Optimal Channel Selection in Hybrid RF/VLC Networks: A Multi-Armed Bandit Approach

Mostafa M. Fouda, *Senior Member, IEEE,* Sherief Hashima, *Senior Member, IEEE,* Sadman Sakib,
Zubair Md Fadlullah, *Senior Member, IEEE,* Kohei Hatano, and Xuemin (Sherman) Shen, *Fellow, IEEE*

***Abstract*—We investigate optimal band/channel selection in hybrid radio frequency and visible light communication (RF/VLC) networks. Particularly, we first develop a robust hybrid RF/VLC based system model for the optimal band/channel selection. We then formulate it as an online stochastic budget-constrained multi-armed bandit (MAB) problem. Two online learning algorithms based on different optimal policies are proposed to choose the appropriate band, i.e., energy-aware band selection with upper confidence bound (EABS-UCB) and energy-aware band selection with Thompson sampling (EABS-TS). The cost/budget is the battery consumption of the transmitting device according to the selected band. Through extensive simulations, it is confirmed that the proposed EABS-TS emerges as the superior technique compared with the random, brute-force, and EABS-UCB band selection schemes, in terms of energy efficiency, average throughput, and convergence performance.**

***Index Terms*—B5G, 6G, millimeter wave, radio frequency (RF), multi-armed bandit.**

I. Introduction

To fulfill the increasing bandwidth demand of mobile users in the beyond fifth-generation (B5G)/sixth-generation (6G) networks, hybrid radio frequency (RF (including 2.4GHz, 5.25GHz, and 38GHz bands)) and visible light communication (VLC), i.e., (RF/VLC) frequency bands, are considered to provide ultra-high capacities and improved connectivity, respectively [1]–[4]. However, these high frequency bands suffer from fast channel fading, signal attenuation, blocking effect of walls and other obstacles, shadowing, limited range, and so forth [5]. Due to the presence of these heterogeneous frequency bands and their dynamically varying channel conditions, optimized channel selection emerges as a difficult-to-model problem.

An RF/VLC linkage selection scheme for data handling in indoor scenarios with quality of service (QoS) guarantee was proposed in [6]. Due to the complex nature of channel assignment in a given frequency band, several researchers resorted to employing AI models to allocate heterogeneous channels [7]–[10]. For instance [7] presented a deep learning model for RF/VLC band selection in a device-to-device (D2D) scenario based on a deep neural network (DNN). Since deriving a closed-form solution to such a problem is challenging, we survey artificial intelligence (AI) techniques, i.e., machine/deep learning considered in our earlier work [8] to address this issue. However, such models need extensive training data so that the trained model may be disseminated to the transmitter and receiver nodes to adaptatively switch to the best bands/channels. Besides, since the B5G/6G network traffic is anticipated to be highly dynamic in nature, such an AI model must be periodically updated. Regarding online channel allocation, researchers handled a different problem in terms of the joint bandwidth, power, and user association optimization problem in hybrid (RF/VLC) systems using deep Q-learning algorithm via targeting optimal policy learning [11].

In this paper, we provide a unified system model, which considers the co-existence of VLC, millimeter wave (mmWave), and wireless local area network (WLAN) bands impacted by variable blocking entities. Due to the high dynamism in such a hybrid RF/VLC system, the band/channel selection appears as a computationally hard problem. Specifically, we aim to design a practical, proactive online heterogeneous band/channel selection strategy to overcome the offline training drawbacks. Recently, multi-armed bandits (MAB) gained a lot of attention in solving distinct wireless communications problems such as multipath selection [12], fast mmWave alignment [13], D2D communications [14], concurrent beamforming [15], and UAV communications [16] due to its ultra fast decisions without the need for offline training [17]. MABs are lightweight stateless reinforcement learning type that can quickly adapt to dynamic environment, which highly fits our problem. The contributions of this paper can be summarized as follows:

- Motivated by evolving environment, we formulate the multi-band/channel selection problem as a stochastic bandits problem.
- We propose two online, energy-aware band selection (EABS) algorithms by customizing the upper confidence bound (UCB) and Thompson sampling (TS) al-

Corresponding author: M. M. Fouda.

M. M. Fouda is with the Department of Electrical and Computer Engineering, College of Science and Engineering, Idaho State University, Pocatello, ID 83209, USA, and the Department of Electrical Engineering, Faculty of Engineering at Shoubra, Benha University, Cairo 11629, Egypt. E-mail: mfouda@ieee.org

S. Hashima is with the Computational Learning Theory Team, RIKEN-Advanced Intelligent Project, Fukuoka 819-0395, Japan, and the Department of Engineering and Scientific Equipment, Egyptian Atomic Energy Authority, Cairo, Inshas 13759, Egypt (e-mail: sherief.hashima@riken.jp).

