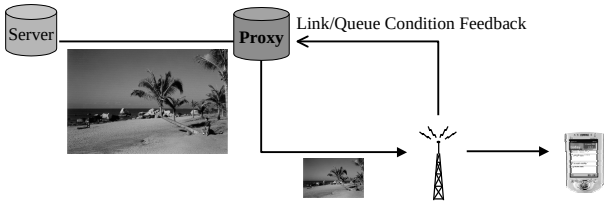

Fig. 2. Image transcoding proxy for cellular users.

A. Link-condition based proxy functions

For any content reduction proxy, two main factors should be considered in determining the desirable degree of data reduction: response time and resulting data quality. First, the degree of data compression must be chosen to meet the radio bandwidth available to the receiving terminal and its capabilities. This factor for transcoding proxy has been pointed out by other researchers [5]. The QoS, as perceived by the users, can be further improved if the transcoding operation is dynamically adjusted according to network conditions. That is, whenever the radio link is congested, the proxy reduces the image quality aggressively, resulting into a large data reduction, as a means for reducing the traffic demand on the radio link. On the other hand, if the link is lightly loaded, the proxy marginally reduces the image quality or even keep the original high resolution images without any data removal.

The second factor needs to be considered in determining the degree of content reduction is the resulting data quality [6]. Content reduction often results in degradation of data quality, thus it needs to be justified by improvements of other performance factors, especially the reduced response time. This further requires the proxy operation to be based on network conditions.

Having proxies that dynamically adjust their operation based on network conditions is particularly important in 3G networks. The reasons are two-fold. First, a radio link will be shared by multiple users in 3G networks [12]. When mobile users' demand for data approaches the link capacity, the response-time performance can degrade significantly. By reducing traffic through image compression based on the network conditions, the proxy will help maintaining satisfactory delay in the system. (It is worth-noting that the use of content reduction to improve response time with slight degradation of data quality in the IP networks draws upon a similar principle for the bit-dropping for packet voice applications in the case of congestion [9], [19].)

Secondly, network-condition based content reduction also serves the purpose of protecting the scared radio resources. This is unique in cellular networks, as the interference caused by excessive traffic can dramatically reduce the efficiency of the system, and further intensify network congestion. The control mechanism offered by channel-condition based content-reduction proxies is very desirable in maintaining network stability and performance. To our knowledge, although proxy functions and architecture for wireless networks have been studied extensively, the issue of how the functions can adapt according to the overall network conditions to control QoS for mobile users has received little research attention and this is the topic addressed in this paper. Without loss of generality, we shall continue to use the example of image compression and, for simplicity, we refer to the content-reduction operation as compression in our discussion.

B. Proxy location and performance

Another issue that has strong impact on proxy functions and their performance is the location of the proxy. As depicted in Figure 1, the proxy can be placed somewhere inside the Internet, which would be very far away from the radio link; or they can be located close to gateway routers, access routers, even base stations (BS's). There are pros and cons in placing proxies at these various locations. The main factors need to be considered include proxy performance, system complexity and flexibility, security, and mobility [15], [17]. Nevertheless, from pure proxy performance viewpoint, since the radio link is often the "bottleneck" of the communication path, it appears to be advantageous to place the proxy close to the BS's so that the proxy operation can adapt responsively to the condition changes of the radio link. The degree of "closeness" is reflected by the control delay, which is defined as from the time when the link congestion is detected until the time when base-station starts to receive compressed packets from the proxy.

Now, the main issue becomes how to measure the proxy performance and how the performance is affected by its location. As to be discussed in detail below, the effectiveness of the network-conditions based proxy functions cannot be reflected by the delay performance and image quality alone. It is important to make sure that the proxy is indeed performing the compression operation only when the link is congested. We shall use a queueing model to study the issues and performance tradeoffs in this paper.

Results from this paper can be utilized in two different ways in dealing with proxy placement problems.
1. Suppose the system has the flexibility of placing proxies at any location. Then, given a desired performance level for the system, our results will reveal the appropriate range of the control delay required, and help determine where the proxy should be placed.
2. Suppose there are certain restrictions on where a proxy function may be located. Results from this pa-

per will help us understand what kind of performance can be achieved and what the performance limitations are at the given location. Based on this understanding, we can then design other components of the system accordingly, to improve the overall system performance.

The rest of this paper is organized as follows. The next section presents a specific algorithm for image compression function at the proxy based on link-congestion conditions. Section 3 details the analytical model and a solution technique for solving the performance measures of interest. Numerical results are presented in Section 4, followed by our conclusion and future work in Section 5.

II. A control algorithm for link-condition based proxies

In this section, we present a specific algorithm to illustrate how proxy server can be used to manage quality of service (QoS) in the case of traffic congestion in wireless IP networks. Although we continue to use the above image compression example for the rest of the paper, our discussions are clearly applicable to other link-condition based content reduction proxies.

Since downlink traffic (from base station to mobile terminals) is typically much higher than uplink traffic (from mobile terminals to the base station), we only consider the radio downlink of the wireless network in this paper. Fig. 3 presents a schematic diagram of the control mechanism described as follows. In this system, mobile terminals request image files from remote servers via the uplink (not shown in the figure). Servers respond by sending the files to the proxy. For simplicity, assume that the main function of the proxy is to compress the image files when instructed by the base station. After possible compression by the proxy, image files are forwarded to the base station for final delivery to the terminals over the radio link. The base station periodically monitors the amount of data (to be referred as to the amount of work and denoted by Q) pending for transmission by the downlink, and compares it with a pre-specified high threshold T_H and low threshold T_L. If $Q \geq T_H$, then the base station sends a control message to instruct the proxy to compress image files until further notice. A random, non-zero delay is incurred from the time when the message is sent until that when the proxy responds to the instruction. This delay is referred to as the control delay (denoted by D) in the following. On the other hand, if the comparison indicates that $Q \leq T_L$, then the base station sends another message to the proxy to stop further image compression.

