IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2020.DOI

Reinforcement Learning-based Power-Saving Algorithm for Video Traffics Considering Network Delay Jitter

DARA RON[1] **and JUNG-RYUN LEE**[1,2] **(Senior Member, IEEE)**
[1]Department of Intelligent Energy and Industry, Chung-Ang University, Seoul 156-756, Republic of Korea.
[2]School of Electronics and Electrical Engineering Engineering, Chung-Ang University, Seoul, 156-756, Republic of Korea.

Corresponding authors: Jung-Ryun Lee (e-mail: jrlee@cau.ac.kr).

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC support program (IITP-2022-2018-0-01799) supervised by the IITP (Institute for Information & communications Technology Planning Evaluation), by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry Energy (MOTIE) of the Republic of Korea (No. 20214000000280), and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No.NRF-2020R1A2C1010929).

ABSTRACT Recent studies on energy efficiency and scheduling of power-saving mode have been considered as key technologies for reducing the energy consumption of device-to-device (D2D) communication. Wi-Fi Direct (P2P), one of the key protocols for D2D communication, defines the on-off power saving mechanic called the notice of absence (NoA) power-saving mode that can be applied to the multimedia video traffic. The on-off power saving mechanic enables the user to transmit or receive the real-time video frame during the awake interval in which the video frame rate should meet the requirement. When the user can wholly transmit one video frame before the end time of a required inter-frame interval, it can switch to the sleep mode to save the power consumption. However, the challenge remaining for the NoA method is the fixed length of awake/sleep interval, even if the traffic load is varied. Therefore, in this study, we proposed a reinforcement learning-based power saving (RLPS) method to enhance the performance of the notice of absence (NoA) power-saving mode in Wi-Fi direct with taking the multimedia video transmission and the network delay jitter into consideration. The proposed RLPS method enables the Wi-Fi direct device to dynamically estimate the length of awake interval for transmitting the future video frame in real-time. In addition, the Wi-Fi direct device may wake up too early before the arrival of the video frame, which is caused by the network delay jitter. Thus, the client device has to wait for receiving the video frame. To tackle this challenge, the proposed RLPS method enables the device to predict the start time of awake interval for the purpose of reducing the delay time for receiving the upcoming video frame. Results show that the proposed RLPS method outperforms the existing NoA power-saving mode in terms of the outage probability, energy consumption, and transmission delay of Wi-Fi Direct devices.

INDEX TERMS Wi-Fi Direct, opportunistic power saving, NoA, reinforcement learning, network delay jitter.

I. INTRODUCTION

WITH drastic increase of mobile data communication and emergence of smart devices, it has become an urgent problem to improve system capacity and strengthen the quality of service (QoS) of users. D2D communication is a key technology for the upcoming future communication systems, which is designed to solve this problem by increasing the system capacity, reducing the transmission delay, and improving the overall spectrum efficiency [1]. However, how to improve energy efficiency is crucial for D2D communication because D2D user typically uses handheld equipment with energy-limited battery.

Since the Wi-Fi direct device uses handheld equipment with energy-limited battery, the method to reduce the energy consumption has become the potential research direction. Wi-Fi Alliance defined the notice of absence (NoA) power-saving method to improve the energy consumption of the device via switching the radio circuitry to the sleep mode

whenever there is no data to be transmitted [2], [3]. Although the original NoA method has some success in reducing the energy consumption, the various studies have attempted to modify the NoA method in the different ways to enhance the energy efficiency because the existing power saving methods do not provide sufficient functions to dynamically cope with varying traffic load [4]. The authors of [5] proved that the performances of the NoA and opportunistic (Opp) power save modes vary according to the type of future wireless USB (WSB) applications. Therefore, the authors proposed an effective method that is capable of dynamically selecting one of two power saving modes (i.e., NoA or Opp power saving modes) based on the WSB application to enhance the energy efficiency. Through the NS-3 simulator, the results verify that the performances of the proposed method (i.e., QoS of various WSB applications and energy efficiency) are better than those of existing NoA and Opp power saving approaches. The authors of [6] optimizes the position of an unmanned aerial vehicles (UAV) in 3-dimensional space to reduce the distance between the UAV and its clients, and thus it can maintain the connectivity of client devices, increase the overall network throughput, and improve energy efficiency. In addition, the authors assume that the P2P group owner (GO) is installed over an UAV, which connects to several clients, so that the UAU and its clients can employ the traditional power saving mechanism in Wi-Fi direct (i.e., opportunistic or notice of absence power saving mode) to enhance the energy efficiency. The con of the traditional power saving mode in Wi-Fi direct is that the length of sleep/awake interval is fixed. Being different from [6], the authors of [7] proposed a Dynamically Synchronized Power Management (DSPM) method that is capable of synchronizing the active time slots with the data transmission intervals. In addition, the DSPM method adjusts the sleep and awake interval according to the traffic pattern in order to increase the energy efficiency. The NS-3 simulation results verify that the proposed DSPM algorithm outperforms the existing NoA method. [8] proposed a traffic-aware parameter tuning scheme to dynamically adjust the awake and absence periods in the NoA power-saving mode and to optimize the client traffic window (CTWindow) in the opportunistic power-saving (OPS) mode according to traffic load.

In addition, with the growing complexity of mobile network architectures, it may have difficulty to address complex control problems in communication networks. The use of machine learning algorithms into future mobile networks is drawing tremendous research attention these days, because it has ability to predict future scenarios, adapt to the network changing environments, and discover the patterns that a human can miss [9], [10]. Particularly, ML enables the device to learn from its experience without intervention of human. Reinforcement learning (RL) is a branch of the ML algorithm, which enables the agent to deal with the dynamic problem using the trial-and-error strategy. With this strategy, the agent can try a possible solution, get the reward from its environment, and transit to a new state. By trying all possible solutions in all states repeatedly, the agent can learn from its experience, and select the best optimal solution according to the optimal decision-making policy. Hence, RL has shown a great potential to solve non-convex and control problems.

With the key features of RL above, various works focusing on the power saving problem in wireless communication have been conducted. [11] proposed a ALOHA and RL-based medium access control (MAC) protocol with Informed Receiving for wireless sensor networks. This method enables a transmitter to inform its receiver about its future slot selection so that it can turn off its radio in other slots to reduce energy consumption. [12] used RL for opponent modeling, and proposed a cooperative communication protocol based on received signal strength indicator and node energy consumption in a competitive context. [13] proposed an RL-based MAC protocol for wireless sensor networks and employed an RL frame to schedule the sleep and active periods of a node to minimize energy consumption. The authors of [14] considered that the sensor node is capable of operating on three modes: 1) transmission, 2) listening, and 3) sleep. In this study, the authors proposed a reinforcement learning (RL) algorithm to select an optimal action based on the decision-making policy, where the actions correspond to those three operation modes. In addition, the RL method is employed to adjust the length of sleep and awake interval in order to improve the energy efficiency while guaranteeing the efficient packet transmission. In [15], P. Verma et al. applied the RL method to handle the lifetime problem in the wireless sensor network because the battery of the sensor node is limited energy and impossible to be recharged. In this study, the RL algorithm enables each node to select an optimal activity by itself, where those activities consist of the sleep, the awake, and the adjustable interval time of sleep/awake to preserve the energy efficiency while ensuring the effective packet transmission. The authors of [16] the authors demonstrated that the fixed length of active/sleep period in IEEE 802.15.4 standard is not suitable with the topology changes in the dynamic sensor network. Therefore, the authors presented the RL-based method to find the optimal duty cycle for the purpose of enhancing the energy efficiency of the IEEE 802.15.4 standard.