S. Sakib and Z. M. Fadlullah are with the Department of Computer Science, Lakehead University; and Thunder Bay Regional Health Research Institute (TBRHRI), Thunder Bay, Ontario, Canada (emails: ssak2921@lakeheadu.ca, Zubair.Fadlullah@lakeheadu.ca)

K. Hatano is with the Computational Learning Theory Team, RIKEN-Advanced Intelligent Project, Fukuoka 819-0395, Japan, and the Faculty of Arts and Science, Kyushu University, Fukuoka 819-0395, Japan (email: hatano@inf.kyushu-u.ac.jp).

X. Shen is with Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada (e-mail: sshen@uwaterloo.ca).

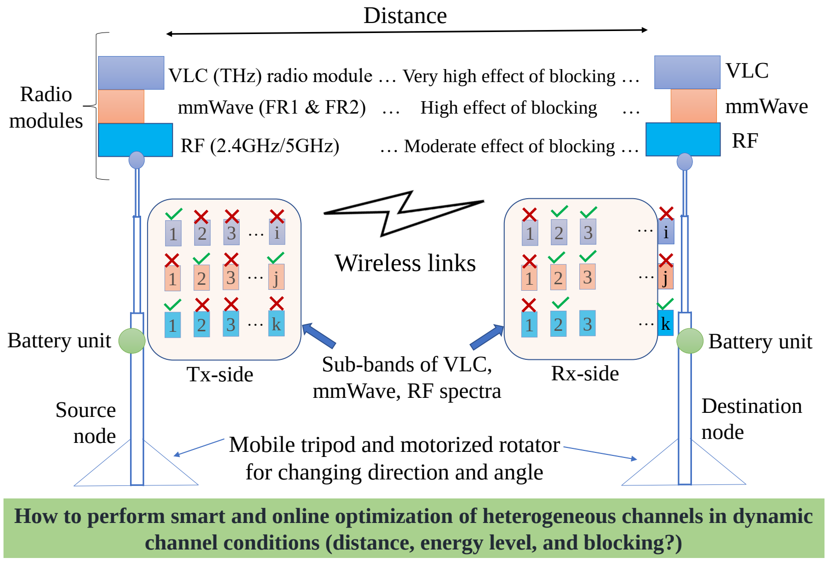

Distance
Radio modules
VLC (THz) radio module … Very high effect of blocking …
mmWave (FR1 & FR2) … High effect of blocking …
RF (2.4GHz/5GHz) … Moderate effect of blocking …
VLC
mmWave
RF
1 2 3 … i
1 2 3 … j
1 2 3 … k
Wireless links
Battery unit
Tx-side
Sub-bands of VLC, mmWave, RF spectra
Rx-side
Battery unit
Source node
Destination node
Mobile tripod and motorized rotator for changing direction and angle
How to perform smart and online optimization of heterogeneous channels in dynamic channel conditions (distance, energy level, and blocking?)

Fig. 1. Considered system model and problem of heterogeneous channel and corresponding sub-band selection.

gorithms [17], referred to as EABS-UCB and EABS-TS, respectively.

- Our proposed algorithms are energy-aware ones that can switch between WLAN, mmWave, and VLC bands adaptatively with different optimal policies to explore and exploit available bands and channels under dynamically changing environments under the effect of no, small, and large blocking entities.
- We conduct extensive simulations to demonstrate our proposals' viability compared to the traditional (e.g., random and brute-force) band/channel selection methods. The results clearly indicate that both EABS-UCB and EABS-TS outperform the traditional approaches with higher convergence speed.

II. System Model

Fig. 1 depicts our considered system model whereby only a pair of many possible source-destination (S-D) nodes is shown for simplicity. The S/D nodes can be mobile terminals, D2D relays, or base stations, equipped with hybrid RF (WLAN/mmWave)/VLC frequency bands.

First, we describe the RF channel model using IEEE 802.11ac/n (WLAN in 5.25GHz and 2.4GHz, respectively). Regarding the general RF channel model, we employ the linkage model of [14]. In WLAN operating at 5.25GHz, the received power (P_D^w in dB) of the receiver node is obtained from [14] with 2.32 path loss exponent value, zero mean log-normal shadowing, and a standard deviation of 6 dB. On the other hand, the 2.4GHz WLAN system parameters are path loss exponent of 2.32 and standard deviation of 2.15 dB [18].

