Quality-of-Service Driven Resource Allocation Based on Martingale Theory

Tingting Liu*†, Jun Li†‡, Feng Shu†, and Zhu Han§¶
*School of Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, CHINA
†School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, CHINA
‡National Mobile Communications Research Laboratory, Southeast University, Nanjing 210009, CHINA
§Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004, USA
¶Department of Computer Science and Engineering, Kyung Hee University, Seoul 02447, South Korea
E-mail: liutt@njit.edu.cn; jun.li@njust.edu.cn; shufeng@njust.edu.cn; zhan2@uh.edu

***Abstract*—One of the key metrics in measuring system quality of service (QoS) is the delay performance. Most existing papers have focused on the studies of decreasing transmission delay. However, as the wireless communication traffic increasing dramatically, queueing delay in the wireless networks becomes a non-negligible issue. Martingale theory, which fits any arrival and service process, providing a much tighter delay bound compared to the effective bandwidth theory, has been proposed to analyze the system queueing delay bound, especially in a bursty traffic scenario. In this paper, we propose to study the resource allocation problem based on the delay bounds derived from martingale theory. In specific, we first revisit some basic knowledge about stochastic network calculus, and present the delay bounds derived from martingale theory in certain typical bursty service models. Then, we setup a resource allocation problem in a computation offloading scenario, where multiple computation nodes with distinct computation capacities are considered. User's computation tasks are usually bursty, and are required to be executed within a limited time. We propose to minimize the system delay violation probability by properly allocating the computation tasks to different computation nodes. A closed-form solution is derived for the computation offloading problem, using a special kind of water-filling policy. Moreover, we discuss two potential models of martingale-based resource allocation, and provide the corresponding system architectures. Finally, numerical results are presented to demonstrate the performances of the proposed scheme. The proposed water-filling scheme achieves a smaller system delay violation probability compared to the benchmark.**

I. Introduction

With the rapid developments of wireless communications, current networks are required to process huge data traffic within a very limited time. Lacking of well-designed resource allocation schemes will induce unsatisfied latency or severe network congestion. How to properly allocate network resources to satisfy the quality of service (QoS) requirements has arisen as a crucial problem in the current and future networks. As one of the critical QoS metrics, delay performance has attracted more and more attentions in recent works. Most existing works focus on the transmission delay. In [1], subcarrier assignment, power allocation and time fraction determination are investigated in a delay sensitive heterogeneous network. In [2], a resource allocation scheme which minimizes the packet transmission delay of the secondary user is proposed. The authors in [3] propose to solve the resource allocation problem concerning the transmission delay without major sacrifices in user's satisfaction.

However, besides the transmission delay, queueing delay is also a non-negligible issue in an end-to-end communication link. Effective bandwidth theory which is associated to a predefined QoS constraint has been proposed to analyze the queueing delay bounds in many papers [4, 5]. The authors of [6, 7] propose to analyze the resource allocation problem based on the effective bandwidth/capacity theory. In [6], the optimal power and rate adaptation scheme is investigated over wireless links. In [8], the authors propose to maximize the effective power efficiency based on the statistical delay-bounded QoS requirements.

However, there is a typical drawback of the effective bandwidth theory. It has a very loose estimation over the delay bounds, which usually lead to inaccurate results for bursty arrivals. Martingale theory is another option in estimating delay bounds [9]. It fits any arrival and service process, and especially, it can provide a very tight bound in a bursty traffic scenario. In [10], the authors provide a theoretical way to measure the end-to-end delay bounds in the multimedia heterogeneous high-speed train networks where the links from the train to the track-side-access points are highly dynamic and bursty. Furthermore, in [11], the authors study the delay bounds in vehicular ad hoc networks where data from vehicles are expected to be bursty. In [12], the authors propose an energy-efficient random access algorithm based on the martingale theory in machine type communication networks.

