Regret-Minimizing Exploration in HetNets with mmWave

Michael Wang[1], Aveek Dutta[2], Swapna Buccapatnam[3], Mung Chiang[1]
[1]Princeton University, NJ, [2]University of Kansas, KS, [3]IBM T.J. Watson Research Center, NY
Email: {mwseven, chiangm}@princeton.edu, aveekd@ku.edu, sbuccap@us.ibm.com

***Abstract*—We model and analyze a User-Equipment (UE) based wireless network selection method where individuals act on their stochastic knowledge of the expected behavior off their available networks. In particular, we focus on networks with millimeter-wave (mmWave) radio. Modeling mmWave radio access technologies (RATs) as a stochastic 3-state process based on their physical layer characteristics in Line-of-Sight (LOS), Non-Line-of-Sight (NLOS), and Outage states, we make the realistic assumption that users have no knowledge of the statistics of the RATs and must learn these while maximizing the throughput obtained. We develop an online learning-based approach to access network selection: a user-centric Multi-Armed Bandit Problem that incorporates the cost of switching access networks. We develop an online learning policy that groups network access to minimize costs for RAT selection, analyze the regret (loss due to uncertainty) of our algorithm. We also show that our algorithm obtains optimal regret and in numerical examples achieves 24% increase in total throughput compared to existing techniques for high throughput mmWave RATs that vary over a fast timescale.**

I. Introduction

Millimeter wave (mmWave) radio is of growing interest for deployment in 5G networks [1]–[3]. With the potential for extremely high throughput compared to sub-28GHz [1], mmWave can exploit the enormous amount of spectrum available in these bands. It is likely mmWave Radio Access Technologies (RATs) will co-exist with existing technologies such as 3G, LTE, and 802.11(Wi-Fi) in a heterogeneous network (HetNet). A user in this HetNet scenario can access different RATs to download data: however, oftentimes users restrict themselves to associating with one RAT at a time. This paper examines the important question of *how should a user optimally select the best access network in a HetNet with mmWave to maximize throughput and minimize switching costs*?

We solve this problem using a user-centric approach which does not require extensive signaling and coordination among the different access networks, an unrealistic assumption if the networks are owned by different operators. In addition, the user is better positioned to monitor both the set of accessible RATs and changes in the UE's own data demands. However, there is a cost associated to switching between RATs: user-incurred RAT-specific overhead due to wireless network handoff, and such costs may add up to be substantial.

As a user switches between different RATs, the lack of perfect information available to the user results in suboptimal throughput and switching costs when maximizing throughput. We model this problem of maximizing throughput and minimizing switching costs with the model of online learning in a stochastic Markovian environment [4], where the user must learn accessible RATs' variable statistics through experimentation. Specifically, we formulate RAT selection in mmWave HetNets by extending the standard Multi-Armed Bandit Problem (MABP) to incorporate two critical elements: (1) HetNet switching overheads and (2) Markovian characteristics of mmWave RATs, thereby accounting for access to high-throughput-but-highly-variable mmWave RATs as shown in Fig. 1.

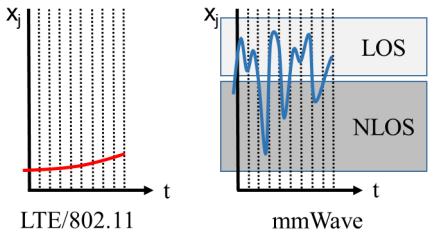


Fig. 1: Example of slower throughput fluctuations (x_j) in LTE/802.11 (left) and faster fluctuations in Millimeter Wave (right).

We present a distributed algorithm for RAT Selection in HetNets with mmWave that balances maximizing aggregate throughput and minimizing wireless switching overheads. It operates on the UE client device, using past empirically-derived throughput values from past RATs the UE has associated with. Our metric is the performance gap between our algorithm and an offline-optimal algorithm, and show that we obtain optimal $O(\ln(t))$ total regret (loss due to uncertainty and switching costs), instead of only the regret component characterizing throughput in [5]. Furthermore, our algorithm achieves better total obtained throughput in simulations compared to other candidates that do not operate under any such switching constraint. This better throughput result while under switching constraints is a direct consequence of the tighter Chernoff bound [6] used by parameters in our algorithm.