Next, we present our considered mmWave linkage model using the 38GHz-based mmWave [14]. The mmWave received power jointly considers the beamforming gain and blockage effect as indicated in [14]. Other parameters of the model include the 38GHz mmWave source power, S/D beamforming gains, and the distance-dependent path loss (expressed in dB) at a reference distance. For the latter, log-normal shadowing with zero mean is considered.

Also, we focus on employing the line of sight (LoS) path since it is typically the most dominant technology in mmWave communications and the gain of the LoS path can be 20 dB higher than those of non-LoS (NLoS) [19]. For transmission/reception, the two-dimensional (2D), steerable antenna model with Gaussian main loop profile is utilized with a given azimuth angle, -3dB beamwidth, and the maximum antenna gain.

On the other hand, regarding the VLC channel model, we utilize the Lambertian model, which is suitable for the light emitting diode (LED) transmitters [1]. The LoS channel gain based on LED with Lambertian patterns is adopted considering the LED beam solid angle, LED beam with the maximum half-angle, the angle between the source receiver line and beam axis, and the angle between the source receiver line and receiver normal. For practically modeling the VLC channel conditions, the optical detector's physical area and the distance between the receiver and the transmitter are also considered. Using these parameters, we derive the gains of the optical filter and concentrator, and then employ these to calculate the received optical power using the obtained parameters.

Furthermore, regarding the impact of blockage, we adopt the urban scenario frequency dependent blocking model in [20] since it closely resembles our scenario. This scenario is almost static, where the source node is fixed and the destination node moves far away and the blocked object (small/large vehicle) is 15 meters away from the source. Hence, our generalized frequency-dependent blocking model is formulated as [20]:

$$Blockageloss[dB] = \beta_a + \alpha_a \log(1 + \frac{f_{c,n}}{1\text{GHz}}), \qquad (1)$$

where α_a, β_a are the slope and the intercept of the line plotted via linear regression, respectively. a indicates the index of the blocker type (i.e., small or large). This blockage loss can be anticipated to the path loss equation as a function of the operated frequency in addition to the blocker type.

III. Problem Description and Formulation

Herein, we formulate the problem of how to design a distributed online channel allocation algorithm in the highly complex hybrid RF/VLC network that can treat user devices as localized decision makers whereby each user may run and converge to the optimal radio resource allocation while achieving an optimal reward. As further constraints to the problem, the user devices may not exchange any information with one another, observe a single band/channel at a time, and sense whether there is a transmission on that channel without decoding transmissions or identifying the transmitting nodes. This can be modeled as follows

$$\arg\max_{i,j} \sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}\mathbb{E}(\psi_{ij}(t)) \tag{2a}$$

$$s.t.$$

$$\Xi_{\mathrm{S},ij}(t) > \Xi_{th}\ ,\forall s \in \{\text{Source nodes}\} \tag{2b}$$

$$\sum_{i=1}^{m} n_i \leq N \tag{2c}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij} = 1\ , \tag{2d}$$

$$x_{ij} \in \{0,1\}\ , \tag{2e}$$

where m and n_i denote the number of heterogeneous frequency bands and the number of channels in band i, respectively. N indicates the total number of available channels across all the bands. $\psi_{ij}(t)$ indicates the throughput in bps of the S-D link at time t utilizing channel j of band i. x_{ij} denotes a decision variable, based on which the optimal band and its corresponding channel is to be selected to maximize the aggregated, expected throughput $\mathbb{E}(\psi_{ij}(t))$. $\psi_{ij}(t)$ is given as follows:

$$\psi_{ij}(t) = \frac{B_{ij}T_D\Gamma_{ij}(t)}{U(t) * T_{h,ij} + T_D}, \tag{3}$$

where T_D refers to the data transmission time while $T_{h,ij}$ denotes the overhead time between the hybrid band S-D pair according to the selected frequency band. B_{ij} is the utilized bandwidth. $\Gamma_{ij}(t)$ is the link spectral efficiency (SE) in bps/Hz upon the chosen band/frequency at time t, which can be expressed as,

$$\Gamma_{ij}(t) = \log_2(1 + \frac{P_{\mathrm{D}}^{ij}(t)}{N_0 + I(t)}), \tag{4}$$