Similar to the scenarios depicted in [10–12], in wireless communication networks, there are a lot of bursty traffic environments. In this paper, we are inspired to study the resource allocation problem based on the delay bounds derived from martingale theory. The main contributions of our work are summarized as follows,

1) We propose to study the resource allocation problem based on the martingale theory. Firstly, we revisit some basic knowledge of the stochastic network calculus. Then we provide the delay bounds derived from the martingale theory employing certain typical service models, such as Aloha and CSMA/CA.

978-1-5386-4727-1/18/$31.00 ©2018 IEEE
Authorized licensed use limited to: NANJING UNIVERSITY OF SCIENCE AND TECHNOLOGY. Downloaded on January 11,2021 at 13:31:37 UTC from IEEE Xplore. Restrictions apply.

2) In a computation scenario, users' computation demands are bursty, and users usually require a sharp task execution time. By considering the influence induced by the transmission and queueing delay, we construct an optimization problem in minimizing the system delay violation probability in the concerned computation offloading scenario. Then, we derive a closed-form solution, using a special type of water-filling policy.
3) Two potential applications in wireless communication networks based on the martingale theory are also discussed. Moreover, we provide two distinct architectures, one of which is suitable for the multimedia networks, and the other one fits the multi-hop networks.
4) Simulation results are provided to demonstrate the performance of the proposed scheme. It can be verified that the proposed scheme has a smaller system delay violation probability compared to the equal offloading scheme.

The remainders of this paper are organized as follows: the delay bounds derived from the martingale theory are presented in Section II. The resource allocation problem in computation offloading scenario and the potential applications based on martingale theory are discussed in Section III. Numerical results are elaborated in Section IV, and finally conclusions are drawn in Section V.

II. Delay Bounds Based on Martingale Theory

A. Preliminary

In this section, we first revisit some definitions and useful results in martingale theory. A represents the arrival process which is defined as follows,

$$A(m,n) = \sum_{k=m}^{n} a_k, \tag{1}$$

where a_k is the arrived data at time k. $A(m,n)$, which is in a bivariate form, is the accumulative amount of arrival source data over time interval $[m,n]$. We use $A(n) \triangleq A(0,n)$ for brevity over time interval $[0,n]$. Service process is also characterized by a bivariate form $S(m,n)$, and the corresponding departure process is denoted as $D(n)$.

The relationship between arrival process A and service process S is depicted in Fig. 1. Service process S plays a critical role in coupling the arrival process A and the departure process D, in terms of a $(\min,+)$ convolution, i.e.,

$$D(n) \geq A * S(n) \triangleq \min_{0\leq m\leq n}\{A(m)+S(m,n)\}. \tag{2}$$

Service process $S(n)$ can be regarded as an impulse response in a linear and time invariant system. (2) provides a lower bound for the departure process D.

Next, we introduce the backlog process $Q(n)$ and delay process $W(n)$. The backlog process $Q(n)$ is the data amount in the system at time n. In other words, $Q(n)$ is the queuing length at time n, i.e.,

$$Q(n) = \sup_{n\geq 0}\{A(n)-S(n)\}, \tag{3}$$

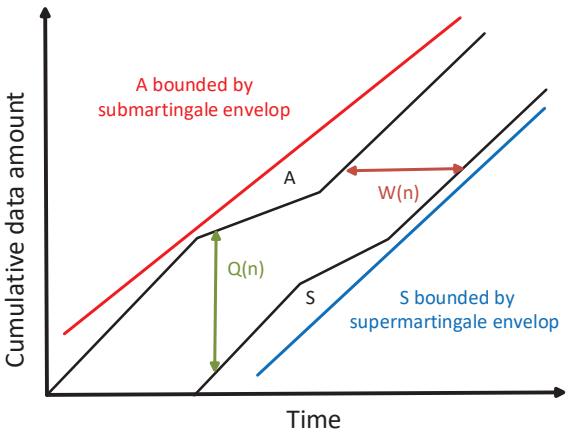

Fig. 1. An illustration of the relationship between arrival process A and service process S.