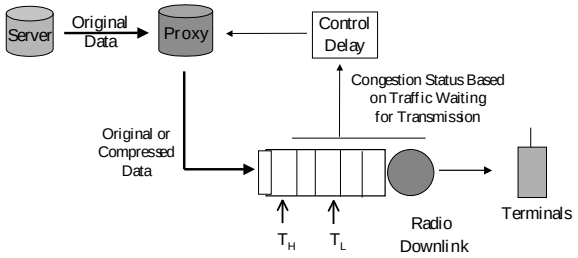

Fig. 3. Analytical model for link-condition based proxies.

It is instructive to understand the relationship between the amount of work for the downlink and its congestion states. As shown in Fig. 4, the downlink status alternates between uncompression and compression period. Image files arriving at the downlink during uncompression and compression period are original (uncompressed) and compressed images respectively. The downlink has four congestion states indexed from 0 to 3 and the state is changed from state i to $i+1$ in a cyclic manner, where it is understood that state $i+1$ becomes 0 when $i+1=4$. These four states are needed to capture the system dynamics with the non-zero delay D. The relationship between the congestion states and uncompression/compression period is described as follows. Assume the downlink is currently in state 3. Once the base station detects that the amount of work Q drops to the low threshold T_L, the downlink enters state 0 and the base station sends a control message to instruct the proxy to stop compression. To consider the effects of the non-zero delay D, the downlink stays in state 0 and enters state 1 at the end of the control delay. As a result, while in state 0, all image files arriving at the downlink are still compressed files as the proxy has not responded to the control message, although the amount of work for the downlink has dropped below T_L.

During state 1, the amount of work for the downlink continues to increase or decrease, depending on the amount of new work arriving and the data transmitted by the radio link. Nevertheless, all arrivals are uncompressed in state 1. When there is a large burst of image arrivals, the amount of work at the base-station may reach the high threshold T_H. At that point, the downlink enters congestion state 2, and the base station sends another control message to request the proxy to resume compression. Again, the purpose of state 2 is to capture the impacts due to the control delay. Clearly, all arrivals at state 2 continue to be uncompressed files because the proxy has not responded

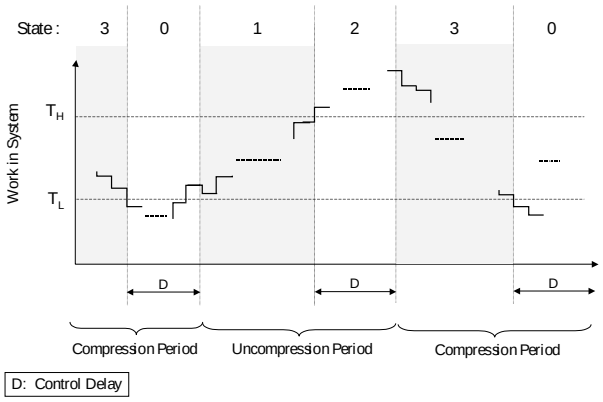

Fig. 4. System state transition diagram.

to the compression message yet. The combined time period in state 1 and 2 corresponds to the uncompression period during which arrivals at the downlink are original (uncompressed) image files.

Similar to the transition from state 0 to 1, the congestion state for the downlink changes from state 2 to 3, when the control message starts to take effect, i.e. when the arrivals become compressed. Compressed files continue to arrive during state 3. When the amount of work drops to T_L, the downlink leaves state 3 and enters state 0, and then the whole process starts over again. The time period while the downlink is in state 3 and 0 is referred to as the compression period as all arrivals at the downlink during the time period are compressed.

Using this algorithm, the proxy is able to compress image files when the radio downlink is congested. This way, the data compression helps reduce the traffic load for the downlink during congestion as compressed files require less service (transmission) time than for uncompressed files statistically. As a result, the QoS in terms of file delivery response time can be improved. Clearly, the control delay D depends on the actual location of the proxy and other system details, such as message transport and processing delay in the network. As explained below, the control delay turns out to be a key parameter in determining the effectiveness of the QoS control mechanism. In addition, there exists performance tradeoffs between response time and image quality. We introduce a queueing model and solution technique to study such performance characteristics and tradeoffs as follows. To our knowledge, the model has not been solved before.

III. Analytical model and solution

To start, let us present our assumptions for the queueing model in Fig. 3:

1. Depending on the instruction from the base station, the proxy compresses image files arriving from the file server or simply forwards the files to the downlink for transmission without compression. Let the probability generating function (PGF) for the service time at radio downlink for compressed and uncompressed files be $X_c(z)$ and $X_u(z)$ respectively, where the service time is assumed to be a positive integer. The compression operation is reflected by the service time chosen from $X_c(z)$ or $X_u(z)$ for the associated files.

2. The downlink is modeled as a discrete-time, single-server queue, where image files are transmitted on a first-come-first-served basis. In this model, time is divided into slots. The actual length of each time slot is properly chosen so that each slot represents the desirable, smallest granularity for the amount of work in the queue. Furthermore, the product of the file arrival rate (in terms of number of files per slot) and the average service time for a compressed file is strictly less than 1 so that the queueing model can reach a steady state.

3. The number of image files arriving at the downlink queue in each time slot is characterized by a general probability distribution. For technical reasons, file arrivals are assumed to actually occur immediately after the beginning of a time slot. Without loss of generality, we assume Poisson distribution for file arrivals during each slot in our numerical results.

4. Let Q_n and S_n be the amount of work in the downlink queue (in unit of the number of time slots) and the congestion state at the end of slot n, respectively. Change of congestion state S_n is assumed to occur only at the end of a slot. For modeling tractability, the possible change of congestion state for slot n is assumed to be based on Q_{n-1} of slot $n-1$. In essence, this causes the control delay to last at least for one time slot.

5. The control delay D is independent of the amount of work for the downlink. Further, for simplicity, D is assumed to have a geometric distribution with α being the probability that the delay expires in one time slot.