where $P_{\mathrm{D}}^{ij}(t)$ denotes the received power at D at time t according to the selected band, N_0 is the noise power at S, and I refers to the interference from nearby devices that utilize the same frequency. In this paper, we consider the interference issued from the two WLAN channels only. The mmWave and VLC systems are directional ones, hence their interference are negligible with respect to the random noise. Hence, there is a conflict between exploring better frequency bands/channels and exploiting the band that is considered to yield the maximum throughput and battery life. An online learning algorithm is required to address this tradeoff with self decision making and appropriate convergence speed. Hence, we relax the aforementioned research problem by reformulating it to a stochastic MAB to obtain ultra-fast decisions over time under the effect of uncertainty and dynamic blockage. MAB presents a robust algorithmic framework in contrast with comparable methodologies that efficiently handles the exploration-exploitation trade-off. The basic MAB framework consists of K possible actions (i.e., arms) to choose within T rounds. In each round $t \in T$, by choosing an arm, the learner collects a reward from the chosen arm. Two of the main types of MAB are stochastic and adversarial MABs. In stochastic MABs, the rewards are i.i.d (independent and identical distribution) drawn. In contrast, adversarial MABs removes all assumptions of reward distributions [17]. Next, $\Xi_{\mathrm{S},ij}(t)$ represents the residual energy (in Joules) of the source device S at time t upon the utilized channel j of band i, and Ξ_{th} is the energy-threshold after which the source devices may no longer be able to establish wireless links, and therefore, is compelled to save its power for its main activity. The energy update formula of the S node according to the chosen band is as follows:

$$\Xi_{\mathrm{S},n^*_{MAB}}(t) = \Xi_{\mathrm{S},n^*_{MAB}}(t-1) - \frac{P_{\mathrm{S}}^{n} L_D}{B_{n^*_{MAB}}\Gamma_{n^*_{MAB}}(t)}, \tag{5}$$

where MAB reflects the utilized MAB scheme (e.g., UCB or TS), $E_{\mathrm{S},n^*_{MAB}}(t-1)$ is the remaining energy of the hybrid-band S at the previous trial ($t-1$), and the term $P_{\mathrm{S}}^{n} L_D / B_{n^*_{MAB}}\Gamma_{n^*_{MAB}}(t)$ defines the energy consumption for transmitting the required data of L_D bits with a data rate of $B_{n^*_{MAB}}\Gamma_{n^*_{MAB}}(t)$ bps using the selected band n^*_{MAB}.

IV. Proposed Online EABS Algorithms

Herein, we present two online, energy-aware band selection (EABS) algorithms, referred to as EABS-UCB and EABS-TS, which operate with UCB and TS MAB policies, respectively.

1) Proposed EABS-UCB algorithm: Our proposed EABS-UCB algorithm uses the upper confidence bound policy, which aims to provide optimism under uncertain channel conditions across heterogeneous frequency bands. It attempts to address the problem arising from the explore-first approach where each arm is explored for the same number of rounds, causing inefficiency. Therefore, the exploration schedule should be based on the observed rewards history. Rather than employing the same confidence range for any arm in a given round, the UCB policy chooses the best arm based on an optimistic estimate [17]. During a specific round, let each arm's reward function be a point estimate based on the average rate of observed rewards. For each point estimate, an upper confidence boundary for each arm may be considered aiming to find the arm with the highest reward rate. At each round t, the UCB reward of all arms is the sum of the current reward return average of arm k at the current round and the number of pulls given to arm k in the history of observations. The best arm is selected as follows:

$$n^*_{EABS-UCB}(t) = arg\max_{n}\{\bar{\psi}_n(t-1) + \sqrt{\frac{2ln\ t}{M_{n,t-1}}} - \frac{d_n}{\Xi_{\mathrm{S},n}(t)}\}, \tag{6}$$

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2022.3163078

where $\bar{\psi}_n(t)$ denotes the average throughput obtained from the transmission band n until time t. $M_{n,t-1}$ refers to the number of times n has been picked until time t. A new term, $\frac{d_n}{\Xi_{\mathrm{Tx},n}(t)}$, (distance/remaining energy) is appended to the main UCB formula to reflect the battery consumption of the hybrid-band S relative to its distance from the hybrid-band D.