which can be interpreted as the vertical distance between $A(n)$ and $D(n)$. The delay process $W(n)$, which is the horizontal distance between the curves $A(n)$ and $D(n)$, is defined as follows,

$$W(n) = \min\{k \geq 0 | A(n-k) \leq D(n)\}. \tag{4}$$

From (4), we have

$$W(n) \geq k \Leftrightarrow A(n-k) \geq D(n). \tag{5}$$

Then, the complementary cumulative distribution function (CCDF) of the delay process $W(n)$ can be represented as

$$\Pr\left(W(n) \geq k\right) = \Pr\left(A(n-k) \geq D(n)\right). \tag{6}$$

Before presenting the backlog and delay bounds, we introduce some important martingale-related definitions in the following.

Definition 1. Martingale Process: *A discrete-time martingale process is a discrete-time stochastic process $X_1, X_2, X_3, \ldots$ that satisfies for any time $n \geq 1$:*

$$\begin{aligned} &E[|X_n|] < \infty, \\ &E[X_{n+1}|X_1, X_2, \ldots, X_n] = X_n. \end{aligned} \tag{7}$$

It means that given all the history observations over time interval $[1,n]$, the conditional expectation of the next observation on time $n+1$ equals to the observation on time n.

Definition 2. Supermartingale Process: *A discrete-time martingale process is a discrete-time stochastic process $X_1, X_2, X_3, \ldots$ that satisfies for any time $n \geq 1$:*

$$\begin{aligned} &E[|X_n|] < \infty, \\ &E[X_{n+1}|X_1, X_2, \ldots, X_n] \leq X_n. \end{aligned} \tag{8}$$

It can be seen that if a process is a supermartingale, the expectation drops as time passes by.

Authorized licensed use limited to: NANJING UNIVERSITY OF SCIENCE AND TECHNOLOGY. Downloaded on January 11,2021 at 13:31:37 UTC from IEEE Xplore. Restrictions apply.

Definition 3. Arrival-Martingales: *A is an arrival martingale process, if for every $\theta > 0$, there is a K_a and a function h_a such that the process*

$$h_a(a_n)e^{\theta(A(n)-nK_a)}, n \geq 0, \tag{9}$$

is a supermartingale.

Definition 4. Service-Martingales: *S is a service martingale process, if for every $\theta > 0$, there is a K_s and a function h_s such that the process*

$$h_s(s_n)e^{\theta(nK_s-S(n))}, n \geq 0, \tag{10}$$

is a supermartingale.

B. Delay Bounds in Different Service Models

Since we are focused on the delay bounds optimization problem, in this section, we will present some main results of the delay bounds derived in recent researches.

Martingale theory fits the environment where the source data flow operates in a bursty model. So, the arrival process A is usually modeled as a Markov-Modulated On-Off (MMOO) process for simplity. The transition matrix of the Markov chain is give by

$$T_a = \begin{pmatrix} 1-p_a & p_a \\ q_a & 1-q_a \end{pmatrix}. \tag{11}$$

The steady state distribution of a_k, which is the arrival data at time k, is $\pi_a = \left(\frac{q_a}{p_a+q_a}, \frac{p_a}{p_a+q_a}\right)$. The arrival process $A(n)$, which has been defined in (1), can be represented by

$$A(n) = \sum_{k=1}^{n} f(a_k), \tag{12}$$

in which $f(0) = 0$ represents that no data arrives, while $f(1) = R$ represents that there are R amount of data arriving.