We remark that these assumptions are made mainly for tractability reasons and they can be relaxed to some extent by introducing additional model complexity. In particular, it is well understood that the packet

arrival process is non-renewal, and has long-range dependence in Internet environment. However, results in [3] reveal that the long-range dependence is mainly due to packets of certain applications such as those with periodic traffic patterns, whereas traffic associated with other applications (such as those with asynchronous access and the image download considered here may be an example) do not process such high autocorrelation property. Thus, in addition to model tractability, our assumption of renewal arrival process by Assumption 3 could be justified for the Web download application. Nevertheless, the modeling assumptions are reasonable as our primary goal here is to examine the first-order performance tradeoffs among system parameters.

A. Model analysis

To analyze this model, let us start by characterizing the amount of work that arrives in one time slot. Based on Assumption 3, let c_i for $i \geq 0$ be the probability that i image files arriving in a time slot. Let $C(z) = \sum_{i=0}^{\infty} c_i z^i$. In addition, we use A and B to denote the random amount of work arriving in a slot during a normal (uncompression) and compression period respectively. The corresponding PGF's for A and B are $A(z)$ and $B(z)$. By standard arguments, we have

$$A(z) = \sum_{i=0}^{\infty} c_i [X_u(z)]^i = C(X_u(z)) \tag{1}$$

and

$$B(z) = \sum_{i=0}^{\infty} c_i [X_c(z)]^i = C(X_c(z)) \tag{2}$$

where $X_u(z)$ and $X_c(z)$ are the PGF for the service time for the uncompressed and compressed files. As stated in Assumption 3, our numerical examples assume Poisson arrival distribution. In this case, (1) and (2) become

$$A(z) = e^{-\lambda(1-X_u(z))} \tag{3}$$

and

$$B(z) = e^{-\lambda(1-X_c(z))}. \tag{4}$$

Let us observe the system state defined by (Q_n, S_n) at the end of each time slot n. Since the arrival process is renewal and the control delay has a geometric distribution, it is clear that $\{(Q_n, S_n), n \geq 1\}$ form a Markov chain. Let $p_n(i,j) = P[Q_n = i \wedge S_n = j]$ and $F_n(z_1, z_2) = E[z_1^{Q_n} z_2^{S_n}]$. In addition, we define $E_n = 0$ if the control delay does not expire in slot n, and 1 otherwise. To solve the Markov chain at steady state, we first derive a recursive relationship between $F_n(z_1, z_2)$ and $F_{n-1}(z_1, z_2)$. Then the probabilities $p_n(i,j)$'s are obtained by an iterative method using discrete Fourier transforms [16].

By considering all the possible system states for slot $n-1$ and using conditional probability, we have

$$\begin{aligned}
F_n(z_1, z_2) = \\
&E[z_1^{(Q_{n-1}-1)^+ + B} z_2^0] P[S_{n-1} = 0 \wedge E_n = 0] + \\
&E[z_1^{(Q_{n-1}-1)^+ + B} z_2^1] P[S_{n-1} = 0 \wedge E_n = 1] + \\
&E[z_1^{(Q_{n-1}-1)^+ + A} z_2^1] P[S_{n-1} = 1 \wedge Q_{n-1} < T_H] + \\
&E[z_1^{(Q_{n-1}-1)^+ + A} z_2^2] P[S_{n-1} = 1 \wedge Q_{n-1} \geq T_H] + \\
&E[z_1^{(Q_{n-1}-1)^+ + A} z_2^2] P[S_{n-1} = 2 \wedge E_n = 0] + \\
&E[z_1^{(Q_{n-1}-1)^+ + A} z_2^3] P[S_{n-1} = 2 \wedge E_n = 1] + \\
&E[z_1^{(Q_{n-1}-1)^+ + A} z_2^3] P[S_{n-1} = 3 \wedge Q_{n-1} > T_L] + \\
&E[z_1^{(Q_{n-1}-1)^+ + A} z_2^0] P[S_{n-1} = 3 \wedge Q_{n-1} \leq T_L]
\end{aligned} \tag{5}$$

where $(x)^+ = x$ if $x > 0$ and 0 otherwise. The first term on the RHS of (5) corresponds to the condition that the congestion state for slot $n-1$ is 0 and the control delay has not expired in slot n as indicated by $E_n = 0$. In this case, the amount of work at the end of slot n is equal to that at slot $n-1$, Q_{n-1}, minus one (that is just transmitted by the downlink, if any) plus the new work arrival B. Since the control delay has not expired, the congestion state for slot n remains in state 0. The net result is captured by the first term. Other terms in (5) have similar physical interpretation.

As the control delay is independent of the amount of work in the queue as in Assumption 5, after some algebraic manipulation, (5) becomes

$$\begin{aligned}
F_n(z_1, z_2) = \\
&B(z_1) \ [1 - \alpha + \alpha z_2] \\
&\qquad [p_{n-1}(0,0) + \frac{1}{z_1} \sum_{i=1}^{\infty} z_1^i p_{n-1}(i,0)] + \\
&A(z_1) z_2 \ [p_{n-1}(0,1) + \frac{1}{z_1} \sum_{i=1}^{T_H - 1} z_1^i p_{n-1}(i,1) + \\
&\qquad \frac{z_2}{z_1} \sum_{i=T_H}^{\infty} z_1^i p_{n-1}(i,1)] + \\
&A(z_1) z_2^2 \ [1 - \alpha + \alpha z_2] \\
&\qquad [p_{n-1}(0,2) + \frac{1}{z_1} \sum_{i=1}^{\infty} z_1^i p_{n-1}(i,2)] +
\end{aligned}$$

$$B(z_1)\ \ [p_{n-1}(0,3) + \frac{1}{z_1}\sum_{i=1}^{T_L-1} z_1^i p_{n-1}(i,3) + \frac{z_2^3}{z_1}\sum_{i=T_L}^{\infty} z_1^i p_{n-1}(i,3)] \quad (6)$$

where $A(z_1)$ and $B(z_1)$ are given by (1) and (2), respectively. As for (5), one can provide the physical interpretation for each term on the RHS of (6). For example, the last factor of the first term represents the PGF of the amount of work in the downlink queue after transmitting one time slot of work. Since the congestion state is 0, the amount of new work associated with compressed files arriving during slot n is $B(z_1)$, which is given as the first factor. Thus, the product of the first and last factor yields the PGF for the amount of work in the queue at the end of slot n. The second factor in the first term reflects a possible change from the current congestion state of 0 to 1, if the control delay expires with probability α in slot n. Otherwise, with probability $1-\alpha$, the congestion state remains at state 0. Other terms in (6) can be interpreted in a similar way.