2) Proposed EABS-TS algorithm: Next, we propose EABS-TS algorithm by using the TS policy, where the samples from the rewards are taken sequentially to update posterior distribution. In other words, EABS-TS aims to maximize the immediate performance and invest in accumulating new information to improve the future performance on the heterogeneous channel selection. We used Gaussian-based TS because of the Gaussian distribution of the reward due to additive white Gaussian noise (AWGN) in the specified bandwidth of the heterogeneous frequency bands considered in our system model. For each observation obtained from an arm and its corresponding reward, a new distribution is generated with the probabilities of success of each arm. Then, further observations are made based on these prior probabilities captured during each round that is used to update the subsequent distributions. When sufficient observations are made, each arm ends up having its own success distribution that can help the EABS-TS algorithm to choose the suitable band and collect the maximum rewards. The EABS-TS main equation is introduced as follows:

$$n^*_{EABS-TS}(t) = arg\max_n\{\theta_n(t) - \frac{d_n}{\Xi_{S,n}(t)}\},$$
$$\theta_n(t) \sim \mathcal{N}(\bar{\psi}_n(t), \frac{1}{M_{n,t}+1}), \quad (7)$$

where $\mathcal{N}(\bar{\psi}_n(t), \frac{1}{M_{n,t}+1})$ is a normal distribution with $\bar{\psi}_n(t)$ mean and $\frac{1}{M_{n,t}+1}$ variance.

The steps of our proposed online algorithms are generalized and simplified in Fig. 2. At the beginning of both algorithms, we initialize to pull each arm (i.e., channel) for each arm ($n = N_{ch}$) and obtain the reward. If the joint conditions ($(N_{ch}+1) < t < T$) and ($E_s > E_{th}$) are not satisfied, the algorithm has no further step to execute and ends. Otherwise, it identifies the better reward (channel index) $n^*_{MAB}(t)$ in the time round $t \in T$ according to UCB and TS policies in EABS-UCB and EABS-TS, respectively. Then, the MAB parameters are updated and remaining energy levels of the user devices are calculated according to the selected band/channel. Then, the next time round starts repeating the aforementioned process when the source needs to send new data frames to destination.

V. Performance Evaluation

Table I lists our considered simulation parameters. The optimal strategy selects the ideal band from the first trial without searching other ones. However, this requires perfect channel state information (CSI) and blockage knowledge. Next, brute force band selection scheme chooses the best band and sub-channel after searching all the available ones, which costs a high decision time. Moreover, the random selection method is considered whereby the channel is picked randomly at each round t, without any channel quality consideration. The performance evaluation metrics are convergence rate, average throughput, and the average energy efficiency which is defined as the average throughput over energy expenditure per selected band/channel in bit/sec/joule.

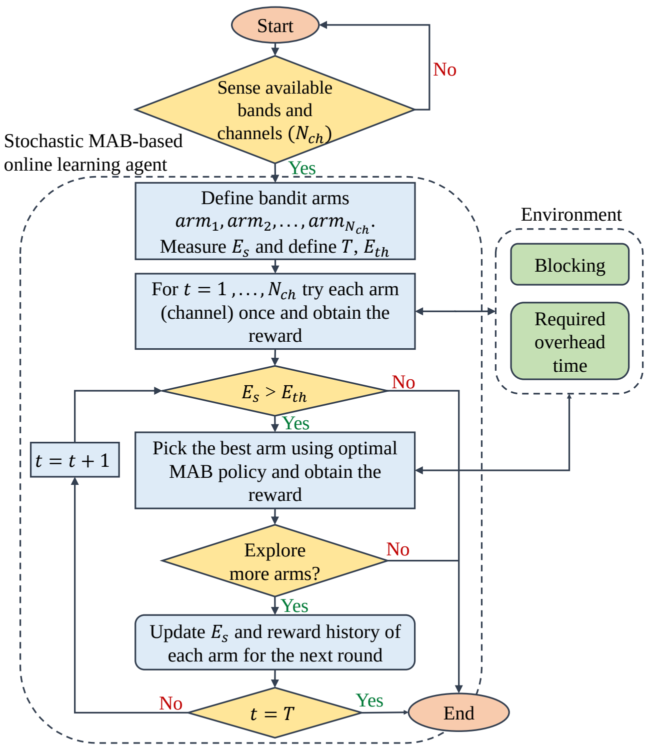

Start
Sense available bands and channels (N_{ch})
No
Yes
Stochastic MAB-based online learning agent
Define bandit arms $arm_1, arm_2, \ldots, arm_{N_{ch}}$. Measure E_s and define T, E_{th}
For $t = 1, \ldots, N_{ch}$ try each arm (channel) once and obtain the reward
$E_s > E_{th}$
No
Yes
Pick the best arm using optimal MAB policy and obtain the reward
$t = t + 1$
Explore more arms?
No
Yes
Update E_s and reward history of each arm for the next round
No
$t = T$
Yes
End
Environment
Blocking
Required overhead time

Fig. 2. Proposed MAB-based predictive channel assignment approach at the relay nodes for simultaneous transmission and reception from source node adopting multiple frequency band and multi-channel networks.