We present the delay bounds of two service models, i.e., Aloha and CSMA/CA, respectively [9],

$$\text{Aloha: } \Pr\left(W(n) \geq k\right) \leq \frac{E[h_a(a_0)]}{H} e^{-\theta^* K_s k}, \tag{13}$$

$$\text{CSMA/CA: } \Pr\left(W(n) \geq k\right) \leq \frac{E[h_a(a_0)]E[h_s(s_0)]}{H} e^{-\theta^* K_s k}. \tag{14}$$

III. Resource Allocation Based on Martingale Theory

Observing (13) and (14), we find that there is only one variable k in the delay bound expressoions. It inspires us to handle the resource allocation problem based on the martingale theory-derived bound.

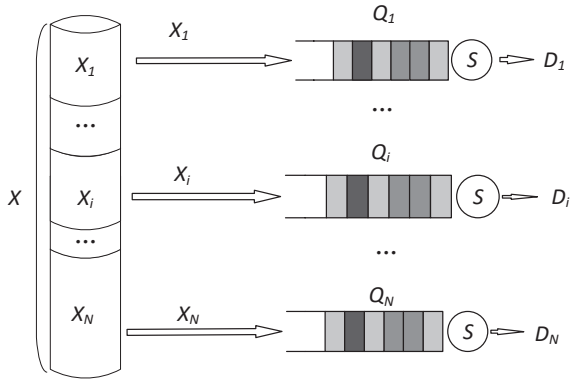

Fig. 2. A simple example in computation offloading.

A. Computation offloading Problem Formulation: A Simple Scenario

In this subsection, we conduct resource allocation in computation offloading scenario, where users' computation requirements are bursty and delay sensitive. The system model is depicted in Fig. 2. In this simple example, we consider only one user who has a total computation task amount of X. The computation task needs to be executed within a limited time. Otherwise, it will be discarded. In this way, due to the limited local computation capacity, this user has to offload its tasks to the other computation nodes, each of which has a distinct computation capacity given by μ_i. These computation nodes are independent from each other. Note that these computation nodes serve multiple users at the same time in a first in first out (FIFO) manner.

In this model, we aim to minimize the system delay violation probability which is a summation of the delay bounds of each computation nodes, i.e.,

$$\begin{aligned} &\min_{X_i} \sum_{i=1}^{N} \Pr\left(W_i(n) \geq (X_i/R)\right), \\ &\text{s.t., } \sum_{i=1}^{N} X_i = X,\ X_i \geq 0, \end{aligned} \tag{15}$$

where X_i is the assigned computation tasks to the ith computation node. In order to simplify the process, we assume that the transmission rate to each node is a uniform constant R. In this scenario, we omit the computation delay in each node, since the computation delay is relatively smaller than communication delay and queueing delay. The best situation is that as soon as the computation task X_i is received completely, the task X_i will be executed. The objective function aims to minimize the probability that the task cannot be executed immediately, when it is received completely.

Authorized licensed use limited to: NANJING UNIVERSITY OF SCIENCE AND TECHNOLOGY. Downloaded on January 11,2021 at 13:31:37 UTC from IEEE Xplore. Restrictions apply.

B. Solution of the Computation Offloading Problem

Given the computation nodes employing the CSMA/CA service model, we rewrite (15) as

$$\min_{X_i} \sum_{i=1}^{N} \frac{E[h_a(a_0)]E[h_s(s_0)]}{H} e^{-\theta_i^* K_i \left(\frac{X_i}{R}\right)}, \quad (16)$$
$$\text{s.t.,} \sum_{i=1}^{N} X_i = X, \; X_i \geq 0,$$

where $\frac{E[h_a(a_0)]E[h_s(s_0)]}{H} = \frac{p_a + q_a h(0)/h(R)}{p_a + q_a} \triangleq \delta$, $K_i = -\frac{1}{\theta_i^*} \ln \mathbb{E}[e^{-\theta_i^* \mu_i}]$. θ_i^* is determined by

$$\eta(\theta_i) = e^{\theta_i \mu_i}. \quad (17)$$

$\eta(\theta_i)$ is the spectral radius of

$$T_a^\theta \triangleq T_a e^{\theta f(a_k)}. \quad (18)$$

It can be verified that (16) is a convex problem. Using the Lagrangian function, we have

