Delay-Aware Predictive Network Selection in Data Offloading

Haoran Yu, Man Hon Cheung, Longbo Huang, and Jianwei Huang

***Abstract*—We tackle the network selection problem in operator-initiated mobile data offloading by considering both the predictive and non-predictive cases. For the non-predictive case, we use the Lyapunov optimization technique to design the *Delay-aware Network Selection Algorithm (DNSA)*, which yields an operation cost within $O\left(\frac{1}{V}\right)$ bound of the minimum value, for any $V > 0$. For the predictive case, we propose a novel approach to incorporate prediction into Lyapunov optimization. The proposed *Delay-aware Predictive Network Selection Algorithm (DPNSA)* efficiently exploits the potential benefit of predicting traffic load and users' mobilities, and avoids the state space explosion issue due to information prediction. Numerical results show that *DPNSA* with prediction window size 5 offers roughly 30% delay saving over *DNSA*.**

I. Introduction

Cellular networks worldwide are now facing an unprecedented growth of mobile data traffic. As predicted by Ericsson, the global mobile data traffic will increase by nearly 10-fold between 2013 and 2019 [1]. To handle such a data explosion, mobile data offloading, which delivers data traffic originally targeted for cellular network over other complementary networks (such as Wi-Fi), is a cost-efficient solution to alleviate the increasingly severe congestion in cellular networks.

There are two main approaches in mobile data offloading: the user-initiated offloading and the operation-initiated offloading. In the user-initiated offloading, the user decides which network (e.g., cellular or Wi-Fi) to connect. In the operator-initiated offloading, the network operator monitors the network load and decides whether to offload some users' traffic from cellular to other complementary networks. This process can be realized through the access network discovery and selection function (ANDSF) specified in the 3GPP standards [2].

In this work, we study the network selection problem in the stochastic operator-initiated offloading scenario, where the traffic arrivals and users' locations vary over time. The operator, who owns several networks (e.g., cellular and Wi-Fi), usually prefers to serve users in the network with the lowest operation cost. However, due to networks' limited coverages and users' mobilities, the intermittent service of these networks cannot support all mobile traffic and may cause a severe delay to the users. Hence, we need to design an efficient network selection policy that meets users' total traffic demands and well balances both the operation cost and traffic delay. We first use the Lyapunov optimization framework to develop an online *Delay-aware Network Selection Algorithm (DNSA)*, which stabilizes the network and achieves a close-to-optimal operation cost. Motivated by some recent work of more accurate estimation of users' mobilities and traffic demands, we extend *DNSA* by considering prediction of traffic arrival and users' locations. Numerical results show that the proposed *Delay-aware Predictive Network Selection Algorithm (DPNSA)* achieves a better cost-delay tradeoff than *DNSA*. The idea of incorporating prediction into Lyapunov optimization is novel, and is promising for dealing with more general predictive scheduling problems in wireless networking.

Haoran Yu, Man Hon Cheung, and Jianwei Huang ({yh012, mhcheung, jwhuang}@ie.cuhk.edu.hk) are with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China.

Longbo Huang (longbohuang@tsinghua.edu.cn) is with the Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China.

II. System Model

We consider a slotted system, where an operator serves L users in N networks[1]. Denote the set of networks by $\mathcal{N} = \{1, 2, \ldots, N\}$, and assume that the availability of networks is location-dependent. Let $\mathcal{S} = \{1, 2, \ldots, S\}$ be the set of locations. We use $\mathcal{N}_s \subseteq \mathcal{N}$ to represent the set of available networks at location $s \in \mathcal{S}$. Let μ_n and u_n be the capacity and unit operation cost[2] for network $n \in \mathcal{N}$, respectively.

We use $\mathcal{L} = \{1, 2, \ldots, L\}$ to represent the set of users. We denote $A_l(t)$ as the traffic arrival for user $l \in \mathcal{L}$ at time slot $t \in \{0, 1, \ldots\}$, and assume that $0 \leq A_l(t) \leq A_{\max}$ for all $l \in \mathcal{L}$ and $t \geq 0$. We denote $S_l(t) \in \mathcal{S}$ as user l's location at time slot t. Let $\boldsymbol{A}(t)$ and $\boldsymbol{S}(t)$ be the vectors of traffic arrival and users' locations. We assume that $(\boldsymbol{A}(t), \boldsymbol{S}(t))$ is random over different time slots[3].