Since the existing NoA method provides the fixed length of sleep/awake interval, which makes it not suitable for the application i.e., the video streaming, our study proposed the reinforcement learning-based power saving (RLPS) scheme to adjust the length of awake interval according to the varying frame size of the video streaming. In addition, although the video frame is sent based on the scheduled time, it may arrive late at the destination, which is caused by the network delay jitter. The proposed RLPS method enables the receiver to predict the start time of each awake interval in order to reduce the effect of the network delay jitter. The benefits of predicting the start time of the awake interval according to the delay jitter can improve the performance gain over the traditional notice of absence (NoA) power-saving method as follows. The traditional NoA method schedules the packet transmis-

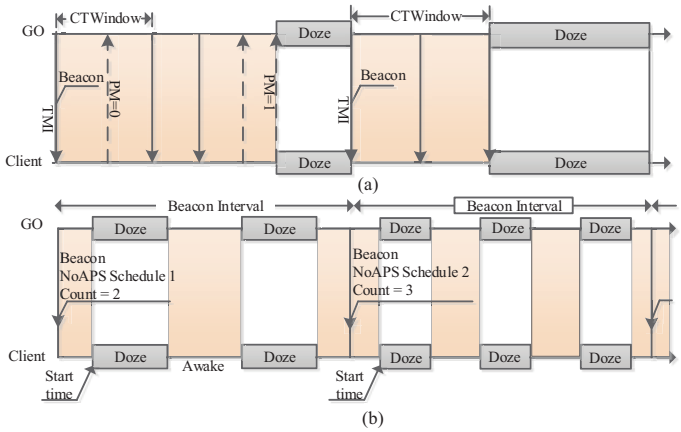


FIGURE 1: Two power-saving mechanisms: (a) opportunistic power-saving (OPS) mode and (b) NoA power-saving mode.

sion time with considering the delay jitter by allowing the group owner (GO) to send a beacon message to its client. Using the NoA method, the start time of awake interval may be faster than the actual arrival frame, which is caused by the jitter delay. Therefore, the client device wastes the energy consumption for waiting for the arrival video frame. In this context, the prediction of the start time of awake interval based on the delay jitter using our proposed RL method can improve the performance gain over the NoA method. Results show that the proposed RLPS method outperforms the existing NoA power-saving mode in terms of the outage probability, energy consumption, and transmission delay of Wi-Fi Direct devices.

The rest of this paper is arranged as follows: Subsection II-A explains the power-saving mechanisms developed by Wi-Fi Direct. Subsection II-B describes the group of picture (GoP) structure and inter-dependency among frames in a GoP structure, and explains the related works. Subsection III explains the operational procedure of the proposed RL algorithm. Subsections III-A, III-B, and III-C present the methods of computing the outage probability of each fame class, transmission delay, and energy consumption, respectively. Section IV describes and compares the performances of the proposed RLPS and the existing NoA power-saving methods. Section V concludes the paper.

II. SYSTEM MODEL AND PREVIOUS WORKS

A. POWER-SAVING MODES IN WI-FI DIRECT

Wi-Fi Direct technology was designed by Wi-Fi Alliance to support the D2D communication without an AP [2]. Wi-Fi Direct devices discover each other by performing a conventional Wi-Fi scan to negotiate which device will be selected as a group owner (GO). Then, the other devices act as clients, and they are referred to as group members. After Wi-Fi Direct devices discover each other, a power-saving mode is implemented for data transmission.

There are two power-saving mechanisms in Wi-Fi Direct, i.e., OPS and NoA modes, as shown in Figure 1. In the OPS mode, the GO periodically sends a beacon message to schedule the transmission time. The beacon message includes the start time of CTWindow, the length and number of CTWindows in a beacon interval, and the power management (PM) bit indicator. The PM bit is used to notify that the client has data to transmit to the GO. The PM bit is set as 0 when the client has data to transmit and as 1 at the end of transmission. The GO and the clients simultaneously wake up to transmit data during CTWindows. When data are completely transmitted during CTWindows, the client can switch to the sleep mode at the end point of the CTWindow. If the GO does not transmit whole data, the client extends the CTWindow to receive the remaining data. In the NoA power-saving mode, communication is initiated when the GO send a beacon message to a client. The beacon message includes scheduled information, such as the start time of doze mode, and the number and length of sleep/awake intervals in a beacon interval. Therefore, the GO and client can simultaneously wake up to exchange data during an awake interval and enter the doze mode to reduce energy consumption. Figure 1 (b) shows that the first and second beacon intervals consist of two and three doze intervals, respectively. Therefore, it should be noted that the NoA power-saving mode provides the flexibility of dynamically adjusting the length of sleep/awake intervals according to traffic load.

B. INTER-DEPENDENCY OF MULTI-LAYERED GROUP OF PICTURE (GOP) STRUCTURE AND RELATED WORKS

A video codec with multi-layered GoP structure such as an MPEG-2 video supports various frame sizes. Usually, video frames in multi-layered GoP structure are categorized into three types: Intra(I), Predictive(P), and Bi-directional(B). Video frames are encoded in a sequence referred to as a GoP to reduce their spatial and temporal redundancies [17], [18]. The GoP structure is generally described as MxNy, where x is the number of frames in the P-P or I-P interval and y is the aggregate number of frames in the GoP. Fig. 2 shows the M3N9 GoP structure, which is encoded as IBB-PBB-PBB. In the GoP structure, an I-frame is encoded without referring to any other frames. A P-frame is encoded with referring to the previous P-frame and I-frame. B-frame is encoded with reference to previous I-frame and next P-frame or the previous P-frame and next P-frame. Therefore, frame loss can occur in the following three manners: 1) The loss of I-frame results in loss of all frames in the GoP. 2) If a P-frame is lost, the following P-frame and B-frame are lost. 3) If a B-frame is lost, there is no more frame loss in the GoP. This inter-dependency among frames decides the frame priority with the order of I-, P-, and B-frames.

The inter-dependency among I-, P-, and B-frames has been utilized to increase transmission efficiency of power-saving mode in wireless networks. In [19], [20], the authors suggested the method to adjust the awake interval according to the frame priority. In these studies, we set the number

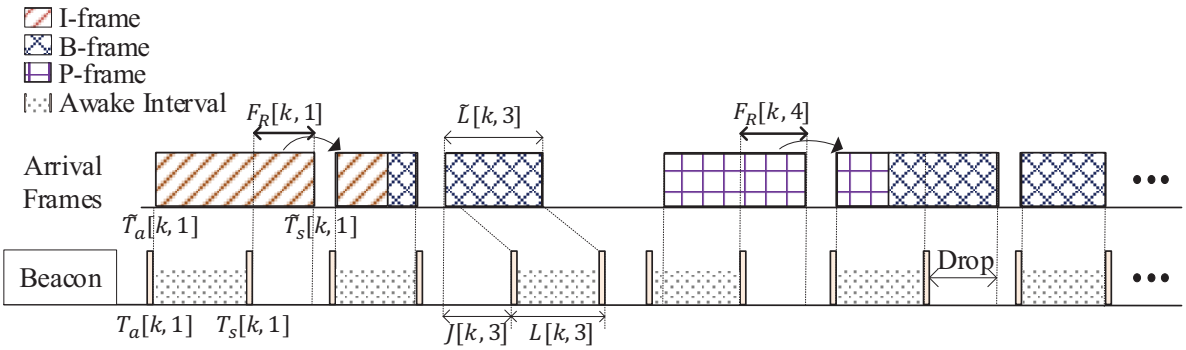


FIGURE 2: Mapping relation between the frames in a GoP and the awake interval in a beacon.

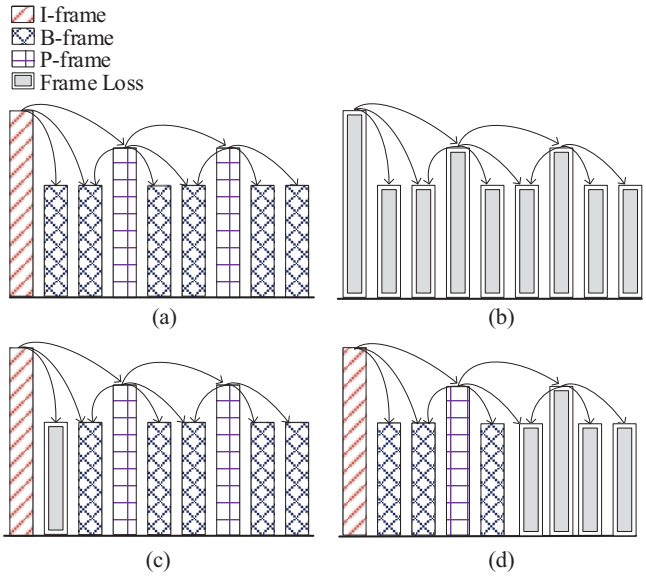


FIGURE 3: (a) No frame loss; (b) Loss of an intra (I)-frame results in the loss of all frames in a GoP; (c) Loss of a predictive (P)-frame causes the loss of the following bidirectional (B)-frames and P-frames; (d) Loss of a B-frame does not affect other frames.

of awake intervals in a beacon interval equal to the number of video frames in the GoP structure to satisfy the one-to-one mapping relation between an awake interval and a video frame. In this circumstance, the trade-off relation between the length of an awake interval and the energy consumption of a device can be explained; longer (shorter) awake interval length decreases (increases) the transmission failure rate while increasing the energy consumption of a device. Since it is impossible to set the length of an awake interval to be exactly fit into the size of each frame due to variable frame sizes, [19] proposed the algorithm which uses the video frame size distribution. In this work, In this work, the length of awake interval is adjusted dynamically according to the probability density function (pdf) of the size of I-, P-, and B-frames, and the transmission failure rate is controlled by the target value which is decided by the mean and the standard deviation of the pdf functions. In addition, frame transmission strategy based on the frame priority in [19] provided the method to deal with remaining fraction of a frame, so that any remaining fraction of an I-frame (P-frame) that is not delivered during an awake interval is concatenated with the immediately following B-frame and transmitted along with that B-frame during the next awake interval. This method can reduce the outage probability of the I-frame (P-frame). However, the operational procedure of the algorithm in [19] assumed that the probabilistic properties of I-, P-, or B-frame are already known and fixed, which is not realistic scenario. To tackle this issue, [20] proposed an Expectation Maximization (EM)-based power-saving method. In this work, the EM algorithm has been employed to update the scale and the shape parameters of the pdf of the video frame size, and the weights of all the gamma mixture components whenever a frame is transmitted. Based on this estimated pdf, the awake interval for each frame is determined by using the target probability.