Our main contributions in this paper are as follows:

- We model the temporal behavior of mmWave RATs as a finite-state, irreducible, aperiodic Markov Chain with a general non-reversible, unknown transition matrix P. (Section II) We formulate the problem of Stochastic RAT Selection in HetNets with mmWave RATs under unknown statistics as a rested MABP (Section III-A);
- We develop an online learning policy that is distribution-independent (mmWave HetNet Selection Algorithm, The-

orem 1) and solves the RAT Selection problem while minimizing cost of switching (Section III-B) by grouping channel access;
- We show that the total regret (loss due to uncertainty and switching costs) of mHS is upper-bounded by $O(\ln(t))$ in time, optimal by [7] (Section III-C); and
- We use real mmWave characteristics in our numerical results to compare our policy with those existing in literature [5], show our solution outperforms existing policies in the total obtained throughput, and discuss the user implications of implementing this system. (Section IV)

From our results, a user using mHS in a HetNet with highly-variable-throughput mmWave RATs obtains the best of both worlds, i.e., order-optimal throughput performance and switching minimization, irrespective of whether mmWave or other traditional RATs are optimal.

II. System Model

A. Millimeter Wave Radio and HetNets

Next-generation mobile networks are expected to make extensive use of millimeter wave (mmWave) radio technology [2]. We model the 3-state mmWave RAT as a 3-state discrete-time stochastic process, illustrated in Fig. 2 with transition probabilities $\{P_{ij}\}$. The mmWave channel is composed of Line-of-Sight (LOS), non-LOS and outage states. In LOS, the mobile device has an unobstructed path for the signal to propagate [8]–potentially obtaining peak rates tens of Gbps. In non-LOS, the mmWave channel severely degrades to a much lower data rate [1], [3] due to obstructed signal paths. Outage occurs when very little signal is observed due to physical obstacles and atmospheric absorption [9], and can be treated as if no signal were received. The mmWave RAT channel is in one of these states at all times, and the non-deterministic transitions between individual states obey Markovian properties [4] when accessed by a mobile user. The channel's data rate is also non-deterministic because of fast-fading and shadowing effects, and can be modeled as a random variable dependent on the number of users and the mmWave channel state.

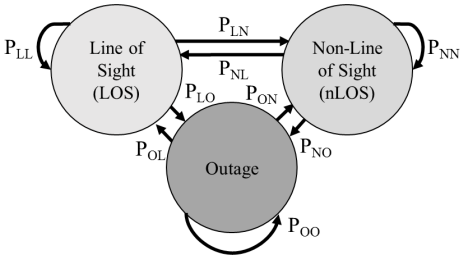


Fig. 2: 3-State mmWave Markovian Channel State Model

Thus, any mobile device that performs RAT selection must do so by predicting the state of the mmWave RAT based on the channel's stochastic parameters, which may not be known to the user. Furthermore, mmWave channels change state and channel quality on the order of a millisecond [3] or less, much faster than comparable technologies like LTE/802.11. These fast state changes demand RAT selection

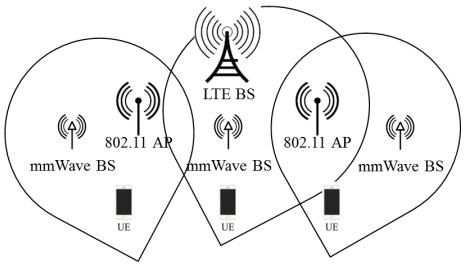


Fig. 3: Heterogeneous Network example with mmWave access: each user has access to a subset of BSs

on a timescale precluding traditional centralized, network-operated control schemes for UE-RAT association, as there is very little time to aggregate all UEs' channel state information, centrally calculate an optimal association, and return UE-RAT configurations to all UEs. These differences necessitate new UE-centric control of HetNets with mmWave.

Due to frequent transitions between mmWave RAT states, many RAT selection policies that converge to an optimal user-BS assignment encounter issues from mmWave's non-i.i.d. nature on the same timescale that LTE/802.11 may be considered independent in time. In the next section we examine a system model and Stochastic RAT selection policy that (1) makes RAT selection decisions on the same timescale as mmWave state transitions, (2) leverages properties of the mmWave transition matrix P, and (3) considers LTE/802.11 RATs consisting of a single state to be degenerate Markov chains used to describe mmWave.

