Analysis of the Energy-Response Time Tradeoff for Delayed Mobile Cloud Offloading

Huaming Wu  
Free University of Berlin, Germany  
huaming.wu@fu-berlin.de

Yi Sun  
Free University of Berlin, Germany  
yi.sun@fu-berlin.de

Katinka Wolter  
Free University of Berlin, Germany  
katinka.wolter@fu-berlin.de

ABSTRACT
We develop a delayed offloading model to leverage the complementary strength of WiFi and cellular networks when choosing heterogeneous wireless interfaces for offloading. Optimality analysis of the energy-delay tradeoff is carried out by using a queueing model with impatient jobs and service interruptions, which captures both energy and performance metrics and also intermittently available access links.

1. INTRODUCTION
There are several ways to offload tasks to a dedicated cloud server, e.g., via a cellular (e.g., 2G or 3G) connection or via intermittently available WLAN hotspots. By delaying transmission until a fast and energy-efficient network (e.g., WiFi) becomes available, it is possible to reduce the transmission time even if extra waiting time is introduced, which could lead to energy savings [2]. However, delayed offloading is still a matter of debate, since it is not known to what extent users are willing to delay a transmission [3]. Different applications usually give different relative importance to the factors of response time and energy consumption. We aim at guiding the decision of how to balance the delay and energy savings for different types of scenarios like delay-tolerant applications (e.g., iCloud and dropbox) and delay-sensitive applications (e.g., chess game and face recognition).

2. DELAYED OFFLOADING MODEL
Figure 1 shows a delayed offloading model, which consists of an Offload Queue for data offloading, a Local Queue denoting the local processing on the mobile device and a Remote Queue representing the remote processing on a cloud server. The jobs are offloaded either via a cellular connection or a WiFi network to the cloud. We consider an $M/M/1$ modulated queue in a 2-phase (fast and slow) Markovian timedelay (sometimes pure delay) station, the probability distribution of the delay being that of the service time.

Given the previously stated assumptions, the delayed offloading model can be modeled with a 2D Markov chain, as shown in Fig. 2.
The states with cellular network are denoted with \{c, i\}, and the states with WiFi connectivity are denoted with \{w, i\}. The variable \(i\) corresponds to the number of jobs in the system (queuing and in service). During the WiFi phase, the system drains at rate \(\mu_w\), and during the cellular phase, the system drains at rate \(\mu_i + \gamma\) since any of the \(i\) queued jobs can abandon the Offload Queue [4]. Writing the balance equations for the cellular and WiFi phases gives:

\[
(\lambda + \eta)\pi_{c,0} = (\mu_c + \gamma)\pi_{c,1} + \xi\pi_{w,0} \\
(\lambda + \eta + \mu_c + i\gamma)\pi_{c,i} = \lambda\pi_{c,i-1} + (\mu_c + (i + 1)\gamma)\pi_{c,i+1} + \xi\pi_{w,i} (i > 0) \\
(\lambda + \xi)\pi_{w,0} = \mu_w\pi_{w,1} + \eta\pi_{w,0} \\
(\lambda + \xi + \mu_w)\pi_{w,i} = \lambda\pi_{w,i-1} + \mu_w\pi_{w,i+1} + \eta\pi_{w,i} (i > 0)
\]

The steady-state probability of finding the offloading system in some region without WiFi availability (with only cellular access) is \(\pi_c = \frac{E[T_{OFF}]}{E[T_{ON}]+E[T_{OFF}] = \xi}{\eta}\). Similarly, the steady-state probability for the periods with WiFi is \(\pi_w = \frac{E[T_{ON}]}{E[T_{ON}]+E[T_{OFF}]} = \frac{\eta}{\eta + \xi}\), which equals to the availability ratio \(AR\).