Although the above two algorithms shows enhanced performance compared to the NoA power-saving method, they have not considered unique phenomena that may occur in wireless network environments. That is, the above algorithms assumed that each frame arrives at the end device consecutively with equal time intervals. However, even though a video codec periodically sends a frame based on a scheduled time, the frame may arrive either early or late at the destination, which results in the network delay jitter. Motivated by this fact, in this study, we propose an RL-based dynamic power-saving (RLPS) method to enhance the performance of the NoA power-saving mode. The proposed algorithm uses the RL algorithm to dynamically adjust the length of awake intervals according to video traffic type and predicts the start point of each awake interval to reduce the effect of network delay jitter.

III. PROPOSED REINFORCEMENT LEARNING-BASED POWER-SAVING MODE

Let $T_a[k,l]$ and $T_s[k,l]$ denote the start and end points of the awake interval allotted to the l-th frame of the k-th GoP,

respectively. The length of this awake interval can be defined as

$$L[k,l] = T_s[k,l] - T_a[k,l]. \tag{1}$$

In real-time video traffic, a video frame may not arrive at the scheduled time because of the variation in the delay time of the network, network delay jitter. Hence, it is necessary to adjust the start point of the awake interval to reduce network delay jitter. The start point of the awake interval for receiving the l-th frame of the $(k+1)$-th GoP can be updated by

$$T_a[k{+}1,l] = T_a[k,l]{+}\alpha(i{-}2)\left|T_a[k,l] - \tilde{T}_a[k,l] + \lambda J[k,l]\right|, \tag{2}$$

where $|\cdot|$ is the absolute value, $\alpha\,(0<\alpha<1)$ is the learning rate, and $\tilde{T}_a[k,l]$ denotes the actual arrival time of the l-th frame in the k-th GoP. $J[k,l]$ represents network delay jitter, which is measured from the start point of the awake interval to the actual arrival time of l-th frame in k-th GoP. λ is the delay factor, which is used to compensate for network delay jitter. The term $err = \left|T_a[k,l] - \tilde{T}_a[k,l] + \lambda J[k,l]\right|$ here is the absolute error of the prediction. The Wi-Fi direct device ought to adjust the start time of awake for the purpose of finding the optimal solution, which minimizes the absolute error of the prediction. The optimal start time of wake can be obtained when the prediction error converges to zero. Thus, the optimal start time of awake is given by $T_a^*[k,l] = \tilde{T}_a[k,l] - \lambda J[k,l]$. Therefore, the optimal start point of awake $T_a^*[k,l]$ decreases as the delay factor λ increases. Since the actual arrival time of video frame varies according the random delay jitter, we introduce the subtraction of $\lambda J[k,l]$ from $\tilde{T}_a[k,l]$ for the purpose of decreasing the start time of awake to ensure that the Wi-Fi direct device can awake up before the video frame arrives. Thus, the delay factor λ is a positive real number $(\lambda \geq 0)$. When we set the value of λ too small, the start time of awake may be later than the arrival time of the video frame, which results in that the Wi-Fi cannot receive a fraction of frame because it is still in sleep mode. However, if we set the value of λ too large, the device may wake up too early. Thus, the Wi-Fi direct device may spend long time for waiting for receiving the video frame. In addition, the large value of λ may result in that the start time of awake is decreased until conflict with the previous end time of awake. Therefore, we should choose a proper value of λ, which guarantees the probability that the device wakes up later than the arrival frame is too small. In addition, i may be set as 1, 2, or 3 to decrease, maintain, or increase the start point of the awake interval, respectively. The end point of the awake interval must be adjusted to reduce the outage probability of the frame and unnecessary energy consumption. The end point of the awake interval can be updated as given by

$$T_s[k{+}1,l] = T_s[k,l]{+}\alpha(j{-}2)\left|T_s[k,l] - \tilde{T}_s[k,l] - \beta\tilde{L}[k,l]\right|, \tag{3}$$

where $\tilde{T}_s[k,l]$ and $\tilde{L}[k,l]$ denote the actual served time of the l-th frame in the k-th GoP and the length of the l-th frame in k-th GoP, respectively, β is a scaling factor, which is used to scale the end point of the awake interval. Our proposed RL-based power saving method enables the Wi-Fi direct device to update the end time of awake interval in order to find the optimal $T_s^*[k,l]$. The optimal end time of awake can be obtained when the absolute error converges to zero. Thus, the optimal $T_s^*[k,l]$ is given by

$$T_s^*[k,l] = \tilde{T}_s[k,l] + \beta\tilde{L}[k,l]. \tag{4}$$

Since the end time of awake for receiving the video frame varies according to the length of video frame, the predicting end time of awake may converges to the average served time $E[\tilde{T}_s[k,l]]$ when we set the scaling factor β to zero. Therefore, the outage probability that the video frame cannot be wholly transmitted should be high. To avoid this issue, we introduce the variation length of video frame weighted by the scaling factor $\beta\tilde{L}[k,l]$ to scale up the optimal end time of awake interval. The outage probability is very close to zero, which means the almost all video frames can be wholly transmitted, when we increase the value of β. Therefore, in practice, we should choose the scaling factor based on the requirement of the system. In addition, j may be set as 1, 2, or 3 to decrease, maintain, or increase the end point of the awake interval, respectively.

Here, we employ the RL algorithm to select the optimal values of i and j, which are denoted by i^* and j^*, respectively. In RL, the problem to solve is described as an Markov Decision Process (MDP). The basic idea of MDP is that the agent in the current state interacts with its environment to take an action according to the policy. As a result, the agent receives a reward and transitions to the next state. From the definition in [22] and [23], an MDP is defined as a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P} \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ are the finite set of states and actions, respectively, $\mathcal{R}$ is the reward function, and $\mathcal{P}$ is the transition probability of moving from the current state $S_{mn}[k,l] \in \mathcal{S}$ to the next state $S_{m'n'}[k+1,l] \in \mathcal{S}$ when using the policy $\mathrm{P}^{\pi}\left(S_{m'n'}[k+1,l]|S_{mn}[k,l]\right)$. $\mathrm{P}\left(S_{m'n'}[k+1,l]|S_{mn}[k,l], A_{ij}[k,l]\right)$ is the transition probability from current state $S_{mn}[k,l]$ to the next state $S_{m'n'}[k+1,l]$ given the action $A_{ij}[k,l] \in \mathcal{A}$, and $\pi\left(A_{ij}[k,l]|S_{mn}[k,l]\right)$ is a mapping from the current state $S_{mn}[k,l]$ to the action $A_{ij}[k,l]$, called the policy. Therefore, $\mathrm{P}^{\pi}\left(S_{m'n'}[k+1,l]|S_{mn}[k,l]\right)$ is defined as the transition probability $\mathrm{P}\left(S_{m'n'}[k+1,l]|S_{mn}[k,l], A_{ij}[k,l]\right)$ weighted by the policy $\pi\left(A_{ij}[k,l]|S_{mn}[k,l]\right)$. The goal of an MDP aims to find an optimal policy π^* to maximize the reward function $\mathcal{R}$. The detailed description of the MDP model in our work is given as follows.

- *Agent*: An agent corresponds to a client, which interacts with a GO to adjust the start and end points of the awake interval, and the length of awake intervals.
- *State Space*: The agent employs RL to obtain the values of i and j. Therefore, the finite state space is defined as the possible values of pair i and j, which is given as $\mathcal{S} = \{S_{11}[k,l], S_{12}[k,l], \ldots, S_{33}[k,l]\}$. It is noted that the integer i (or j) $= 1, 2,$ or 3 mean the start (or the

end) point of awake should be decreased, maintained, or increased, respectively. For instance, if the agent transitions to the state $S_{12}[k,l]$ ($i^* = 1$ and $j^* = 2$), it means that the client decreases and maintains the start and end of points of the awake interval for receiving the l-th frame in the k-th GoP, respectively.