The convergence points for the proposed EABS-UCB and EABS-TS techniques are plotted in Fig. 3. The source-destination distance was set to 10 meters in this scenario. The simulation results indicate that the proposed EABS-TS algorithm demonstrated superior performance throughout all the rounds from $t = 1$ to 1000. When the value of round t approaches 500, the proposed EABS-TS technique converges to 99.5% of the average throughput of the optimal scenario, while EABS-UCB obtains 97.2% at that point. After complet-

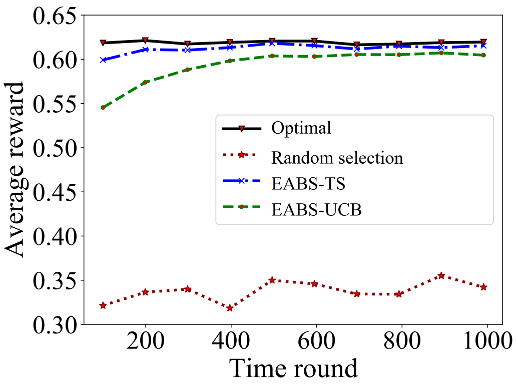
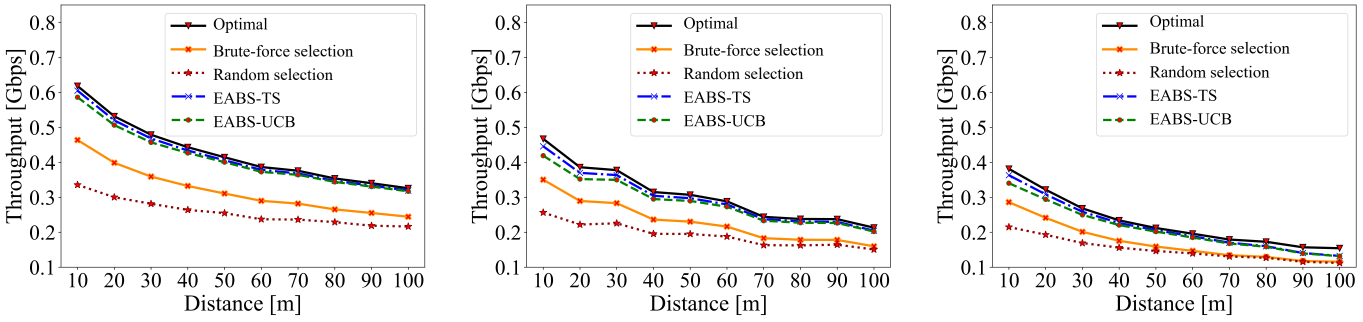


Fig. 3. Convergence rate evaluation.

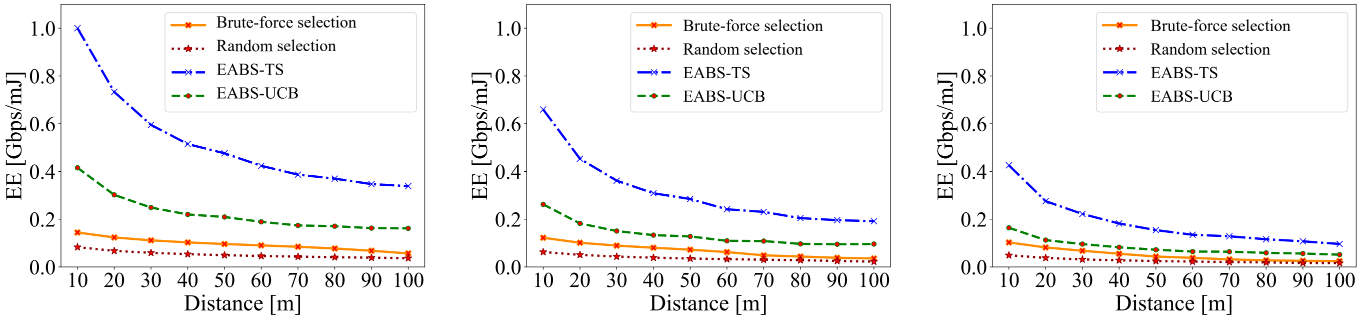


(a) No blockage scenario.

(b) Small blockage scenario.

(c) Large blockage scenario.