The probability generating functions for both cellular and WiFi states are defined as:

\[
G_c(z) = \sum_{i=0}^{\infty} \pi_{c,i} z^i \quad \text{and} \quad G_w(z) = \sum_{i=0}^{\infty} \pi_{w,i} z^i, \quad |z| \leq 1.
\]

After some calculation and algebraic manipulations in (2), we obtain:

\[
G_w(z)\beta(z) = \eta z G_c(z) - \mu_w (1-z) \pi_{w,0},
\]

where \(\beta(z) = (\lambda - \mu_w) (1-z) + \xi z\). The roots \(z_1, z_2\) of the quadratic polynomial \(\beta(z) = -\lambda (z - z_1)(z - z_2)\) are positive. According to [5], we obtain:

\[
\pi_{c,0} = \frac{S\gamma\kappa_2(1)}{\mu_c(\xi + \eta)(SV - TU)}, \quad \pi_{w,0} = -\frac{T\gamma\kappa_2(1)}{\mu_w(\xi + \eta)(SV - TU)}.
\]

where we define \(S = \int_0^1 \frac{\kappa_1(x)}{\beta(x)} dx\), \(T = \int_0^1 \frac{\kappa_1(x)}{x} dx\), \(U = \int_{z_1}^{z_2} \frac{\kappa_1(x)}{\beta(x)} dx\) and \(V = \int_{z_1}^{z_2} \frac{\kappa_1(x)}{x} dx\). Accordingly, \(\kappa_1(z)\) and \(\kappa_2(z)\) are represented as follows:

\[
\kappa_1(z) = e^{-\frac{\mu_c}{\lambda} z (z - 2)} \frac{\xi (z - 2)}{\xi + \eta} (z - 1) \kappa_2(z - 1)(z - 2) - \frac{\eta}{\xi + \eta} (z - 2), \quad z \leq z_1, \\
\kappa_2(z) = e^{-\frac{\mu_c}{\lambda} z (z - 2)} \frac{\xi (z - 2)}{\xi + \eta} (z - 1) \kappa_2(z - 1)(z - 2) - \frac{\eta}{\xi + \eta} (z - 2), \quad z \geq z_1.
\]

By the definitions of \(\kappa_1(z)\), \(\kappa_2(z)\), and \(\beta(z)\), it follows that \(T, U > 0\) and \(S < 0\). Therefore, \(\pi_{c,0}\) and \(\pi_{w,0}\) are positive. It can be shown formally that the system is ergodic.

Let \(\mu\) be defined as: \(\mu = \pi_c \cdot \mu_c + \pi_w \cdot \mu_w\). According to [5], we obtain:

\[
E[N_c] = \frac{\lambda - \mu + \mu_i\pi_{c,0} + \mu_w\pi_{w,0}}{\gamma}
\]

\[
E[N_w] = \frac{\eta(\lambda - \mu) + \gamma(\lambda - \mu_w) + \mu_i\pi_{c,0} + \mu_w(\eta + \gamma)\pi_{w,0}}{\gamma}
\]

As shown in Fig. 2, the expected number of jobs served per unit of time in the slow phase and the fast phase are \(\mu_i(\pi_c - \pi_{c,0})\) and \(\mu_w(\pi_w - \pi_{w,0})\), respectively. An arbitrary job arriving to the Offload Queue may leave and join the Local Queue due to impatience in the slow phase. Therefore, the rate of jobs executed locally on the mobile device, \(\lambda_m\), is given by:

\[
\lambda_m = \lambda - \mu_c(\pi_c - \pi_{c,0}) - \mu_w(\pi_w - \pi_{w,0}) = \lambda - \mu + \mu_i\pi_{c,0} + \mu_w\pi_{w,0}. \quad (7)
\]

Further, the probability to abandon the queue is defined as:

\[
Pr\{\text{abandonment}\} = \frac{\lambda_m}{\lambda} = \frac{\lambda - \mu + \mu_i\pi_{c,0} + \mu_w\pi_{w,0}}{\lambda}
\]

where \(\Pr\) denotes the probability operation.