- *Actions Space*: The agent takes an action to observe a next state before making the decision. By trying all actions to observe all next states, the agent knows which action is the best for selection. Therefore, the number of possible actions should be equal to the number of feasible states. It is given by $\mathcal{A} = \{A_{11}[k,l], A_{12}[k,l], \ldots, A_{33}[k,l]\}$. For example, if $A_{12}[k,l]$ is selected, the agent transitions from the current state, $S_{i^*j^*}[k,l]$, to the next state $S_{12}[k+1,l]$. Hence, i^* and j^* for the $k+1$-th GoP are set as 1 and 2, respectively.
- *Reward*: According to the ϵ-greedy policy, the action with the maximum reward is selected with the highest probability. The proposed reward function, $R_{ij}[k,l]$, is defined as

$$R_{ij}[k.l] = -\Big|F_a[k,l] + \alpha(i-2)\big|F_a[k,l]\big|\Big| - \Big|F_s[k,l] + \alpha(j-2)\big|F_s[k,l]\big|\Big|, \quad (5)$$

where $F_a[k,l] = T_a[k,l] - \tilde{T}_a[k,l] + \lambda J[k,l]$ and $F_s[k,l] = T_s[k,l] - \tilde{T}_s[k,l] - \beta\tilde{L}[k,l]$. $T_a[k,l]$ and $\tilde{T}_a[k,l]$ is the estimated and actual arrival time of the l-th frame in the k-th GoP, respectively. The information including the estimated and actual arrival time, and the network delay jitter $J[k,l]$ is used to predict arrival time of the l-frame in the next $k+1$-th GoP. Since the actual arrive time of the next video frame is caused by the network delay jitter, we use $\lambda J[k,l]$ to compensate this variety. λ here represent the delay factor. Therefore, $F_a[k,l]$ measures the difference between the estimate arrival time $T_a[k,l]$ and the actual arrival time $\tilde{T}_a[k,l]$ compensated by $\lambda J[k,l]$. $T_s[k,l]$ and $\tilde{T}_s[k,l]$ represent the estimated and actual served time of the l-th frame in the k-th GoP. The served time vary according to the distribution of the video frame. Similar to the compensation of the arrival time, we use $\beta\tilde{L}[k,l]$ to compensate the variety of the served time, where $\tilde{L}[k,l]$ is the length of video frame and is the scaling factor, which is use to scale length of awake interval. The length of awake interval gets longer when the scaling factor gets higher. Therefore, $F_s[k,l]$ measures the difference between the estimated and actual served time compensated by $\beta\tilde{L}[k,l]$. For instance, if $F_a[k,l] = \phi, \forall(\phi > 0) \cap (\alpha > 0)$, and $F_s[k,l] = \theta, \forall(\theta < 0) \cap (\beta > 0)$, $\max_{ij} R_{ij}[k,l] = R_{13}[k,l]$. Thus, the agent will select the action A_{13} with the highest probability $\epsilon/v + 1 - \epsilon$. If $A_{13}[k,l]$ is selected, the agent transitions from the current state, $S_{i^*j^*}[k,l]$, to the next state $S_{13}[k+1,l]$. Thus, the Wi-Fi direct device will reduce the arrival time and increase the served time to receive the l-th frame in the next $k+1$-th GoP. The method to reduce the arrival time and increase the served time is given in (2) and (3), where i and j are set to 1 and 3, respectively. In sum up, the Wi-Fi direct device in the current state has to select an optimal action from the action spaces, which maximizes the reward. In our design, the maximum reward is equal to zero ($\max_{ij} R_{ij}[k,l] = R^*_{ij}[k,l] = 0$), and it can be achieved when $F^*_a[k,l] = 0$ and $F^*_s[k,l] = 0$. That means the agent tries to minimize the error between the estimated and actual arrival time compensated by the variety of the network delay jitter, and simultaneously minimizes the error between the estimated and actual served time compensated by the various service time.

In our study, we calculate the Q-value function as the function of the current reward and the maximum Q-value function of the previous. In the practical network, when the Wi-Fi direct device transmits the video frame in real-time, it can store the previous information including the start/end time of awake, the length of awake interval, state, action, reward, and Q-value function. Then, the Q-value function can be updated according to that previous information. From [21] and [23], the one-step Q-value of the current action is defined as

$$Q_{ij}[k,l] = R_{ij}[k,l] + \gamma \max_{A_{ij}[k-1,l]\in\mathcal{A}} \big(Q_{ij}[k-1,l]\big), \quad (6)$$

where γ is the discount factor and $\max_{A_{ij}[k-1,l]\in\mathcal{A}} Q_{ij}[k-1,l]$ is the maximum Q-value of the previous action. From [23], the ϵ-greedy policy is used to select the greedy action, $A_{i^*j^*}[k,l]$, with probability $\pi(A_{ij}[k,l]|S_{i^*j^*}[k,l])$, which is defined as

$$\pi(A_{ij}[k,l]|S_{i^*j^*}[k,l]) = \begin{cases} \epsilon/v + 1 - \epsilon & \text{if} \quad A_{i^*j^*}[k,l] = \text{argmax}_{A_{ij}[k,l]\in\mathcal{A}} Q_{ij}[k,l], \\ \epsilon/v & \text{otherwise}, \end{cases} \quad (7)$$

where $v = 9$ represents the total number of actions. The proposed RLPS algorithm is summarized in **Algorithm 1**.

Algorithm 1 Q-learning for action i and state j selection in order to compute $T_a[k+1,l]$ and $T_s[k+1,l]$, where are are given in Eq. 2 and 3, respectively.

Initialize $Q_{ij}[0,l]$, $T_a[1,l]$, and $T_s[1,l]$
for $k = 1 : N_I$
1. Take action $\mathcal{A}$, observe reward $R_{ij}[k,l]$
2. Compute $Q_{ij}[k,l]$, where is given in Eq. 6.
3. Choose a greed action $A_{i^*j^*}[k,l]$ from the action space using policy, which is derived from Eq. 7
4. Obtain the i^* and j^* according to the selected greedy action $A_{i^*j^*}[k,l]$
5. Compute $T_a[k+1,l]$ and $T_s[k+1,l]$, where are given in (2) and (3), respectively.

end for

A. OUTAGE PROBABILITY

As discussion in the section II-B, the loss of I-frame results in loss of all frames in the GoP, and the loss of P-frame results in the loss of the following P-frame and B-frame. But the loss of B-frame does not influence to other frames in the GoP. To reduce the outage probability, the remaining fraction of an I-frame (P-frame) is concatenated with the following B-frame and transmitted along with that B-frame during the next awake interval. In addition, the scaling factor (β) given in (3) is used to adjust the length of awake interval. The increment of the scaling factor results in increasing the length of the awake interval, and thus reducing the outage probability. Therefore, the performance evaluation of the proposed method in terms of the outage probability and the scaling factor (or the length of awake interval) is very essential in that it can verify how much the outage probability can be achieved for a given the scaling factor. Therefore, This subsection describes the method of obtaining the outage probabilities of I-frames, P-frames, and B-frames. To obtain the outage probability of video frame, we first determine the residual frame that cannot be completely transmitted in its awake interval. Let $F_R[k,l]$ be the residual frame, which is given by

$$\begin{aligned} F_R[k,l] = &\left(\tilde{L}[k,l] - L[k,l]\right) \mathbb{I}\left(\tilde{L}[k,l] > L[k,l]\right) \\ &\mathbb{I}\left(\tilde{T}_a[k,l] < T_a[k,l]\right) + \left(\tilde{T}_s[k,l] - T_s[k,l]\right) \\ &\mathbb{I}\left(\tilde{T}_s[k,l] > T_s[k,l]\right) \mathbb{I}\left(\tilde{T}_a[k,l] \geq T_a[k,l]\right), \end{aligned} \tag{8}$$

where $\mathbb{I}(.)$ is the indicator function. $\tilde{L}[k,l]$ is the length of the l-th frame in the k-th GoP, which is calculated by

$$\tilde{L}[k,l] = \tilde{T}_s[k,l] - \tilde{T}_a[k,l] = \frac{Z[k,l]}{D_R}, \tag{9}$$

where $Z[k,l]$ is the size of the l-th frame in the k-th GoP in [bits], and D_R is the data rate in [bps]. The outage probability of an I-frame is the percentage of residual I-frames that cannot be wholly transmitted in the following B-frame interval. This outage probability is expressed as

$$P_I = \frac{1}{N_I} \sum_{k=1}^{N_I} \mathbb{I}\left(F_R[k,1] > L[k,2]\right), \tag{10}$$

where N_I is the total number of I-frames. The loss probability of the P-frame is defined as

$$P_P = 1 - \tilde{P}_P, \tag{11}$$

where $\tilde{P}_P$ is the probability that the P-frame is fully transmitted. According to the inter-dependency between video frames in the GoP structure and the priority-based frame transmission strategy, a current P-frame is completely transmitted when the previous I-frame and P-frame are successfully transmitted, and the length of the current P-frame is shorter than the following B-frame interval. The successful transmission probability of the P-frame is defined as

$$\begin{aligned} \tilde{P}_P = &\frac{1}{N_P} \sum_{k=1}^{N_I} \sum_{l=1}^{(n/m)-1} \prod_{r=1}^{l} \mathbb{I}\left(F_R[k,1] \leq L[k,2]\right) \\ &\mathbb{I}\big(F_R[k,rm+1] \leq L[k,rm+2]\big), \end{aligned} \tag{12}$$

where $N_P = \left(\frac{n}{m} - 1\right) N_I$ is the total number of P-frames and $L[k,rm+2]$ is the length of the awake interval allotted to the combined residual P-frame and B-frame.