It is worth-noting that since $F_n(z_1, z_2)$ is the PGF of the probabilities $p_n(i,j)$'s, they are one-one corresponding. Thus, one can see that (6) actually represents a recursive relationship between $F_n(z_1, z_2)$ and $F_{n-1}(z_1, z_2)$. That is, given the probabilities $p_{n-1}(i,j)$'s for slot $n-1$, substituting them into (6) yields the PGF $F_n(z_1, z_2)$ of all $p_n(i,j)$'s. In other words, we can find the time-dependent state probabilities as the recursion is executed once for each time slot. Furthermore, since the queueing model has a steady state by Assumption 2, when $n \to \infty$, $F_n(z_1, z_2) = F_{n-1}(z_1, z_2)$. Thus, using (6) as a basis for iterations, we can apply a technique based on discrete Fourier transforms [16] to solve for all probabilities $p_n(i,j)$'s with $n \to \infty$ as follows.

B. Solution by Discrete Fourier Transforms

Since the queueing model has a steady state, one can estimate the "maximum" amount of work for the downlink to be N (if the chosen N value is large enough, the $p_n(i,j)$'s should be very small, e.g. on the order of 10^{-6}, for all i close to N and $j = 0$ to 3). Then, we closely approximate $F_n(z_1, z_2)$ by

$$F_n(z_1, z_2) \approx \sum_{i=0}^{N}\sum_{j=0}^{3} z_1^i z_2^j p_n(i,j). \quad (7)$$

Since the number of all possible system states (Q_n, S_n) now becomes finite due to the truncation of state space, we can use discrete Fourier transforms (DFT's) to represent $F_n(z_1, z_2)$ for computation. To obtain the DFT's, we define $\omega_1 = e^{-2\pi j/(N+1)}$ and $\omega_2 = e^{-2\pi j/4}$ where $j = \sqrt{(}-1)$. (Note that j is also used an integer index.) Let $\{F_n^*(k_1,k_2)|k_1 = 0$ to N, $k_2 = 0$ to $3\}$ be the DFT's corresponding to $F_n(z_1, z_2)$. By the definition of DFT, for all $k_1 = 0$ to N and $k_2 = 0$ to 3, we have

$$F_n^*(k_1,k_2) = \sum_{i=0}^{N}\sum_{j=0}^{3} \omega_1^{ik_1}\omega_2^{jk_2} p_n(i,j). \quad (8)$$

Similarly, the DFT's associated with $A(z)$ and $B(z)$ in (1) and (2) can be defined and found as $A^*(k_1)$ and $B^*(k_1)$, respectively. As a result, one can convert the recursive relationship in PGF in (6) into one in terms of DFT's as follows:

$$\begin{aligned} F_n^*(k_1,k_2) = \\ B^*(k_1)\ \ & [1-\alpha+\alpha\omega_2^{k_2}] \\ & [p_{n-1}(0,0) + \frac{1}{\omega_1^{k_1}}\sum_{i=1}^{\infty}\omega_1^{ik_1}p_{n-1}(i,0)] + \\ A^*(k_1)\omega_2^{k_2}\ & [p_{n-1}(0,1) + \frac{1}{\omega_1^{k_1}}\sum_{i=1}^{T_H-1}\omega_1^{ik_1}p_{n-1}(i,1) + \\ & \frac{\omega_2^{k_2}}{\omega_1^{k_1}}\sum_{i=T_H}^{\infty}\omega_1^{ik_1}p_{n-1}(i,1)] + \\ A^*(k_1)\omega_2^{2k_2}\ & [1-\alpha+\alpha\omega_2^{k_2}] \\ & [p_{n-1}(0,2) + \frac{1}{\omega_1^{k_1}}\sum_{i=1}^{\infty}\omega_1^{ik_1}p_{n-1}(i,2)] + \\ B^*(k_1)\omega_2^{3k_2}\ & [p_{n-1}(0,3) + \frac{1}{\omega_1^{k_1}}\sum_{i=1}^{T_L-1}\omega_1^{ik_1}p_{n-1}(i,3) + \\ & \frac{\omega_2^{3k_2}}{\omega_1^{k_1}}\sum_{i=T_L}^{\infty}\omega_1^{ik_1}p_{n-1}(i,3)]. \end{aligned} \quad (9)$$

Furthermore, one can choose and substitute an initial solution for the probabilities $p_0(i,j)$'s with $i = 0$ to N and $j = 0$ to 3 (e.g., those corresponding an empty queue) in (9) to obtain the DFT's $F_1^*(k_1,k_2)$. Naturally, inverting these DFT's yields the probabilities $p_1(i,j)$'s. This process is repeated by substituting the new results into (9) as the argument again. Since the model has a steady state, after a sufficiently large number of iterations, probabilities $p_n(i,j)$'s converge to the steady-state solution, as one would expect. A convergent criterion can be that the corresponding new and old probabilities differ less than a very small number (e.g., 10^{-7}). Once these probabilities are obtained, other important performance measures such as

the response-time distribution and fraction of image files compressed become known as discussed below.

C. Response Time & Fraction of Image Compression

To obtain the response-time distribution, we define

$$\hat{F}_n(z_1, j) = \sum_{i=0}^{\infty} z_1^i p_n(i, j). \quad (10)$$

Clearly, $\hat{F}_n(z_1, j)$ is the PGF for the amount of work for the downlink, given that its congestion state is j at the end of slot n.