$$L = \sum_{i=1}^{N} \left(\delta e^{-\theta_i^* K_i \left(\frac{X_i}{R}\right)} + \lambda X_i + \gamma_i X_i \right), \quad (19)$$

where $\{\gamma_1, ..., \gamma_N\}$, and λ are the Lagrangian multipliers. We obtain the necessary and sufficient Karush-Kuhn-Tucher (KKT) conditions as

$$\begin{cases} \frac{\partial L}{\partial X_i} = -\frac{\delta \theta_i^* K_i}{R} e^{-\frac{\theta_i^* K_i X_i}{R}} + \lambda + \gamma_i = 0, \\ \gamma_i \geq 0, \\ \gamma_i X_i = 0. \end{cases} \quad (20)$$

The condition $\gamma_i X_i = 0$ results in $X_i = 0$, or $\gamma_i = 0$,

$$X_i^* = \frac{R}{\theta_i^* K_i} \left(\ln \frac{\delta}{\lambda^* R} - \ln \frac{1}{\theta_i^* K_i} \right). \quad (21)$$

The unknown variable λ^* is chosen such that the constraints $X_i^* \geq 0$ and $\sum_{i=1}^{N^*} X_i = X$ are satisfied.

We rewrite (21) as the water-filling solution as,

$$X_i^* = \frac{R}{g_i} \left(\ln \frac{1}{g_o} - \ln \frac{1}{g_i} \right)^+. \quad (22)$$

where $g_o = \frac{\lambda^* R}{\delta}$ and $g_i = \theta_i^* K_i$.

The illustration of the solution is depicted in Fig. 3. Observing (22), it can be seen that the solution of the proposed computation offloading problem is a special type of water-filling. As shown in Fig. 3, the value of $\ln(\frac{1}{g_o})$ is the water level which is determined by the optimal λ^*. $\ln(\frac{1}{g_i})$ can be regarded as a kind of resource measurement indicating the state of node i. When $\ln(\frac{1}{g_i})$ is smaller than $\ln(\frac{1}{g_o})$, it indicates that node i has extra computation resources, and therefore it can be allocated with computation tasks. If $\ln(\frac{1}{g_i})$ is larger than $\ln(\frac{1}{g_o})$, node i will not be allocated. Moreover, the filled water is determined by both the height value of $\ln(\frac{1}{g_o}) - \ln(\frac{1}{g_i})$ and the width value of $\frac{R}{g_i}$.

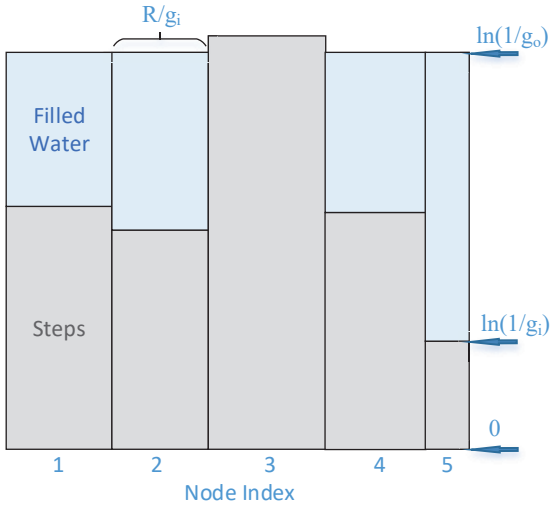

Fig. 3. Illustration of the solution in (22).

C. Discussions

By employing the proposed computation allocation scheme, the system delay violation probability will be minimized. However, the proposed scheme provides no guarantee that the computation tasks will be executed within a given time. We will investigate the resource allocation problem concerning the execution time constraint based on martingale theory in our future work.