The outage probability of B-frames is defined as the ratio of the number of lost B-frames to the total number of B-frames, which is given by

$$P_B = \frac{N_B^{Loss}}{N_B} = \frac{N_B - N_B^{Suc}}{N_B}. \tag{13}$$

N_B^{Loss} denotes the number of lost B-frames, N_B^{Suc} represents the number of B-frames that are successfully transmitted, and N_B is the total number of B-frames. The successful transmission of B-frames occurs in three cases. First, a B-frame is successfully transmitted when the length of the combined residual I-frame and B-frame is shorter than the length of the awake interval allotted to these combined frames, which is given by

$$N_{I_R B}^{Suc} = \sum_{k=1}^{N_I} \mathbb{I}\Big(F_R[k,1] + \tilde{L}[k,2] \leq L[k,2]\Big). \tag{14}$$

Second, a B-frame is fully transmitted if the I-frame and previous P-frame are successfully transmitted and the length of the combined residual P-frame and B-frame is shorter than that of the awake interval allotted to these frames. In this case, the number of B-frames that are successfully transmitted is given by

$$\begin{aligned} N_{P_R B}^{Suc} = &\sum_{k=1}^{N_I} \sum_{\ell=1}^{(n/m)-1} \prod_{r=1}^{\ell} \mathbb{I}\left(F_R[k,1] \leq L[k,2]\right) \\ &\mathbb{I}\big(F_R[k,rm+1] \leq L[k,rm+2]\big) \\ &\mathbb{I}\Big(F_R[k,\ell m+1] + \tilde{L}[k,\ell m+2] \leq L[k,\ell m+2]\Big). \end{aligned} \tag{15}$$

Finally, a B-frame that is not combined with the residual I-frame (or P-frame) is successfully transmitted when the length of this B-frame is shorter than its awake interval and the previous I-frame and P-frame are fully transmitted. Thus, the number of B-frames that are successfully transmitted is calculated as

$$\begin{aligned} N_{NB}^{Suc} = &\sum_{k=1}^{N_I} \mathbb{I}\left(F_R[k,1] \leq L[k,2]\right) \Bigg(\sum_{\ell=1}^{(n/m)-1} \sum_{l=3+(\ell-1)m}^{\ell m} \prod_{r=1}^{\ell} \\ &\mathbb{I}\big(F_R[k,rm+1] \leq L[k,rm+2]\big) \mathbb{I}\Big(\tilde{L}[k,l] \leq L[k,l]\Big) \\ &+ \sum_{l=n-m+3}^{n} \prod_{r=1}^{(n/m)-1} \mathbb{I}\big(F_R[k,rm+1] \leq L[k,rm+1]\big) \\ &\mathbb{I}\Big(\tilde{L}[k,l] \leq L[k,l]\Big) \Bigg). \end{aligned} \tag{16}$$

Finally, the total number of B-frames that can be fully transmitted is defined as

$$N_B^{Suc} = N_{I_RB}^{Suc} + N_{P_RB}^{Suc} + N_{NB}^{Suc}. \tag{17}$$

Therefore, the outage probability of B-frames can be derived using (13)-(17).

B. TRANSMISSION DELAY

The metric "transmission delay" is very essential to evaluate how long we ought to wait for transmitting the remaining fraction of the I-frame (P-frame) during the next awake interval. On the other hand, in the case that the video frame arrives before the scheduled time, the Wi-Fi direct device has to wait until the scheduled time to initiate the frame transmission. This transmission delay is caused by the network delay jitter. Thus, the performance metric of transmission delay is very crucial to verify the performance gain of our proposed algorithm over traditional NoA method. The network delay jitter in the case that a video frame arrives before the scheduled time is given by

$$D_J = \frac{1}{N}\sum_{k=1}^{N_I}\sum_{\ell=1}^{n}\Big(T_a[k,\ell] - \tilde{T}_a[k,\ell]\Big)\mathbb{I}\Big(T_a[k,\ell] > \tilde{T}_a[k,\ell]\Big). \tag{18}$$

Let T be the length of the inter-awake interval. The transmission delay that is caused by the residual I-frame and P-frame is defined as

$$D_{wait} = \frac{1}{N_I + N_P}\sum_{k=1}^{N_I}\Bigg(\big(T - L[k,1]\big)\,\mathbb{I}\left(R[k,1] > 0\right) + \sum_{r=1}^{n/m-1}\big(T - L[k,rm+1]\big)\,\mathbb{I}\left(R[k,rm+1] > 0\right)\Bigg). \tag{19}$$

Hence, the average total transmission delay of a frame is given by

$$D_T = D_J + D_{wait}. \tag{20}$$

C. ENERGY CONSUMPTION

This metric is very significant because we can calculate how much the device consumes the energy for receiving one video stream in real-time. Most importantly, we can verify that our proposed algorithm can reduce much energy consumption of the device compared to the traditional NoA power-saving method. In addition, the method used to measure the energy consummation of the proposed RL power saving method is given in the section III-C. The average energy consumption is defined as the sum of the energy consumed during the awake and sleep intervals plus an additional energy, which is consumed to switch from the sleep to the awake mode.

The power saving method enables the Wi-Fi direct device to power off the circuitry to save the energy consumption. Thus, the energy consumption varies according to the length of sleep/awake interval. The metric "energy consumption" is very significant because we can calculate how much the device consumes the energy for receiving one video stream in real-time. Most importantly, we can verify that our proposed algorithm can reduce much energy consumption of the device compared to the traditional NoA power-saving method. The method to obtain the energy consumption is described as follows. The average energy consumption is defined as the sum of the energy consumed during the awake and sleep intervals, and the additional energy used to switch from the sleep to the awake mode. The average energy consumption during an awake interval is defined as

$$E_{awake} = \frac{P_{awake}}{N}\sum_{k=1}^{N_I}\sum_{l=1}^{n} L[k,l], \tag{21}$$

where $N = N_I + N_P + N_B$ is the total number of frames of an MPEG-2 video. The average energy consumption during a sleep interval is determined as

$$E_{sleep} = \frac{P_{sleep}}{N}\sum_{k=1}^{N_I}\sum_{l=1}^{n}\big(T - L[k,l]\big). \tag{22}$$

The average energy consumption of frame transmission is defined as

$$E_{average} = E_{awake} + E_{sleep} + E_{switch}, \tag{23}$$

where E_{switch} is the total energy required to switch frames from the sleep mode to the awake mode.

IV. PERFORMANCE EVALUATION

For performance evaluation, we decoded the movie titled 'Jurassic World (2015)' using the Elecard StreamEye Studio software, which is a video quality test software for the analysis of stream structures and the inspection of code parameters [24]. The GoP structure for this video is encoded as M3N30. The standard of this video is MPEG-2, which requires a frame rate of 24fps to support a resolution of 1920×1080 [25]. It is noted that the frame rate of 24fps is equivalent to an inter-frame interval of 41.7ms when Wi-Fi Direct devices use the 802.11ac standard, which achieves a high PHY data rate of 58.5 Mbps using a channel bandwidth of 160 MHz along with a BPSK modulation scheme and a code rate of 1/2 [26]. The power consumption during the awake and sleep intervals is set to 432mW and 0.3mW, respectively [27]. The energy required to switch from the sleep to the awake mode is 0.6mJ [28]. In addition, choosing too small value of α may result in that the algorithm may converge very slow. In addition, choosing too large value of α may cause the algorithm to overshoot the optimal solution and diverge. So, we choose an appropriate learning rate, which causes the start/end time of awake interval to guarantee the convergence with a better convergent speed, according to the observation of the simulation results. To evaluate the performance of the proposed method, we set the value of α to 0.2. we set the value of discount factor to 0.1 because the immediate rewards are very important to be used to predict the start time and length of awake interval of the future. In addition, we aim to scale down the previous Q-value to avoid the problem of

TABLE 1: Simulation parameters

Parameters	Values	Parameters	Values
N_I	1619	N_P	46951
N_B	97140	Frame rate	24fps
T	41.7ms	D_R	58.5 Mbps
γ	0.1	ϵ	0.2
α	0.2	λ	0.5
UDP-jitter	3.8 ~ 4.4 ms	P_{sleep}	0.3mW
P_{awake}	432mW	E_{switch}	0.6mJ

divergence to negative infinity. All parameters used in the simulation runs were summarized in Table 1.