Recall that the probability of i file arrivals in a slot is c_i. According to the batch-size biasing argument [20], the probability of a randomly selected files being the k^{th} file arrival in a slot is given by $\frac{1}{\bar{c}}\sum_{i=k}^{\infty} c_i$, where $\bar{c}$ is the average number of file arrivals per slot. Since the arrival process is independent of the congestion state and the service time is chosen from the PGF's $X_u(z)$ or $X_c(z)$, when the file arrives during a normal (uncompressed) and compression period. In either case, the response time (from the arrival until its transmission completion by the downlink) of a tagged image file is the sum of the amount of work in the queue at the beginning of the slot at which the file arrives and the service time for those files that arrive in the same slot but prior to the tagged file. In terms of PGF, the response time for an arbitrary file is thus given by

$$\begin{aligned} T(z_1) = \\ [\hat{F}(z_1, 0) + \hat{F}(z_1, 3)] \textstyle\sum_{k=1}^{\infty} [X_c(z)]^k \frac{1}{\bar{c}} \sum_{i=k}^{\infty} c_i + \\ [\hat{F}(z_1, 1) + \hat{F}(z_1, 2)] \textstyle\sum_{k=1}^{\infty} [X_u(z)]^k \frac{1}{\bar{c}} \sum_{i=k}^{\infty} c_i \end{aligned} \quad (11)$$

where $\hat{F}(z_1, j) = \lim_{n\to\infty} \hat{F}_n(z_1, j)$ for $j = 0$ to 3.

After some algebraic manipulation, (11) becomes

$$\begin{aligned} T(z_1) = \\ [\hat{F}(z_1, 0) + \hat{F}(z_1, 3)] \tfrac{1}{\bar{c}} \tfrac{X_c(z_1)[1 - C(X_c(z_1))]}{1 - X_c(z_1)} + \\ [\hat{F}(z_1, 1) + \hat{F}(z_1, 2)] \tfrac{1}{\bar{c}} \tfrac{X_u(z_1)[1 - C(X_u(z_1))]}{1 - X_u(z_1)}. \end{aligned} \quad (12)$$

Since $\hat{F}(z_1, j) = \lim_{n\to\infty} \sum_{i=0}^{N} z_1^i p_n(i, j)$ and all $p_n(i, j)$'s at steady state have been obtained from the iterations, we can obtain the DFT's $\hat{F}^*(k_1, j)$ corresponding to $\hat{F}(z_1, j)$ by

$$\hat{F}^*(k_1, j) = \lim_{n\to\infty} \sum_{i=0}^{N} \omega_1^{ik_1} p_n(i, j) \quad (13)$$

for all $k_1 = 0$ to N and $j = 0$ to 3. Similarly, the DFT's for $C(X_u(z_1))$ and $C(X_c(z_1))$ can be computed. Finally, the DFT's associated with $T(z_1)$ are obtained. Inverting the latter DFT's yields the probability distribution for the file response time.

Let the fraction of files compressed by the proxy be denoted by β. By definition, the probability of staying in congestion state j at steady state is $\lim_{n\to\infty} \sum_{i=0}^{N} p_n(i, j)$, which is given by $\hat{F}^*(0, j)$ in (13). As discussed above, the arrival process is independent of the congestion state. Hence, we have $\beta = \hat{F}^*(0, 0) + \hat{F}^*(0, 3)$, which is known as the last two terms have been found by (13).

Let us consider the mean time between sending two control messages to reflect the overhead of the control mechanism. We observe that two control messages are sent during the time period for the link to go through the whole cycle of 4 congestion states once. Let the average length of such a time period be $\bar{K}$. As $\bar{K}$ includes two average control delay $\bar{D}$, by renewal theory we have $\beta = 2\bar{D}/\bar{K}$. Since two control messages are sent during $\bar{K}$, the mean time between two control messages is $\bar{K}/2$, which is given by $\bar{D}/\beta$, where β has been found above.

IV. Numerical results and discussions

We now use the queueing model described above to illustrate the possible performance tradeoffs for the proxy function under consideration. In our numerical examples, the average service time for a original (uncompressed) and compressed files are 10 and 7 time slots, respectively. Unless stated otherwise, the thresholds T_L and T_H are set to be 20 and 25 respectively. These parameters are actually derived for a possible operation scenario in the EDGE system [12] and each time slot corresponds to 20 msec. The details of how the parameters are generated are omitted here due to space limitation. In the following discussion, we simply use D to stand for the average control delay, which has a geometric distribution by Assumption 5.

A. Response Time Performance

The primary goal of having content reduction proxies is to reduce the response time involved in downloading data, where the response time is defined as the time from the arrival of an image file (uncompressed or compressed) at the downlink until it is completely transmitted by the link. Fig. 5 shows the average response time for cases with and without the proxy function as a function of file arrival rate at steady state. The average control delay for the cases with proxy is set to 5, 25 and 50 slots respectively. As standard queueing behavior, the figure reveals that

without proxy, the average response time grows very rapidly when the arrival rate approaches 0.09 (i.e., corresponding to 90% link utilization). On the other hand, with proxy, as the traffic load increases, the proxy starts to compress image files. As a result, the reduced traffic load maintains the reasonable response time well beyond the arrival rate of 0.1 where the system without the proxy could have become overloaded and unstable. Clearly, by performing data compression based on link-congestion condition, the proxy can effectively control the response time and improve the system stability significantly. We have also obtained results for 90 percentile response time, which have shapes very similar to those in Fig. 5.