We model a user's heterogeneous wireless environment to be composed of M Base Stations (BSs). Here, "BS" represents a generic Node-B in 3G, eNodeB in 4G, Access Point in 802.11, and mmWave Base Stations, and $\mathbf{M} = \{1, ..., M\}$ denotes the set of BSs. n_j denotes the number of users on BS j, and each user has access to a subset of the BSs. We assume that all RATs are interference-free due to frequency separation between different-RAT BSs, and frequency reuse among same-kind-RAT BSs. An example of this HetNet environment is shown in Fig. 3. We model beacon signals belonging to different BSs received on the same RAT by a single user as being received from multiple RATs to simplify analysis, using the term "RAT" to mean "BS" interchangeably.[1] Furthermore, due to battery considerations for mobile devices, we assume that each user is willing to use at most one RAT at any given time. In addition, we assume that the mmWave RATs are rested: the mmWave channel does not undergo transitions unless the user interacts with it. This can be assumed because the user is stationary over the timescale of the algorithm, which can be preformed on the order of sub-milliseconds, and user mobility is the primary driver of mmWave state changes.

B. Online Learning and Regret in highly variable HetNets

In the HetNets scenario, there are multiple choices of RATs (arms) for a user (player), but the user can only associate with

[1]For example, an 802.11g interface may be capable of receiving data on Channels 1, 6 and 11 in the 2.4GHz ISM band, but we consider the user has having 3 distinct RATs.

a one RAT in any given discrete time step. At each time step, the user is able to download data (obtain a reward) from its associated RAT. The obtained throughput for a user from each RAT is unknown *a priori* to that user and differs between RATs. The goal of the user is therefore to maximize the expected sum data over a period of time–in other words, how should one select a RAT to download data from by learning the statistics of each RAT from its past observations. Furthermore, this is complicated by the existence of fixed, RAT-specific overhead costs incurred from switching.

This is an example of the fundamental tradeoff between *exploration* and *exploitation*. The user needs to explore, or sample/associate with, all available RATs to discover the best option–and to prevent repeated sampling of suboptimal RATs that it wrongly believes to be best. On the other hand, the user also needs to maximize total obtained throughput, as well as to avoid excessive switches. The user's *total regret* is a concept typically used to measure how well an algorithm performs.

Regret [10] is the difference between the expected total obtained throughput from an ideal policy with perfect knowledge, and the expected total obtained throughput of the policy in question. It is a measure of the inefficiency generated by uncertainty in the network and the decision-making process. We focus on regret relative to the infeasible "ideal best single RAT" policy, which relies on perfect *a priori* knowledge of the throughput (reward) distributions of all RATs. The key challenge is to upper-bound regret as tightly as possible.

Our design goals for such a UE-centric RAT selection algorithm for HetNets with mmWave require the following:

- A UE-centric design to remove inefficient signaling and delay overhead present in network-centric control designs, such that UEs may respond more quickly to channel conditions changing on the faster mmWave timescale.
- Maximize aggregate throughput obtained in order to efficiently obtain data for UE operation.
- Minimize total switching costs from moving between different RATs (e.g., soft/hard handover, session handover).

There exists an extensive body of work on applying online learning-type algorithms to regret minimization in MABPs [5], [7], [10], [11]. However, many of these are insufficient to meet all the design goals: [7], [10] provide policies to maximize aggregate throughput but do not take into account the Markovian nature of state transitions or aggregate switching costs, essential to minimizing overhead in today's highly-loaded wireless network infrastructure. [11] minimizes switching between available options assuming i.i.d. rewards from all options, without accounting for frequent and fast-changing mmWave RATs in HetNets. The work in [5] comes closest to ours: they address reward maximization given Markovian (not i.i.d.) rewards, but any analysis of switching costs and modeling their interaction with throughput maximization is absent.

III. Online Learning for Stochastic RAT Selection

In this section, we describe the user-centric RAT selection problem with mmWave, our proposed mHS (mmWave HetNet Selection) algorithm, and show our main result of optimal $O(\ln(t))$ regret in Theorem 1.

A. Problem Formulation

We formulate the problem as a rested Multi-armed Bandit, where a UE selects a single RAT to associate with at every time slot. At each time slot, the UE obtains a throughput dependent on the distribution for that given RAT but does not know the state of the RAT. Under the rested assumption, RATs with Markovian state transitions do not evolve or change states over time when not played: the RAT evolves according to the underlying Markov law only when played.