At each time slot t, the operator makes the network selection decision for each of the users. We denote the decision by $\boldsymbol{\alpha}(t) = (\alpha_l(t), \forall l \in \mathcal{L})$, where $\alpha_l(t)$ is the network that user l is connected to at time t. If the operator does not assign user l to any network, then $\alpha_l(t) = 0$. Therefore, together with the network's availability constraint, we have $\alpha_l(t) \in \mathcal{N}_{S_l(t)} \cup \{0\}$. Assume that the capacity of a network at time slot t is evenly shared among all users connected to the network, then a user l's transmission rate $r_l(\boldsymbol{\alpha}(t))$ is given by

$$r_l(\boldsymbol{\alpha}(t)) = \frac{\mu_{\alpha_l(t)}}{m_{\alpha_l(t)}(\boldsymbol{\alpha}(t))}, \tag{1}$$

where $m_n(\boldsymbol{\alpha}(t)) = |\{i \in \mathcal{L} : \alpha_i(t) = n\}|$ is the number of users connected to network $n \in \mathcal{N}$ from the entire coverage of network n. We assume that $0 \leq r_l(\boldsymbol{\alpha}(t)) \leq r_{max}$, for all $l \in$

[1] For example, there is one cellular network and $N-1$ Wi-Fi networks.

[2] By unit operation cost, we mean the cost that the operator pays for serving users with one unit of transmission rate.

[3] In this section, we do not make assumption on the type of the randomness.

978-1-4673-1017-8/14/$31.00 ©2014 IEEE

$\mathcal{L}, t \geq 0$. Let $\boldsymbol{Q}(t) = (Q_l(t), \forall l \in \mathcal{L})$ be the queue backlog vector at time t, where $Q_l(t)$ denotes user l's unserved traffic. Assume that $Q_l(0) = 0$ for all $l \in \mathcal{L}$. The queue changes according to the traffic arrival rate and transmission rate as

$$Q_l(t+1) = \max\left[Q_l(t) - r_l(\boldsymbol{\alpha}(t)), 0\right] + A_l(t), \forall l \in \mathcal{L}, t \geq 0. \quad (2)$$

The operator's instant operation cost at time t is defined as

$$c(\boldsymbol{\alpha}(t)) \triangleq \sum_{l=1}^{L} r_l(\boldsymbol{\alpha}(\tau))\, u_{\alpha_l(t)}. \quad (3)$$

Assume that there exists a constant $c_{\max}$ such that $c(\boldsymbol{\alpha}(t)) \leq c_{\max}$, for all $t \geq 0$. Then the goal of the operator is to design an online network selection algorithm that minimizes the time average operation cost, while keeping the network stable. This can be formulated as the following optimization problem:

$$\min \quad \bar{c} \triangleq \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{c(\boldsymbol{\alpha}(\tau))\} \quad (4)$$

$$\text{s.t.} \quad \overline{Q_l} \triangleq \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{Q_l(\tau)\} < \infty, \ \forall l \in \mathcal{L}, \quad (5)$$

$$\text{var.} \quad \alpha_l(t) \in \mathcal{N}_{S_l(t)} \cup \{0\}, \ \forall l \in \mathcal{L}, t \geq 0. \quad (6)$$

III. Network Selection without Prediction

We first assume that there is no prediction of traffic arrival and users' locations, and design the following algorithm.

Delay-aware Network Selection Algorithm (DNSA): At every slot t, the operator observes $\boldsymbol{Q}(t)$, $\boldsymbol{A}(t)$, and $\boldsymbol{S}(t)$, and chooses user-network association $\boldsymbol{\alpha}(t)$ that solves

$$\min \quad \sum_{l=1}^{L} r_l(\boldsymbol{\alpha}(t)) \left(V u_{\alpha_l(t)} - Q_l(t)\right) \quad (7)$$

$$\text{var.} \quad \alpha_l(t) \in \mathcal{N}_{S_l(t)} \cup \{0\}, \ \forall l \in \mathcal{L}. \quad (8)$$

We next state its performance under independent and identically distributed (i.i.d.) network randomness[4].

Theorem 1: Suppose $(\boldsymbol{A}(t), \boldsymbol{S}(t))$ are i.i.d. over slots, and $\mathbb{E}\{\boldsymbol{A}(t)\}$ is strictly interior to the capacity region Λ[5], where there exists a positive ε such that $\mathbb{E}\{\boldsymbol{A}(t)\} + \varepsilon \cdot \mathbf{1} \in \Lambda$. *DNSA* achieves a long-term average cost c_{av}^{DNSA} that satisfies

$$c_{av}^{DNSA} \leq c_{av}^{*} + \frac{B}{V}, \quad \sum_{l=1}^{L} \overline{Q_l} \leq \frac{B + V c_{\max}}{\varepsilon}, \quad (9)$$

where c_{av}^{*} is the optimal long-term operation cost of problem (4)-(6) and $B = \frac{L\left(A_{\max}^2 + r_{\max}^2\right)}{2}$.