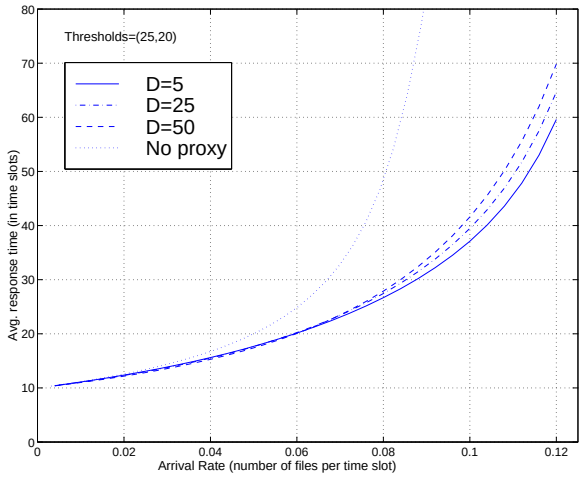

Fig. 5. Average response time Vs. file arrival rates

As one would intuitively expected, when the control delay decreases, the system can react faster to link congestion conditions, thus lowering the response time somewhat. It is interesting to observe from Fig. 5 that although this intuition is correct, the decrease in response time is only marginal even when the average control delay is reduced by one order of magnitude from 50 to 5 time slots. However, one should not conclude at this point that the control delay has little impacts on system performance, as we will see later in this section.

B. Probability of file compression

Another performance measure is the percentage of image files compressed by the proxy. Since data reduction often results in loss in display quality, for a given response time, the less number of files are compressed, the better the proxy performs.

For the same average control delays considered above, Fig. 6 plots the fraction of files compressed by the proxy as a function of file arrival rate at steady state. It shows that, for a given control delay, the percentage of compression increases with arrival rate. This is expected because high traffic loading causes the radio link to be congested, thus triggering image compression (i.e., degraded image) more often. As one can observe from the figure, the biggest discrepancy in the compression fraction for the selected control delays occurs at medium arrival rates. The reason for this is that, at medium loadings, the radio link often moves in and out of the congestion state while its buffer occupancy fluctuates between the threshold T_H and T_L. And increased control delay causes the control mechanism to be less responsive to changes of the congestion state. As the arrival rate further increases, the link constantly stays in the congestion state (and files are compressed most of the time), thus the impacts of the control delay is reduced.

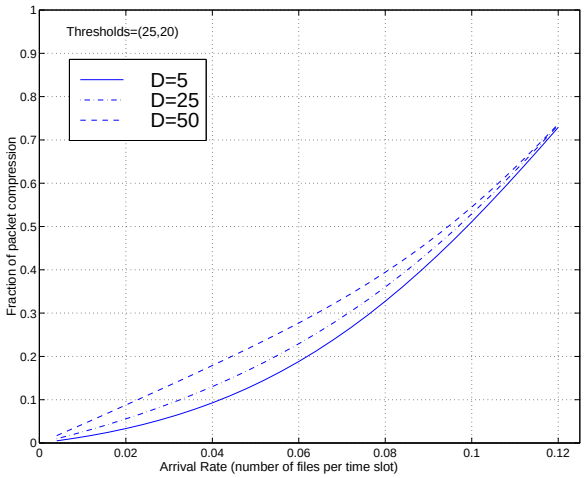

Fig. 6. Compression probability Vs. file arrival rate

It is worth-noting that by comparing the above figure with Fig. 5, we find that with a larger control delay, not only more files experience longer response time, but also more files end up being compressed by the proxy and thus having a lower quality. This may seem to be counter-intuitive at the first glance, since more content reduction means less traffic overall, which could lower the response time. This calls for a closer examination of proxy performance as described next.

C. Probability distribution of system states

We have so far looked at two performance measures separately: response time and fraction of files compressed. As mentioned previously, our results show that when the control delay increases, both re-

sponse time and image quality degrade. What is not captured when considering the performance measures separately is whether the system is doing the right thing in the right situation. In other words, whether the proxy performs data compression when and only when the system is congested. We now examine this aspect of proxy performance by looking at the probability distribution of the congestion states.

As illustrated in Fig. 4, the radio link has four possible congestion states, indexed from 0 to 3. Due to the non-zero control delay, the link stays in state 0 and receives compressed image files, despite the fact that the amount of work for the system has dropped below the threshold T_L. Similarly, the link is in state 2 and receives uncompressed files while the link has been considered to be congested. For this reason, we call state 0 and 2 to be undesirable states. In contrast, state 1 and 3 are referred to as the desirable states because uncompressed or compressed files arrive at appropriate times when the link is not congested or congested, respectively. As one would expect, the more the link stays in the desirable states, the more efficient the proxy performs the QoS control function.

Fig. 7 plots the probability of the system being in the desirable (state 1 or 3) and undesirable (state 0 and 2) states, with an average control delay of 5, 25 and 50 time slots respectively. As expected, when the control delay increases, it takes longer for the proxy to react to the changes of link condition, and the system stays in the undesirable states longer than with a lower control delay. A striking observation from the figure is that the amount of time that the system spends in the undesirable states increases sharply with the average control delay. For instance, with an average delay of 50 slots, the system can stay at the undesirable states nearly 50% of time at certain loads.

For a large control delay, delayed response to compress images by the proxy causes the downlink queue to build up further than with a smaller control delay. Although the proxy eventually starts file compression, it will take a longer time to clear out the queue to below T_L. Thus more files are compressed overall compared to lower control delays. Moreover, the files in the queue experience longer response time. This explains why both compression probability as well as the response time go up with an increasing control delay.

In addition, because the system also spends more time in state 0 for a large control delay, while files are compressed by the proxy when not needed, these files experience lower response time. So overall, increased control delay does not make a big impact on the file response time.

To further confirm these observations, we consider the correlation between the amount of work in the system (denoted by Q) at the end of slot prior to the arrival of an arbitrary file and whether or not it is compressed. Specifically, let a flag $G = 1$ if the file is compressed, and 0 otherwise. Based on the equilibrium probabilities $p_n(i, j)$'s obtained from the model, the coefficient of correlation between Q and G for different arrival rates is computed and plotted in Fig. 8. For any given arrival rate, the correlation coefficient for a smaller average control delay is consistently larger than the one with a larger delay, indicating that files are more likely to be compressed when and only when the system is congested.

D. Control overhead and transient behavior

Another important aspect of system performance is the overhead of sending control messages involved in the mechanism shown in Fig. 3.