We describe the parameterized throughput distributions for each of the M RATs by $f_1(x;\theta_1,n_1),...,f_M(x;\theta_M,n_M)$ over a measure v for the channel throughput. Both $f_j(\cdot;\cdot)$ and n_j are known from the RAT, and θ_j are unknown channel conditions belonging to some bounded parameter set Θ. Assuming that the distribution is bounded, the mean is

$$\mu(\theta_j) = \int_{-\infty}^{\infty} x f(x;\theta) d\mathrm{v}(x) \tag{1}$$

and we define the optimal mean over all j to be

$$\mu^* = \max\{\mu(\theta_1),...,\mu(\theta_M)\} = \mu(\theta^*) = \mu(\theta_{j^*}) \tag{2}$$

The total data obtained by a RAT selection policy $\alpha = \{\alpha(1),\alpha(2),...\}$ by time t, where $\alpha(t) \in \{1,...,M\} \forall t$, is

$$X(t) = \sum_{i=1}^{t} x(i) \tag{3}$$

The expected total data obtained by time t is therefore

$$E_\alpha[X(t)] = \sum_{i=1}^{t} \mu(\theta_j) E_\alpha[T^j(t)] \tag{4}$$

where $T^j(t) = \sum_{i=1}^{t} 1(\alpha(t)=j)$ is the total number of time slots that policy α has sampled the distribution $f_j(x;\theta_j,n_j)$.

We characterize R_α, total regret for the policy α, as the sum of the sampling regret R_α^T (loss due to suboptimal throughput) and the switching regret R_α^S (loss due to total switching costs):

$$R_\alpha = R_\alpha^T + R_\alpha^S \tag{5}$$

$$\begin{aligned} \text{where} \quad R_\alpha^T(t) &= t\mu^* - E_\alpha[X(t)] \\ &= \sum_{j:\mu<\mu(\theta_j)} (\mu^* - \mu(\theta_j)) E_\alpha[T^j(t)] \end{aligned} \tag{6}$$

$$\text{and} \quad R_\alpha^S(t) = \sum_{j=1}^{M} c_a^j E_\alpha[s_a^j(t)] + \sum_{j=1}^{M} c_d^j E_\alpha[s_d^j(t)] \tag{7}$$

where c_a^j, c_d^j are constant RAT-specific costs of association and dissociation and $s_a^j(t), s_d^j(t)$ are the number of times up to t the user has associated and dissociated from RAT j.

The UE wishes to maximize its total throughput less the switching regret, $E_\alpha[X(t)] - R_\alpha^S(t)$: this is equivalent to minimizing the total regret over time $R_\alpha(t)$.

Throughout the rest of the paper, we make the following assumptions:

M	Number of RATs	n_j	Number of users on RATs j
$f_j(x;\theta_j,n_j)$	Throughput reward distribution given θ_j, n_j	θ_j	Hidden (from the UE) parameters for RAT j
x_j	Throughput obtained by user on RAT j	$\mu(\theta_j)=\mu^j$	Mean value of x_j
$\alpha(t)$	RAT chosen by selection policy at time t	μ^*	Highest mean throughput over all RATs
$X(t)$	Total data by user over $t \in [1,t]$	$\bar{X}^j(t)$	Sample mean data for RAT j over $t \in [1,t]$
R_α	Total Regret	$T^j(t)$	Number of samples of RAT j by policy α over $t \in [1,t]$
R_α^T	Sampling Regret	R_α^S	Switching Regret
m_j	Peak throughput obtainable on RAT j	P_j	Probability Transition matrix for RAT j
c_a^j	Cost of Association to RAT j	c_d^j	Cost of Dissociation from RAT j
s_a^j	Number of Switches to RAT j	s_d^j	Number of Switches away from RAT j
N_{fk}	Timeslots the user can switch RATs	f	Frame number
b_f	Length of block in frame f	k_f	Number of blocks in frame f
$t(f,k)$	First time slot in frame f, block k	N_f	Last time slot at the end of frame f

TABLE I: Main Notation

1) The parameters $\{\theta_j\}$ are such that there exists $\mu(\theta_j) = \mu^j < \mu^* = \mu(\theta_{j^*})$, for all $j \neq j^*$.
2) $|x_j(t) - E[x_j(t)]| \leq m_j \ \forall j$.
3) P_j and $P_j^* P_j$ are irreducible, where $P_j^* = \text{adjoint}(P_j)$.

The first assumption implies there exists a unique "best" RAT amongst set of available RATs: it is unlikely for multiple RATs to be perfectly identical. The second implies that the deviation of the throughput values are bounded by m_j, because throughputs must be finite. Finally, the last assumption is weaker than reversibility of Markov Chains. Note that we do not require the transition matrices P_j to be reversible, since reversibility is an unrealistic assumption for all mmWave RATs.