According to Little's law, the average queue length is proportional to the average delay. Hence, Theorem 1 implies that, by increasing parameter $V > 0$, the operator can push the operation cost arbitrarily close to c_{av}^{*}, at the expense of an increased average traffic delay.

[4] *DNSA* achieves the similar performance bounds for Markovian network randomness, while the detailed results are omitted here due to space limit.

[5] The capacity region is defined as the closure of the set of arrival vectors that can be stably supported, considering all network selection algorithms.

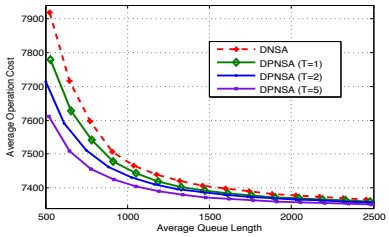

Fig. 1. Performance Comparison

IV. Network Selection with Prediction

We next incorporate prediction into the network selection. Assume that the operator can exactly predict the traffic arrival and users' locations for the future T slots, where T is the size of prediction window. We propose the following algorithm.

Delay-aware Predictive Network Selection Algorithm (DPNSA): At every slot t, the operator observes $\boldsymbol{Q}(t)$, $\boldsymbol{A}(t)$, and $\boldsymbol{S}(t)$, predicts $\boldsymbol{A}(\tau)$ and $\boldsymbol{S}(\tau)$, $\tau \in \{t+1, t+2, \ldots, t+T\}$, runs the following algorithm, and chooses $\boldsymbol{\alpha}^K(t)$ as the network selection at time slot t.

Initialization:

Set $\boldsymbol{\alpha}^0(\tau) = \mathbf{0}$, $\tau = t, t+1, \cdots, t+T$;

Iteration:

1: **for** $k = 0$ to $K-1$ **do**

2: **for** $\tau = t$ to $t+T$ **do**

3: Set $H_l(\tau) = Q_l(t) + \sum_{\tau'=t,\tau'\neq\tau}^{t+T} A_l(\tau') - \sum_{\tau'=t,\tau'<\tau}^{t+T} r_l\left(\boldsymbol{\alpha}^{k+1}(\tau')\right) - \sum_{\tau'=t,\tau'>\tau}^{t+T} r_l\left(\boldsymbol{\alpha}^{k}(\tau')\right)$ for all $l \in \mathcal{L}$;

4: $\boldsymbol{\alpha}^{k+1}(\tau) = \underset{\substack{\alpha_l(\tau)\in\mathcal{N}_{S_l(\tau)} \\ \cup\{0\},\forall l\in\mathcal{L}}}{\arg\min} \sum_{l=1}^{L} r_l(\boldsymbol{\alpha}(\tau)) \left(V u_{\alpha_l(\tau)} - H_l(\tau)\right)$;

5: **end for**

6: **end for**

7: **return** $\boldsymbol{\alpha}^K(t)$;

By generalizing the problem (7)-(8) to the $T+1$ slots case, we get a non-convex optimization problem, which possesses an exponentially large solution space and aims at balancing the cost and delay over the current and future T slots. The low-complexity iteration process in *DPNSA* actually corresponds to an approximation algorithm for solving such an optimization problem. Therefore, with the proper number of iterations K, $\boldsymbol{\alpha}(t)$ will converge to a point that efficiently balances the cost and delay for both the current and future T slots. Figure 1 illustrates *DPNSA*'s performance, which achieves a better cost-delay tradeoff than *DNSA*. For instance, when the target of the operation cost is set as 7400 and $T = 5$, the average queue length under *DPNSA* and *DNSA* are 1090 and 1568, respectively. According to Little's law, *DPNSA* offers 30.48% delay saving over *DNSA*.

V. Future Work

We plan to analytically characterize *DPNSA*'s performance bound, and understand the impact of the prediction window size and variance of randomness. We will further study the robustness of *DPNSA* under a prediction error.

References

[1] Ericsson, "Ericsson mobility report," Tech. Rep., November 2013.

[2] A. Aijaz, H. Aghvami, and M. Amani, "A survey on mobile data offloading: Technical and business perspectives," *IEEE Wireless Communications*, vol. 20, no. 2, pp. 104–112, April 2013.