Besides the application in computation offloading scenario, martingale-based resource allocation scheme can be extended to many other scenarios. It matches the multimedia networks very well, if we change the system model to a multiplexing architecture as shown in Fig. 4 (a). Also, it can be used in a multi-hop network, if we model the system architecture as shown in Fig. 4 (b). There are various applications to be exploited. We will elaborate them in our future works.

IV. Numerical Results

In this simulation, we present some numerical results in the computation offloading scenario. The elements in transition matric are $p_a = 0.4$ and $q_a = 0.5$. The transmission rate is given by $R = 4$ cycles/s. Here we define the unit of transmission rate as cycles/s, and this is due to the reason that the transmitted computation tasks are in unit of cycles for simplicity. Suppose we have 3 computation nodes, and the node computation capacities are given by $\mu_1 = 2.1$ cycles/s, $\mu_2 = 2.2$ cycles/s, and $\mu_3 = 2.3$ cycles/s.

In Fig. 5, we compare the optimal offloaded computation amounts in each node. It can be seen that as the parameter λ decreases, i.e., the water level increases, computation tasks are firstly allocated to node 3, corresponding to the node with the highest computation capacity μ_3. This can be verified at point $\lambda = 0.07$. At point $\lambda = 0.04$, three nodes are all allocated with computation tasks, while node 3 is allocated with the most amount of tasks. However, at point $\lambda = 0.01$, node 1 is allocated with the most amount of computation tasks. This can

Authorized licensed use limited to: NANJING UNIVERSITY OF SCIENCE AND TECHNOLOGY. Downloaded on January 11,2021 at 13:31:37 UTC from IEEE Xplore. Restrictions apply.

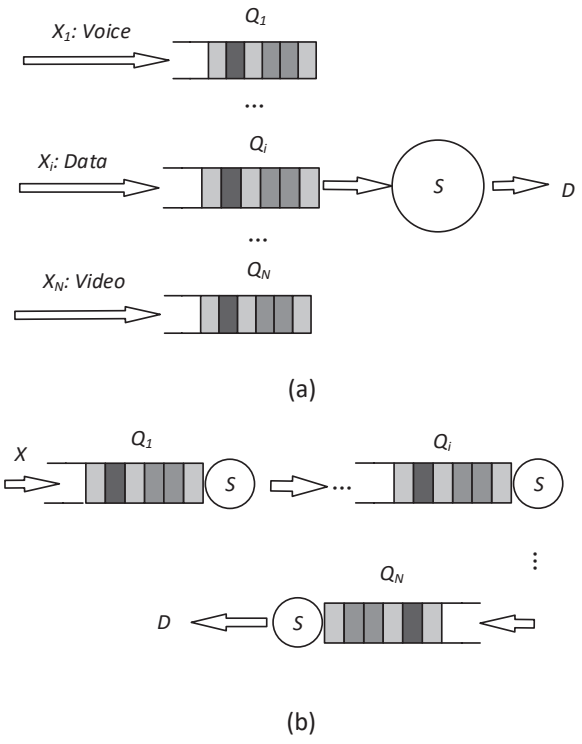

Fig. 4. (a) multiplexing architecture. (b) multi-hop architecture.

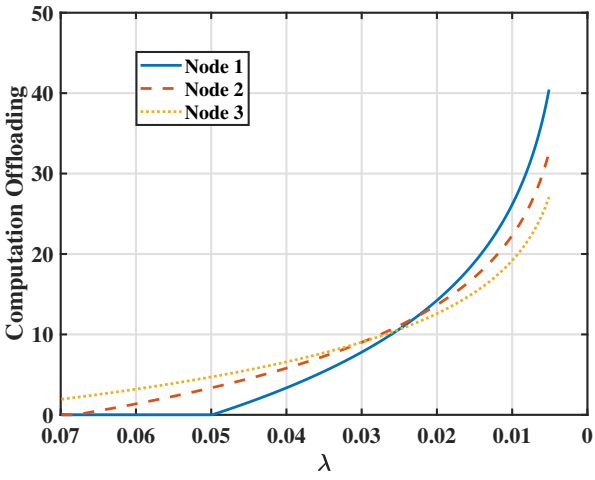

Fig. 5. Optimal Computation downloading of three nodes.