Fig. 4. Throughput comparison for distinct source-destination distances and blockage cases.



(a) No blockage scenario.

(b) Small blockage scenario.

(c) Large blockage scenario.

Fig. 5. Energy efficiency comparison for distinct source-destination distances and blockage cases.

TABLE I
SIMULATION PARAMETERS

Simulation parameters		Value
Number of channels		4 (WLAN 2.4GHz, 5.25GHz, mmWave, VLC)
Total rounds, data length, E_{th}		1000, 1TB, 1%
Operating frequencies of each channel		5.25, 2.4, 38, 10^5GHz
Bandwidth of each channel		40, 20, 100, 20 MHz
Distance between sender and receiver		{10 – 100} meters
Blocking model [20]	Small blocking: {length, width, height}	{5.07, 1.69, 1.93} meters
	Large blocking: {length, width, height}	{7.01, 2.04, 2.63} meters

ing just half of the total rounds, the encouraging performance indicates the viability of our proposed algorithms for multi-band, multi-channel selection in B5G/6G networks.

Figure. 4 demonstrates the throughput performance comparison of our proposed EABS-UCB and EABS-TS algorithms with the conventional band selection methods (optimal, brute-force, and random) against distinct source-destination separation distances at three different blocking scenarios (i.e., none, small, and large). For all methods, the average throughput achieved is decreased as the blocking ratio increased which is evident in the downward transition of the average throughput values due to blocked objects, i.e., no-blocking, small blocking, and large blocking, as shown in Figs. 4(a), 4(b), and 4(c), respectively. The proposed EABS-UCB and EABS-TS schemes demonstrated encouraging performance (near optimal one) in contrast with the brute force and random selection methods. Note that the obtained average throughputs across all considered distances have a very minimal difference compared to the optimal case. Compared to the optimal case, the EABS-TS and EABS-UCB algorithms achieved up to 97.89% and 96.17% average throughput, respectively, across all considered distances. Thus, EABS-TS emerges as the best method for the multi-band, multi-channel allocation in the hybrid RF/VLC network due to the Bayesian policy of TS. However, the random channel selection achieved only 60.28% of the average throughput across all distances. The random selection exhibited the worst performance with a comparatively large drop in the average throughput with increasing distances due to the random frequency band selection policy that might encounter poor channel conditions.

To indicate the applicability of our proposals in an energy-constrained situation, the adopted methods' average energy efficiency are determined and demonstrated in Fig. 5. Here, Figs. 5(a), 5(b), and 5(c) represent the considered case with no-blocking, small blocking, and large blocking, respectively. In all three scenarios, our proposed EABS-TS algorithm outperformed all the other techniques. This robust performance of EABS-TS can be credited to selecting suitable channels with the best band selection policy while considering the remaining energy of the device. Because of the path loss, all the compared methods' energy efficiency showed a decreasing trend as the distance between the source and destination nodes increased for all considered blocking scenarios. As the distance between the source and destination devices was increased, all the conventional methods' energy efficiency drastically declined, whereas the proposed methods persisted with considerably superior performance. Among all the compared methods, the random channel selection technique suffered from the least energy efficiency for all distances due to the inefficient selection of channels and lack of energy-awareness. Furthermore, the encouraging experimental results indicate the acceptability of the proposed EABS-TS technique in multi-band/channel selection B5G/6G networks.

VI. Conclusion

In this paper, we reformulated the computationally hard heterogeneous frequency band selection problem as a stochastic budget constrained MAB, and designed online, energy-aware algorithms for ultra-fast, self decision-making by the user devices to select the most appropriate band/channel. Our developed EABS-UCB and EABS-TS search optimal policies for WLAN, mmWave, or VLC band selection with energy-awareness to optimally maintain the battery life of the devices according to the selected band. Simulation results demonstrated the superior performance of our proposed algorithms over the conventional channel allocation methods. For the future works, we will study the multi-player scenarios and consider the highly dynamic environments, where adversarial MABs will be more applicable.

References

[1] H. Abuella, M. Elamassie, M. Uysal, Z. Xu, E. Serpedin, K. A. Qaraqe, and S. Ekin, "Hybrid RF/VLC systems: A comprehensive survey on network topologies, performance analyses, applications, and future directions," *IEEE Access*, vol. 9, pp. 160 402–160 436, 2021.

[2] S. Sakib, T. Tazrin, M. M. Fouda, Z. M. Fadlullah, and N. Nasser, "A deep learning method for predictive channel assignment in beyond 5G networks," *IEEE Network*, vol. 35, no. 1, pp. 266–272, 2021.