A. PERFORMANCE EVALUATION OF THE PROPOSED RLPS METHOD

We consider that two Wi-Fi Direct devices are initiated when a GO sends a beacon message to a client to schedule the transmission time. Even though there is a scheduled time for the client to transmit a frame, the actual frame arrival time may shift forward with a UDP-jitter varying from 3.8 to 4.4 ms [29]. Hence, the client uses the proposed RLPS method with a delay factor (λ) of 0.5 to predict the arrival time of the frame to reduce network delay jitter. In the first beacon interval, the lengths of awake and sleep intervals are set as equal, and the client wakes up according to the scheduled time. In the next beacon interval, the GO and client simultaneously use the proposed RLPS method to predict the start and end points, and the length of awake intervals to transmit and receive a frame, respectively.

Fig. 4 shows the performance of the proposed RL power saving method in term of the average delay, outage probability, and energy consumption of a frame under two different movies. The result verify that the average delay and outage probability decrease as the energy consumption increases. The increasing energy consumption is caused by scaling up the length of awake interval, which results in reducing the outage probability because most of the video frames can be wholly transmitted during the current awake interval. Furthermore, scaling up the awake interval length also decreases the average delay because the number of the residual frames that have to wait to be transmitted during the next awake interval is decreased. Fig. 4 shows the average delay and outage probability of a frame for the movies titled 'Amazing Mary Gifted' and 'Jurassic World', assuming the same simulation parameters. Since 'Jurassic World' has more active scenes than 'Amazing Mary Gifted', the average frame size of 'Jurassic World' is larger than that of 'Amazing Mary Gifted', as shown in Table 2. As a result, Fig. 4 shows that the average delay and outage probability of 'Amazing Mary Gifted' is lower than that of 'Jurassic World'.

TABLE 2: The average frame size [bytes]

Movies	I-frame	P-frame	B-frame
A. M. Gifted	98638	20061	6636.7
J. World (2015)	128250	33772	11353

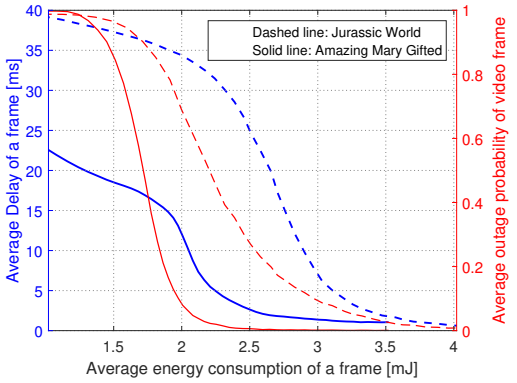
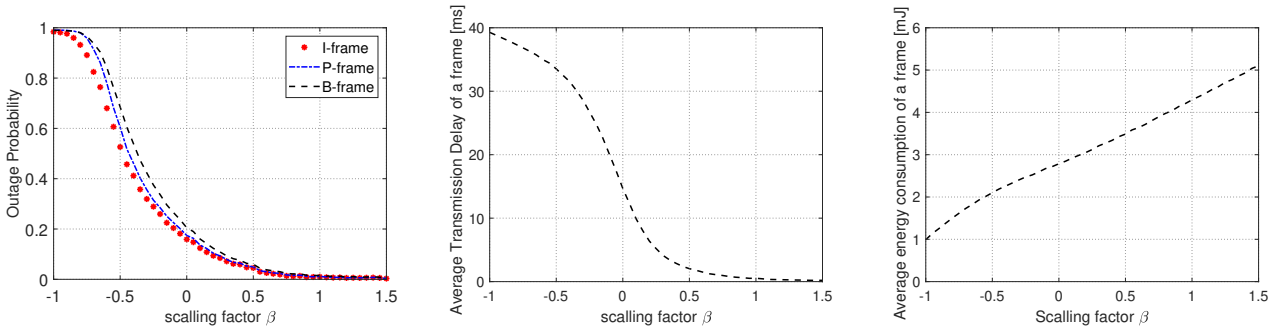


FIGURE 4: The dashed and solid lines represent the movie titled 'Jurassic World' and 'Amazing Mary Gifted', respectively.

Figure 5a shows the comparison between the outage probabilities of I-, P-, and B-frames as a function of scaling factor (β). Since the loss of an I-frame results in the loss of all frames in the GoP, the outage probability of I-frames is lower than that of P-frames and B-frames. The loss of B-frames is caused by the loss of P-frames; hence, the outage probability of B-frames is higher than that of P-frames. The proposed RLPS method uses a coefficient to scale the length of awake intervals, which increases with the value of the coefficient. Hence, the outage probability of frame transmission decreases as the scaling factor increases, as shown in Figure 5a. Figure 5b shows the average transmission delay as a function of the scaling factor. As the scaling factor increases, the transmission delay of a frame decreases because the length of awake intervals increases. Figure 5c shows the average energy consumption as a function of the scaling factor. The energy consumption increases as the scaling factor increases because the energy consumption during awake intervals is considerably higher than that during sleep intervals.

B. PERFORMANCE EVALUATION OF THE EXISTING NOA POWER-SAVING MODE

In the existing NoA power-saving mode, a GO sends a beacon message to a client. The message includes the start point of awake intervals, and the number and length of awake/sleep intervals. The length of awake intervals in each beacon interval are set to be equal. The client periodically wakes up based on the scheduled time, which is received from the GO. Network delay jitter is measured from the start point of awake intervals to the actual frame arrival time. The delay of I-frame (P-frame) transmission is measured from the end point of the I-frame (P-frame) interval to the start point of the following B-frame interval if the I-frame (P-frame) cannot be fully transmitted during the I-frame (P-frame) interval. The delay is zero if the I-frame is fully transmitted. The energies consumed during awake and sleep intervals are set to 432mW and 0.3mW, respectively. An energy of 0.6 mJ is required to switch from the sleep mode to the



(a) (b) (c)

FIGURE 5: RLPS method: (a) the outage probabilities of I-frames, P-frames, and B-frames, (b) the average transmission delay of a frame, and (c) the average energy consumption of frame transmission vs. scaling factor β.

awake mode. The outage probability is computed using the algorithm described in subsection III-A, where the lengths of awake intervals, $L[k, l]$, are set to be equal.

Figure 6a shows the comparison of the outage probabilities of I-frames, P-frames, and B-frames under various length of the awake interval. The outage probabilities of I-frames, P-frames and B-frames decrease as the length of awake intervals increases. The increase in the length of awake intervals decreases the probability that the residual of I-frames and P-frames will occur. This reduces the transmission delay of a frame, as shown in Figure 6b. Figure 6c verifies that the increment of the length of awake interval results in linearly increasing the average energy consumption of frame transmission.

C. THE COMPARABLE RESULTS OF THE PROPOSED RLPS AND THE EXISTING NOA POWER-SAVING METHODS

Figure 7 shows the average transmission delay and the outage probability, respectively, as a function of the average energy consumption. The dashed and solid lines represent the performances of the proposed RL and the NoA methods, respectively. The curves in this figure are highlighted in difference colors to clearly distinguish between the metrics “average delay of a frame" and “average outage probability of a video frame". The blue line indicates the metric “average delay of a frame", whereas the red line shows the average outage probability. The results show that the performance of the proposed RLPS method is better than that of the existing NoA power-saving method.