be verified from Fig. 3 that the offloaded computation amounts are determined not only by the height $\ln(\frac{1}{g_o})-\ln(\frac{1}{g_i})$, but also by the width $\frac{R}{g_i}$. Node 3 has a larger μ_i, and therefore it has a smaller $\ln(\frac{1}{g_i})$. This will help it be filled with computation tasks firstly. However, since the width $\frac{R}{g_i}$ of node 3 is small, the increment speed of the filled water will not as fast as node 1 which has a wider width.

In Fig. 6, we compare the proposed martingale-based computation offloading scheme with the equal offloading scheme. In the equal assignment scheme, we assign the 3 nodes with

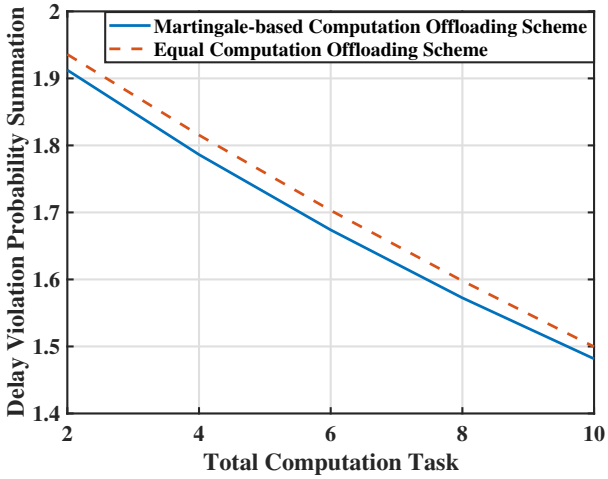

Fig. 6. Delay Violation Probability Comparison between the martingale-based computation offloading scheme and the equal computation offloading scheme.

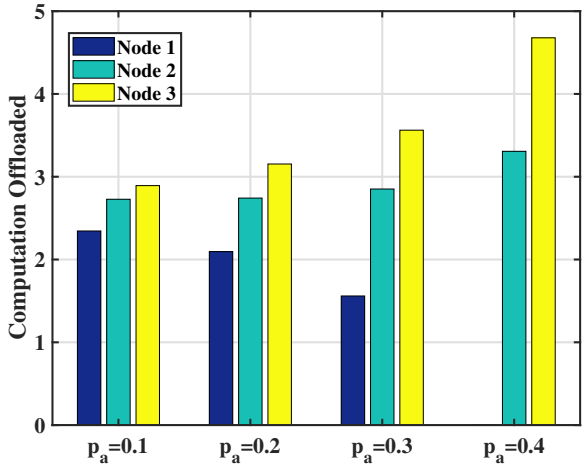

Fig. 7. Computation Offloading Comparison, when p_a varies from 0.1 to 0.4.

equal computation amount, without consideration of their distinct computation capacities. As the total computation task amount X increases, the system delay violation probabilities in both schemes decrease. Also, it can be seen that the proposed martingale-based scheme always has a smaller violation probability summation compare to the equal offloading one.

In Fig. 7, we vary p_a from 0.1 to 0.4 to verify the variation of the offloaded computation tasks among three different nodes. Given the total computation tasks $X = 8$, it can be seen that as p_a increases, i.e., the steady state π_1 increases, node 3 which has a larger computation capacity is prone to be allocated with more tasks. While node 1 with a smaller computation capacity is allocated with less tasks.