Fig. 9 shows the average time between two control messages sent from the BS to the proxy as a function of T_H (where T_L is set to be 5 less than the chosen T_H value), when the file arrival rate is 0.05. As one would expect, these results reveal that, for any given T_H, the mean time between messages increases as the average control delay D does. This is so because a large control delay makes the time period for the link to go through the whole cycle of 4 congestion states once longer. Since there are two control messages sent per cycle (one message upon entering state 0 and 2 respectively), longer cycle time implies longer mean time between control messages. Furthermore, for a given D value, the mean time between messages increases with T_H. Since the file arrival rate is fixed, a larger T_H requires a longer time to reach the congestion state 2 and 3. As a result, the cycle time increases, and so does the mean time between control messages.

Last but not least, the performance model presented above also enables us to examine the transient performance of the proxy server. In particular, we consider an overload condition where the downlink has reached a steady state at time slot 0 with file arrival rate of 0.05. From time 1 to 100, the arrival rate is increased three folds to 0.15 and the rate is re-adjusted back to 0.05 after time 101. Figure 10 shows how the probability of file compression changes in time for the average control delay of 5, 25 and 50, respectively. These results reveal that for small D, the congestion status quickly reaches the proxy. As a

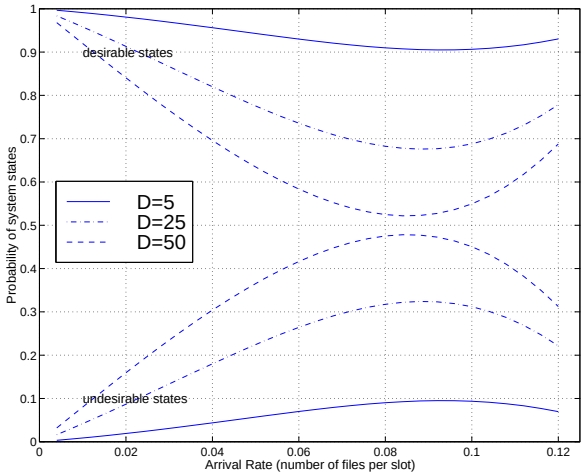

Fig. 7. System states probability Vs. arrival rate

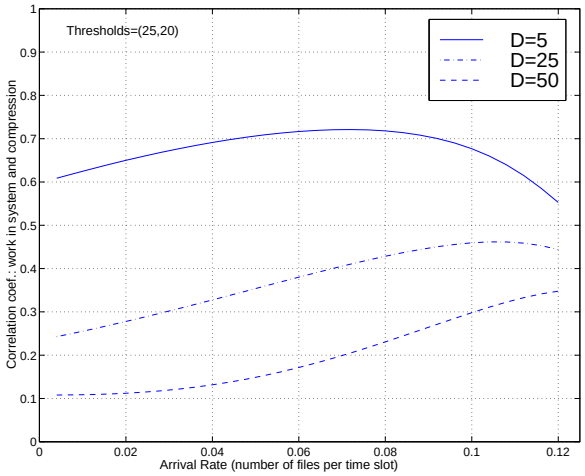

Fig. 8. Correlation coefficient for response time and compression operation.

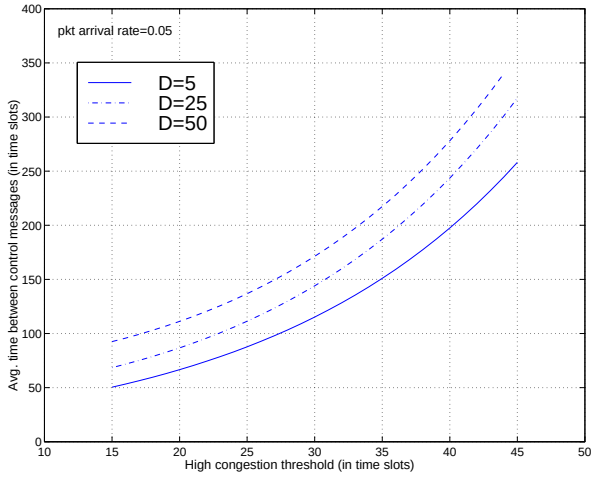

Fig. 9. System control overhead Vs. control thresholds

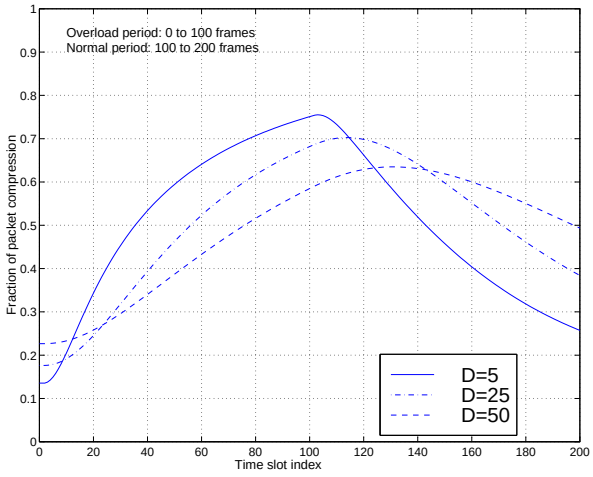

Fig. 10. Transient behavior of image compression

result, the proxy starts to compress image files right away following the beginning of the overload period. Similarly, the proxy also responds quickly by stopping unnecessary compression after time 101 at which the overload period ends. Results in this figure again show that small control delays help the proxy adapt to the network conditions more efficiently.

V. Conclusion

We have proposed and analyzed the performance of a proxy for QoS control in wireless networks, where the proxy performs data/content reduction according to the network conditions. Specifically, when the radio link is congested, the proxy is made to reduce data aggressively, as a means to throttle traffic to control the quality of service.

A new queueing model has been constructed for analyzing and studying the performance tradeoffs of the proxy system. In particular, the model enables us to examine how the performance is affected by the control delay between base station and the proxy. Based on our numerical results for image compression, we found that when compared with no proxy function, the proxy can serve as an efficient control mechanism to improve response time, hence protecting the scared radio resources against excessive traffic load. The control delay turns out to be a key performance parameter. It is observed that the control mechanism can operate efficiently if the delay is reasonable small (e.g., on the order of 100 msec for the 3G wireless networks with EDGE as the air interface [12]).