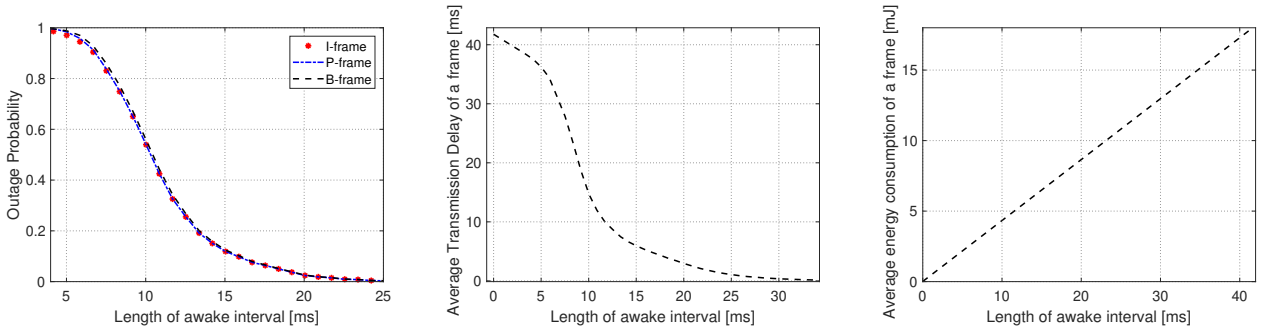
Fig. 8 shows the comparable results between the proposed RL and NoA power-saving methods in terms of the average delay of a frame and the frame rate. In this simulation, we assume that the frame rate of the video varies from 16 to 32 frames per second. Here, it is noted that when the number of frame transmissions per second increases, the length of an inter-frame interval decrease, which results in reducing the time delay for transmitting the residual I-frame and P-frame. Therefore, the average delay of a frame decreases as the number of frame transmissions per second increases, as shown in Fig. 8 below. Most importantly, the simulation result verifies that the performance of the proposed RL method is better than that of the existing NoA method.

Fig. 9 show the comparison result of the proposed RL and exiting NoA methods in terms of the average delay jitter and delay factor λ. In our study, we assume that the actual frame arrival time may shift forward with a UDP-jitter varying from 3.8 to 4.4 ms [29]. Since the existing NoA power-saving method fixes the start time of awake, the average jitter delay of this method is equal to 4.0993 ms. The delay factor λ here is used to compensate for the random delay jitter. Thus, the delay jitter varies according to the value of λ. Our proposed method aims to reduce the start time of awake when we increase the value of the delay factor to ensure that the device can wakes-up before the video frame arrives the destination. Thus, the delay jitter decreases as the delay factor increases. According to the result illustrated in Fig 9, we can verify that the delay jitter of the proposed method is less than that of the existing NoA method.

Fig. 10 shows the comparison results between the proposed RL, NoA, and EM methods in terms of the average delay and energy consumption of a frame. The results verify that the performance of the proposed algorithm is better than that of EM method. The degraded performance of the EM method may be caused by two main reasons. First, although the EM method in [20] can predict the statistical distribution of each video frame class (i.e., I-, P-, or B-frame class) and adjust the length of awake interval accordingly, the length of awake intervals scheduled for the video frame transmissions in a same class are equal to each other. Second, the study in [20] only focuses on the method to regulate the length of awake interval without considering the network delay jitter.

V. CONCLUSION

In this paper, we proposed the RL-based power-saving algorithm which adjusts the scheduling of awake intervals (the start/end time and lengths of awake intervals) according to the frame arrival time and the frame sizes. We designed the RLPS algorithm considering the network delay jitter and evaluated the performance of this proposed method using the



(a) (b) (c)

FIGURE 6: NoA power-saving method: (a) the outage probabilities of I-frames, P-frames, and B-frames, (b) the average transmission delay of a frame, and (c) the average energy consumption of frame transmission vs. the length of awake interval.

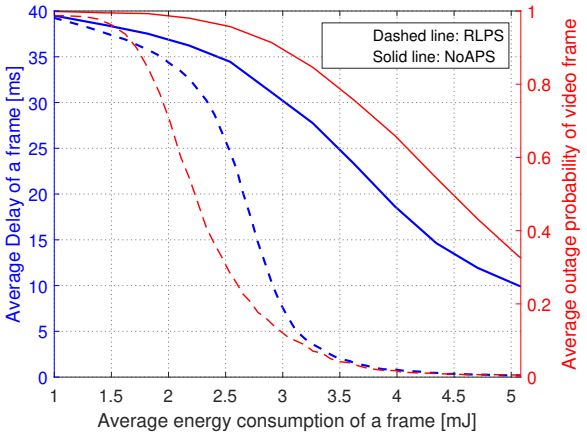


FIGURE 7: Comparison of RLPS and NoA power-saving methods: the dashed and solid lines represent the performance of the RLPS and existing NoA methods, respectively, in term of the average delay (blue line) and outage probability (red line).

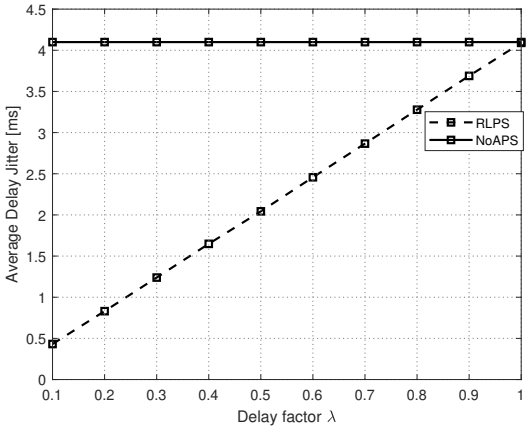


FIGURE 9: Comparison of RLPS and NoA power-saving methods in term of the average delay jitter and the delay factor.

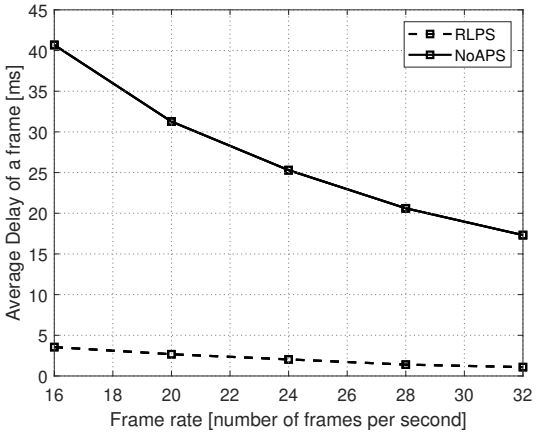


FIGURE 8: Comparison of RLPS and NoA power-saving methods in term of the average delay and the frame rate.

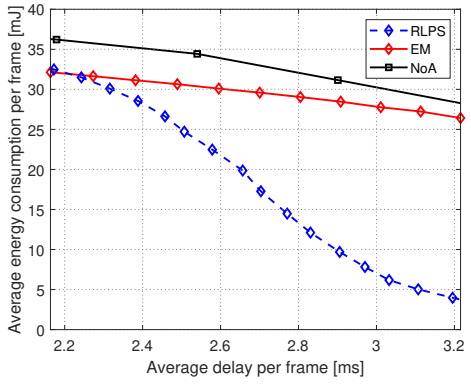


FIGURE 10: Comparison results between the proposed RL, NoA, and EM methods.

video quality test software. Simulation results showed that the proposed RLPS method outperforms the existing NoA power-saving method in terms of the outage probability, average delay, and energy consumption of frame transmission. According to the survey on Wi-Fi direct in [30], M. A. Khan et al. reveal that the Wi-Fi Direct devices establish the communication by discovering each other, setting up the security and the IP configuration, and implementing the power-saving protocol. Therefore, we build a plan to propose the machine learning algorithm-base latency and energy minimization in Wi-Fi direct with taking the device discovery, the security setup, the IP configuration, and the power save protocol implementation into consideration as the future work. In addition, unmanned aerial vehicles (UAVs) have attracted the research attention in the last decades because their mobility makes them to be easy deployment over every location and they can also establish line of sight (LOS) links with the users. The challenge remaining for UAVs is battery-limited. Therefore, M. A. Khan et al. employed the existing NoA power-saving method to improve the energy efficiency of UAVs and client association [6]. The performance degradation of the existing NoA method is caused by the fixed length of awake interval that is scheduled for packet transmission or reception. The results in our study verified that the length of awake interval should be adjusted according to the variation of the packet size to improve the performance of the existing NoA method. Therefore, we build a plan to propose the RL power-saving method-based energy efficiency maximization in UAVs communication network as the future work. Device-to-device (D2D) communication has been emerged to offer many advantages for cellular networks, such as enhancing energy efficiency, offloading the overloaded cellular traffic, wide cellular coverage, reducing delay or latency, and higher spectral efficiency [31]. However, the challenge remaining for D2D communication is the co-channel interference due to the coexistence of the direct D2D communications in the same frequency band. In addition, to enhance the performance of the networks, the D2D device should make a decision to establish the direct communication or indirect communication mode based on the channel condition. The indirect communication mode refers to the technique that uses the base station (BS) as a relay between D2D transmitter and receiver. However, BS may have insufficient to provide the wireless service for the D2D users in the disaster area, because it has a difficulty to be deployed in that area. Unlike BS, unmanned aerial vehicles (UAVs) is feasible and easy to deploy in various scenarios to provide the wireless service for users. Therefore, UAVs can act as a relay instead of the BS to forward the message from the D2D transmitter to receiver. Therefore, we will build a plane to study the application of RL to the joint optimization problem of mode selection, trajectory, and resource management for D2D communication underlaid multi-UAVs for the purpose of maximizing the energy efficiency.