V. Conclusion

In this paper, we provide a new idea on resource allocation problem. Martingale theory derived delay bound, which is

Authorized licensed use limited to: NANJING UNIVERSITY OF SCIENCE AND TECHNOLOGY. Downloaded on January 11,2021 at 13:31:37 UTC from IEEE Xplore. Restrictions apply.

much tighter than the one derived from the effective bandwidth theory, is proposed as an optimization objective. With the purpose to minimize the system delay violation probability, firstly, we setup an optimization problem in the computation offloading scenario by properly assigning computation tasks to different computation nodes. Then, we provide a closed form solution, i.e., the modified water filling formula, to this problem. Numerical results are provided to demonstrate the effectiveness of the proposed method. It can be verified that the proposed martingale-based computation offloading scheme has a smaller delay violation probability compared to the equal computation offloading one.

ACKNOWLEDGMENT

This work is supported in part by the National Natural Science Foundation of China under Grant No. 61702258, 61727802, 61771244, 61472190, 61501238, in part by the National Key R&D Program under the grant number 2018YFB1004802, in part by the Jiangsu Provincial Science Foundation under Project BK20150786, in part by the Specially Appointed Professor Program in Jiangsu Province, 2015, in part by the Fundamental Research Funds for the Central Universities under Grant 30916011205, in part by the Open Research Fund of National Mobile Communications Research Laboratory, Southeast University, under grant No. 2017D04, in part by the China Postdoctoral Science Foundation under grant 2016M591852, in part by Postdoctoral research funding program of Jiangsu Province under grant 1601257C, in part by the China Scholarship Council Grant 201708320001, in part by US MURI, US NSF CNS-1717454, CNS-1731424, CNS-1702850, CNS-1646607.

REFERENCES

[1] Y. Li, Y. Shi, M. Sheng, C. X. Wang, J. Li, X. Wang, and Y. Zhang, "Energy-efficient transmission in heterogeneous wireless networks: A delay-aware approach," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 9, pp. 7488–7500, Sep. 2016.

[2] G. Saleh, A. El-Keyi, and M. Nafie, "Cross-layer minimum-delay scheduling and maximum-throughput resource allocation for multiuser cognitive networks," *IEEE Transactions on Mobile Computing*, vol. 12, no. 4, pp. 761–773, Apr. 2013.

[3] M. Li and X. Wang, "Delay and rate satisfaction for data transmission with application in wireless communications," *IEEE Network*, vol. 29, no. 5, pp. 70–75, Sep. 2015.

[4] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.

[5] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[6] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Transactions on Wireless Communications*, vol. 6, no. 8, pp. 3058–3068, Aug. 2007.

[7] ——, "Quality-of-service driven power and rate adaptation for multichannel communications over wireless links," *IEEE Transactions on Wireless Communications*, vol. 6, no. 12, pp. 4349–4360, Dec. 2007.

[8] W. Cheng, X. Zhang, and H. Zhang, "Statistical-qos driven energy-efficiency optimization over green 5g mobile wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 12, pp. 3092–3107, Dec. 2016.

[9] F. Poloczek and F. Ciucu, "Service-martingales: Theory and applications to the delay analysis of random access protocols," in *2015 IEEE Conference on Computer Communications (INFOCOM), Hong Kong, China*, Apr. 2015, pp. 945–953.

[10] Y. Hu, H. Li, Z. Chang, and Z. Han, "Scheduling strategy for multimedia heterogeneous high-speed train networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 4, pp. 3265–3279, Apr. 2017.

[11] ——, "End-to-end backlog and delay bound analysis for multi-hop vehicular ad hoc networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 10, pp. 6808–6821, Oct. 2017.

[12] L. Zhao, X. Chi, and Y. Zhu, "Martingales-based energy-efficient d-aloha algorithms for mtc networks with delay-insensitive/urllc terminals co-existence," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1285–1298, Apr. 2018.

Authorized licensed use limited to: NANJING UNIVERSITY OF SCIENCE AND TECHNOLOGY. Downloaded on January 11,2021 at 13:31:37 UTC from IEEE Xplore. Restrictions apply.