[3] B. Mughal, Z. M. Fadlullah, M. M. Fouda, and S. Ikki, "Allocation schemes for relay communications: A multiband multichannel approach using game theory," *IEEE Sensors Letters*, vol. 6, no. 1, Art no. 7500104, 2022.

[4] S. Sakib, T. Tazrin, M. M. Fouda, Z. M. Fadlullah, and N. Nasser, "An efficient and light-weight predictive channel assignment scheme for multi-band B5G enabled massive IoT: A deep learning approach," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5285–5297, 2021.

[5] Y. Chen, B. Ai, Y. Niu, R. He, Z. Zhong, and Z. Han, "Resource allocation for device-to-device communications in multi-cell multi-band heterogeneous cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 5, pp. 4760–4773, 2019.

[6] M. Hammouda, S. Akın, A. M. Vegni, H. Haas, and J. Peissig, "Link selection in hybrid RF/VLC systems under statistical queueing constraints," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2738–2754, 2018.

[7] M. Najla, P. Mach, and Z. Becvar, "Deep learning for selection between RF and VLC bands in device-to-device communication," *IEEE Wireless Communications Letters*, vol. 9, no. 10, pp. 1763–1767, 2020.

[8] Z. M. Fadlullah, Y. Kawamoto, H. Nishiyama, N. Kato, N. Egashira, K. Yano, and T. Kumagai, "Multi-hop wireless transmission in multi-band WLAN systems: Proposal and future perspective," *IEEE Wireless Communications*, vol. 26, no. 1, pp. 108–113, 2019.

[9] F. Tang, B. Mao, Y. Kawamoto, and N. Kato, "Survey on machine learning for intelligent end-to-end communication toward 6G: From network access, routing to traffic control and streaming adaption," *IEEE Communications Surveys Tutorials*, vol. 23, no. 3, pp. 1578–1598, 2021.

[10] B. Mao, F. Tang, Y. Kawamoto, and N. Kato, "Ai models for green communications towards 6G," *IEEE Communications Surveys Tutorials*, 2021.

[11] S. Shrivastava, B. Chen, C. Chen, H. Wang, and M. Dai, "Deep Q-network learning based downlink resource allocation for hybrid RF/VLC systems," *IEEE Access*, vol. 8, pp. 149 412–149 434, 2020.

[12] F. Wu, W. Yang, M. Sun, J. Ren, and F. Lyu, "Multi-path selection and congestion control for NDN: An online learning approach," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1977–1989, 2021.

[13] W. Wu, N. Cheng, N. Zhang, P. Yang, W. Zhuang, and X. Shen, "Fast mmwave beam alignment via correlated bandit learning," *IEEE Transactions on Wireless Communications*, vol. 18, no. 12, pp. 5894–5908, 2019.

[14] S. Hashima, K. Hatano, H. Kasban, and E. Mahmoud Mohamed, "Wi-Fi assisted contextual multi-armed bandit for neighbor discovery and selection in millimeter wave device to device communications," *Sensors*, vol. 21, no. 8, p. 2835, 2021.

[15] S. Hashima, K. Hatano, H. Kasban, M. Rihan, and E. M. Mohamed, "Multiagent multi-armed bandit techniques for millimeter wave concurrent beamforming," in *2020 8th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC)*, 2020, pp. 56–59.

[16] S. Hashima, K. Hatano, and E. M. Mohamed, "Multiagent multi-armed bandit schemes for gateway selection in uav networks," in *2020 IEEE Globecom Workshops (GC Wkshps*, 2020, pp. 1–6.

[17] Q. Zhao, *Multi-Armed Bandits: Theory and Applications to Online Learning in Networks*. Morgan & Claypool, 2019.

[18] S. Kaddouri, M. E. Hajj, G. Zaharia, and G. E. Zein, "Indoor path loss measurements and modeling in an open-space office at 2.4 GHz and 5.8 GHz in the presence of people," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, 2018.

[19] E. M. Mohamed, S. Hashima, K. Hatano, S. A. Aldossari, M. Zareei, and M. Rihan, "Two-hop relay probing in WiGig device-to-device networks using sleeping contextual bandits," *IEEE Wireless Communications Letters*, vol. 10, no. 7, pp. 1581–1585, 2021.

[20] M. Boban *et al.*, "Multi-band vehicle-to-vehicle channel characterization in the presence of vehicle blockage," *IEEE Access*, vol. 7, pp. 9724–9735, 2019.