4. METRIC-BASED ANALYSIS

4.1 Mean Response Time

The total cost for offloading a job is composed of the cost for sending the job to the cloud and idly waiting for the cloud to complete the job in the slow phase. According to Little’s Law, \(E[N] = \lambda E[T]\), the mean response time can be calculated as:

\[
E[T] = E[T_j] = \sum_j \frac{\lambda_j}{\lambda} E[T_j] = \frac{1}{\lambda} \sum_j E[N_j], \quad (9)
\]

where \(j \in \{c, w, m, r\}\) represents the cellular phase, the WiFi phase, the mobile device and the remote cloud, respectively.

The average number of jobs in the Local Queue and Remote Queue can be calculated as: \(E[N_{cm}] = \rho_{cm}/(1 - \rho_{cm})\) and \(E[N_c] = \lambda_c/\mu_c\), respectively, where \(\rho_{cm} = \lambda_{cm}/\lambda_c\) and \(\lambda_c = \lambda - \lambda_m\) is the arrival rate to the Remote Queue.

4.2 Mean Energy Consumption

Since \(E[P] = \lambda E[\xi]\) is the mean power consumption, we can calculate the mean energy consumption for the delayed offloading model as:

\[
E[\xi] = E[E[\xi_k]] = \sum_k \frac{\lambda_k}{\lambda} E[\xi_k] = \frac{1}{\lambda} \sum_k E[P_k], \quad (10)
\]

where \(k \in \{c, w, m\}\) represents the cellular phase, the WiFi phase and the mobile device, respectively. The corresponding average power consumption can be calculated as:

\[
E[P_k] = p_k \cdot Pr\{N_i > 0\} = p_k \cdot \rho_k. \quad (11)
\]

Since the utilization of the queue is the probability that the server is busy, we have \(Pr\{N_i > 0\} = \rho_k\), i.e., the energy cost is only incurred during the fraction of the time the server is busy. The parameters \(p_c\) and \(p_w\) are the utilization of the cellular and WiFi network. According to Fig. 2, they can be separately calculated as: \(\rho_c = \pi_c - \pi_{c,0}\) and \(\rho_w = \pi_w - \pi_{w,0}\).
4.3 The ERWP Metric

We define a new metric, the Energy-Response time Weighted Product (ERWP) as:

\[ ERWP = E[E|^\omega|] \cdot E[T]^{1-\omega} \]  

(12)

where \( E[T] \) and \( E[E] \) are the mean response time and mean energy consumption, respectively, and \( \omega \) (ranging between 0 and 1) is a weighting parameter that represents the relative significance of energy consumption and response time.

Further, by substituting (9) and (10), into (12), we can formulate the explicit expressions and the optimization problem of the ERWP metric:

\[ \gamma^* = \arg \min \gamma \]  

we seek the abandonment rate \( \gamma^* \) such that ERWP is minimized.

5. PERFORMANCE EVALUATION

Using measurements from real traces in [2], the average data rates of the cellular and WiFi networks are set as \( s_c = 200 \) Kbps and \( s_w = 2 \) Mbps, respectively. The average duration of WiFi availability period is 52 min \( (\xi = 1/52 \text{ min}^{-1}) \), while the average duration with only cellular network coverage is 25.4 min \( (\eta = 1/25.4 \text{ min}^{-1}) \). The availability ratio is thus 67%. The mean job size is assumed to be 10 MB.

We set the power coefficients \( p_c = 2.5 \) W, \( p_w = 0.7 \) W and \( p_m = 2 \) W, respectively. Besides, suppose that the total job arrival rate is \( \lambda = 0.5 \) packet/min, the mobile service rate \( \mu_m = 0.2 \) and the cloud service rate \( \mu_r = 1 \).

As shown in Fig. 4, the non-delayed offloading model has smaller queueing delays. Therefore, by optimally choosing the abandonment rate \( \gamma \), the delayed offloading model can achieve the smallest ERWP value as depicted in Fig. 4. As shown in Fig. 5, the non-delayed offloading model is more sensitive to the job arrival rate. When considering response time as a more important factor (e.g., for delay-sensitive applications), it is better to use the delayed offloading model; while at higher \( \omega \) when considering energy consumption more important (e.g., for delay-tolerant applications), the non-delayed offloading model is preferred.

6. REFERENCES