Comparing results obtained for different control delays, we found that both the response time and the average data quality degrade with an increasing control

delay. More importantly, we also found that studying the response time and compression rate separately for the proxy system is not sufficient for capturing the entire picture regarding proxy performance. Rather, the effectiveness of the control mechanism should be assessed by examining whether the proxy performs data reduction when and only when link congestion occurs. We therefore computed the system-state probabilities and the coefficient of correlation between response time and compression, which reveal that the proxy performance is far more undesirable with a relatively large control delay in that respect. Although we consider the proxy adaptation based on the overall link conditions here, the control delay is expected to have similar performance impacts for cases where the proxy functions are adjusted according to the link quality for individual mobile users.

In cellular networks, the available capacity for the entire downlink and for the individual users fluctuate in time, so does the quality of service as perceived by mobile users. For future work, we plan to examine how proxy functions can provide further help in improving the response time and data quality with this type of channel conditions.

VI. Acknowledgment

We would like to thank Paul Henry for his invaluable comments on this work and thank Li Fung Chang, J. Kim, and Hui Luo for the discussions on proxies.

References

[1] H. Balakrishnan, V. Padmanabhan, S. Seshan, and R. H. Katz, "A comparison of mechanisms for improving TCP performance over wireless links," *IEEE/ACM Transactions on Networking,* vol.5, no.6, Dec. 1997, pp.756-69.

[2] H. Bharadvaj, A. Joshi, and S. Auephanwiriyakul, "An active transcoding proxy to support mobile web access," *Proceedings of 7th IEEE Symposium on Reliable Distributed Systems,* Oct. 1998, pp. 118-23.

[3] V. Bolotin, J. Coombs-Reyes, D. Heyman, Y. Levy and D. Liu, "Traffic Characterization for Planning and Control," *International Teletraffic Congress (ITC-16)*, Edinburgh, Scotland, June 1999.

[4] J. Border, M. Kojo, J. Griner, and G. Montenegro, "Performance enhancing proxies," Internet-draft, http://search.ietf.org/internet-drafts/draft-ietf-pilc-pep-02.txt, March 2000.

[5] E. A. Brewer, et al., "A Network Architecture for Heterogeneous Mobile Computing," *IEEE Personal Communications Magazine*, Oct. 1998, pp. 8-24.

[6] S. Chandar and C. Ellis, "JPEG Compression Metric as a Quality Aware Image Transcoding," *Proceedings of 2nd USENIX Symposium on Internet Technologies and Systems*, Oct. 1999.

[7] C. Chi, J. Deng, and Y.H. Lim, "Compression Proxy Server: Design and Implementation," *Proceedings of 2nd USENIX Symposium on Internet Technologies and Systems*, Oct. 1999.

[8] C. Chien, M. B. Srivastava, R. Jain, P. Lettieri, V. Aggarwal, and R. Sternowski, "Adaptive radio for multimedia wireless links," *IEEE Journal on Selected Areas in Communications*, vol.17, no.5, May 1999, pp. 793-813.

[9] S. Dravida and K. Sriram, "End-to-End Performance Models for Variable Bit Rate Voice Over Tandem Links in Packet Networks," *IEEE Journal on Selected Areas in Communications*, Vol. 7, No. 5, 1989, pp. 718-728.

[10] R. Floyd, B. Housel, and C. Tait, "Mobile Web access using eNetwork Web Express," *IEEE Communications Magazine*, vol.5, no.5, Oct. 1998, pp. 47-52.

[11] A. Fox, S. D. Gribble, Y. Chawathe, and A. E. Brewer, "Adapting to network and client variation using infrastructural proxies: lessons and perspectives," *IEEE Personal Communications Magazine,* vol.5, no.4, Aug. 1998, pp. 10-19.

[12] A. Furuskar, S. Amzur, F. Muller, and H. Olofsson, "EDGE, Enhanced Data Rates for GSM and TDMA/136 Evolution," *IEEE Personal Communications*, vol. 6, no. 3, June 1999, pp. 56-66.

[13] R. Han, P. Bhagwat, R. LaMaire, T. Mummert, V. Perret, and J. Rubas, "Dynamic adaptation in an image transcoding proxy for mobile Web browsing," *IEEE Personal Communications Magazine,* vol.5, no.6, Dec. 1998 pp. 8-17,

[14] C. Y. Hsu, A. Ortega, and M. Khansari, "Rate control for robust video transmission over burst-error wireless channels," *IEEE Journal on Selected Areas in Communications*, vol.17, no.5, May 1999, pp. 756-73.

[15] Z. Jiang, L. Chang, J. Kim, and K. K. Leung, "Incorporating proxy services into wide area cellular IP networks," *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC) 2000*, Sept. 2000.

[16] K. K. Leung, "Cyclic-Service Systems with Probabilistically-Limited Service," *IEEE Journal of Select. Areas on Commun.* , Vol. 9, No. 2, Feb. 1991, pp. 185-193.

[17] B. Li, M. J. Golin, G. F. Italiano, X. Deng, and K. Sohraby, "On the optimal placement of web proxies in the Internet," *Proceedings of INFOCOM'99*, pp.1282-90, Mar. 1999.

[18] S. Sen, J. Rexford, and D. Towsley, "Proxy prefix caching for multimedia streams," *Proceedings of INFOCOM'99*, March 1999, pp. 1310-19.

[19] K. Sriram and D. M. Lucantoni, "Traffic Smoothing Effects of Bit Dropping in a Packet Voice Multiplexer," *IEEE Transactions on Communication*, Vol. 37, No. 7, 1989, pp. 703-712.

[20] R. W. Wolff, *Stochastic modeling and the theory of queues*, Prentice Hall, New Jersey (1989), pp. 68-69.

[21] B. Zenel and D. Duchamp, "General purpose proxies: solved and unsolved problems," *Proceedings of The Sixth Workshop on Hot Topics in Operating Systems*, May 1997, pp. 87-92.