REFERENCES

[1] A. Asadi, Q. Wang, and V. Mancuso, "A Survey on Device-to-Device Communication in Cellular Networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1801 - 1819, 2014.

[2] D. Camps-Mur, A. Garcia-Saavedra, and P. Serrano, "Device-to-device communications with Wi-Fi Direct: overview and experimentation," *IEEE Wireless Commun.*, vol. 20, no. 3, pp. 96 - 104, June 2013.

[3] Wi-Fi Alliance, Wi-Fi Protected Setup Specification v2.0.8, 2020.

[4] K.-W. Lim, W.-S. Jung, H. Kim, J. Han, and Y.-B. Ko, "Enhanced Power Management for Wi-Fi Direct," *IEEE Wireless Commun. Networking Conf.*, Shanghai, China, Apr. 2013.

[5] K.-W. Lim, Y. Seo, Y.-B. Ko, J. Kim, and J. Lee, "Dynamic power management in Wi-Fi Direct for future wireless serial bus", *Wireless Networks*, vol. 20, pp. 1777–1793, 2014.

[6] M. A. Khan, R. Hamila, M. S. Kiranyaz and M. Gabbouj, "A Novel UAV-Aided Network Architecture Using Wi-Fi Direct," *IEEE Access,* vol. 7, pp. 67305-67318, 2019.

[7] K.-W. Lim, W.-S. Jung, Y.-B. Ko, "Energy efficient quality-of-service for WLAN-based D2D communications", *Ad Hoc Networks*, vol. 25, pp. 102-116, 2015.

[8] H. Y., S. K., S. Lee, J.-Y. H., and D. Kim, "Traffic-aware parameter tuning for Wi-Fi direct power saving," *In Proc. 6th Int. Conf. Ubiquitous Future Networks*, Shanghai, China, 2014, pp. 1-2.

[9] C. Zhang, P. Patras, and H. Haddadi, "Deep Learning in Mobile and Wireless Networking: A Survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2224-2287, 2019.

[10] A. Geron, "Hands-On Machine Learning With Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, N. Tache, Ed., 1st ed. Sebastopol, CA, USA: O'Reilly Media, 2017, p. 543. [Online]. Available: http://shop.oreilly.com/product/0636920052289.do

[11] Y. Chu, P. D. Mitchell, and D. Grace, "ALOHA and Q-Learning based medium access control for Wireless Sensor Networks,"*Int. Symp. Wireless Commun. Systems Conf.*, Paris, French, 2012, pp. 511-515.

[12] M. Maalej, H. Besbes, and S. Cherif, "A cooperative communication protocol for saving energy consumption in WSNs," *3rd Int. Conf. Commun. Networking*, Hammamet, Tunisia, 2012, pp. 1-5.

[13] Z. Liu and I. Elhanany, "RL-MAC: A QoS-Aware Reinforcement Learning based MAC Protocol for Wireless Sensor Networks," *IEEE Int. Conf. Networking Sensing Control*, Ft. Lauderdale, FL, USA, 2006, pp. 768-773.

[14] D. Ye and M. Zhang, "A Self-Adaptive Sleep/Wake-Up Scheduling Approach for Wireless Sensor Networks," *IEEE Transactions on Cybernetics*, vol. 48, no. 3, pp. 979-992, March 2018.

[15] P. Verma, A. Dumka, D. Vyas, and A. Bhardwaj, "Reinforcement Learning based Node Sleep or Wake-up Time Scheduling Algorithm for Wireless Sensor Network" *Int. J. Math. Engineering Management Sciences,* vol. 5, no. 4, pp. 707-731, 2020.

[16] S. Sarwar, R. Sirhindi, L. Aslam, G. Mustafa, M. M. Yousaf and S. W. U. Q. Jaffry, "Reinforcement Learning Based Adaptive Duty Cycling in LR-WPANs," *IEEE Access*, vol. 8, pp. 161157-161174, 2020.

[17] M. F. Alam, M. Atiquzzaman, and M. A. Karim, "Traffic shaping for MPEG video transmission over the next generation internet," *J. Computer commun.*, vol. 23, no. 14-15, pp. 1336-1348, 30 Aug. 2000.

[18] A. Huszak and S. Imre, "Analysing GOP Structure and Packet Loss Effects on Error Propagation in MPEG-4 Video Streams," *4th Int. Symp. Commun. Control Signal Processing*, Limassol, Cyprus, 2010, pp. 1-5.

[19] M. Jin, J.-Y. Jung, and J.-R. Lee, "Dynamic Power-Saving Method for Wi-Fi Direct Based IoT Networks Considering Variable-Bit-Rate Video Traffic," *Sensors*, vol. 16, no. 10, Oct. 2016.

[20] D. Ron and J.-R. Lee, "Expectation Maximization Based Power-Saving Method in Wi-Fi Direct," *IEEE Access*, vol. 8, pp. 158600 - 158611, Aug. 2020.

[21] G. A. Rummery and M. Niranjan, "On-Line Q-Learning Using Connectionist Systems," Cambridge Univ. Engineering Dept., Cambridge, U.K., CUED/F-INENG/TR 166, 1994.

[22] M. Wiering and M. V. Otterlo, "Reinforcement Learning: State-of-the-Art", Berlin, Germany: Springer-Verlag, 2012.

[23] Sutton, R. S. and Barto, A. G., Reinforcement learning: An introduction, MIT press, 2018.

[24] Elecard, http://www.elecard.com/

[25] H. Guo, C. Zhu, S. Li, and Y. Gao, "Optimal Bit Allocation at Frame Level for Rate Control in HEVC," *IEEE Trans. Broadcasting*, vol.65, no. 2, pp. 270-281. June 2019.

[26] L. Verma, M. Fakharzadeh, and S. Choi, "WiFi on Steroids: 802.11ac and 802.11ad," *IEEE Wireless Commun.*, vol. 20, no. 6, pp.30-35, Dec. 2013.

[27] D. C.-M., X. P.-C., and S. S.-Ribes, "Designing energy efficient access points with Wi-Fi Direct," *J. Computer Networks*, vol. 55, no. 13, pp. 2838-2855, Sep. 2011.

[28] A. Bhojan and G. W. Tan, "Mumble: Framework for seamless Message Transfer on Smartphones," *1st International Workshop on Experiences with the Design and Implementation of Smart Objects*, Paris, France, 7–11 September 2015; pp. 43–48.

[29] J. A. R. Pacheco de Carvalho, H. Veiga, N. Marques, C. F. F. Ribeiro Pacheco, and A. D. Reis, "Performance measurements of IEEE 802.11 b, g laboratory WEP and WPA point-to-point links using TCP, UDP and FTP," *Int. Conf. Applied Electronics*, Pilsen, Czech Republic, 2011, pp. 1-6.

[30] M. A. Khan, W. Cherif, F. Filali, R. Hamila, "Wi-Fi Direct Research-Current Status and Future Perspectives", *Journal of Network and Computer Applications*, vol. 93, pages 245-258, 2017.

[31] M. S. M. Gismalla et al., "Survey on Device to Device (D2D) Communication for 5GB/6G Networks: Concept, Applications, Challenges, and Future Directions," *IEEE Access,* vol. 10, pp. 30792-30821, 2022.

DARA RON received the B.S. degree from the Department of Electrical and Energy Engineering, Institute of Technology of Cambodia (ITC), Phnom Penh, Cambodia, in 2017. He is currently pursuing the integrated M.S. and Ph.D. degrees with the School of Electrical and Electronics Engineering, Chung-Ang University, South Korea. His current research interests include bio-inspired algorithms, LoRaWAN protocol, and artificial intelligent-based wireless networks.

JUNG-RYUN LEE (Senior Member, IEEE) received the B.S. and M.S. degrees in mathematics from Seoul National University, in 1995 and 1997, respectively, and the Ph.D. degree in electrical and electronics engineering from the Korea Advanced Institute of Science and Technology (KAIST), in 2006. From 1997 to 2005, he was a Chief Research Engineer with LG Electronics, South Korea. From 2006 to 2007, he was a full time Lecturer of electronic engineering with the University of Incheon. Since 2008, he has been a Professor with the School of Electrical and Electronics Engineering, Chung-Ang University, South Korea. His research interests include energy-efficient networks and algorithms, bio-inspired autonomous networks, and artificial intelligence-based networking. He is a Regular Member of IEICE, KIISE, and KICS. He received the Excellent Paper Award at ICUFN 2012, the Best Paper Award at ICN 2014, the Best Paper Award at QSHINE 2016, and the Excellent Paper Award at ICTC 2018.