B. Network Selection Algorithm

For this type of *exploration* vs *exploitation* problem with constant RAT-specific costs per switch, it is intuitive to see that any asymptotically efficient policy must ensure that samples from the same RAT are grouped as much as possible to minimize the number of switches between different RATs. We use a Switching Scheduler Algorithm (Alg. 1) inspired by the block switching algorithm in [11]. This algorithm outputs N_{fk}, a schedule of time slots at which a UE is allowed to switch RATs to minimize switching costs.

Algorithm 1: Switching Scheduler Algorithm

Input: Number of accessible RATs M
Initialization: Switching Schedule $N_{fk} = \{1, ..., M\}$, frame index $f = 1$, block index $k = 1$
while $f \geq 1$ **do**
  while $k \leq k_f = (N_f - N_{f-1})/b_f$ **do**
    $t(f,k) = N_{f-1} + ((k-1)b_f + 1)$;
    $N_{fk} = N_{fk} \cup t(f,k)$;
    $k = k + 1$;
  end
  $f = f + 1$;
end

The Switching Scheduler scheme for mmWave HetNets, inspired by [11], first divides discrete time into "frames" $f = 0, 1, 2, ...$ Each frame is further divided into "blocks" of equal duration b_f, numbered $k = 1, 2, ..., k_f$, labeled by (f, k). Table II shows how to determine block and frame lengths. N_f is the final time slot in frame f, and $t(f,k)$ is be the first time slot in the block (f,k). An example schedule for an $(M = 3)$-RAT system is shown in Fig. 4.

Frame (f)	b_f	$N_f - N_{f-1}$
0	1	M
1	1	$\lceil \frac{e^{1^2} - e^{0^2}}{1} \rceil \cdot M \cdot 1$
2	2	$\lceil \frac{e^{2^2} - e^{1^2}}{2} \rceil \cdot M \cdot 2$
f	f	$\lceil \frac{e^{f^2} - e^{(f-1)^2}}{f} \rceil \cdot M \cdot f$

TABLE II: Switching Scheduler Algorithm: Block/Frame Lengths

The algorithm initializes N_{fk} to be all time slots $1, ..., M$, to allow the UE to sample all M RATs. Next, in each frame, the number of blocks is determined by dividing the frame duration $N_f - N_{f-1}$ by k_f, the block duration. The intuition for Alg. 1 is to restrict switching opportunities as time goes on, because the need to refine throughput estimates of other RATs decreases with time due to averaging in the sample mean throughput in mHS (Alg. 2).

Once N_{fk} is determined *a priori*, it is fed to the online mHS Algorithm (Alg. 2), which calculates which RAT to switch to at the allowed switching opportunities $t \in N_{fk}$. The algorithm starts by sampling throughputs for all M RATs during the first M time slots to initialize sample means for each RAT. Next, at each $t \in N_{fk}$, the algorithm calculates a ranking index, index$^j(t)$, for all RATs, and switches to the RAT with the highest rank. This ranking index balances the user's need for exploitation (selecting the RAT with the largest $\bar{X}_j(t)$), and an exploration term that encourages occasional samples of suboptimal RATs to update sample mean throughputs. This term depends on m_j, the maximum throughput obtainable, L_j, a parameter depending on the state transition matrix of the UE-RAT channel, and $T^j(t)$, the number of samples of RAT j up to time t. This exploration term grows with $\sqrt{\ln(t)}$, causing the UE to associate with less-sampled RATs whenever it becomes large relative to all other ranking indexes in order to update sample throughputs.

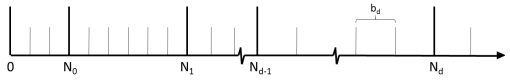


Fig. 4: Example of switching schedule for (M=3)-RAT HetNets.

Algorithm 2: mHS: mmWave HetNet Selection Algorithm

Input: Switching Schedule: $\{N_{fk}\}$, RAT-specific parameters: $\{m_j, L_j\}$

Initialization: time $t = 1$
while $t > 0$ **do**
  if $t \in N_{fk}$ **then**
    if $t \leq M$ **then**
      Play RAT t ;
    else
      Calculate for all $j \in M$:
      $\bar{X}^j(t) = \frac{x_j(1)+x_j(2)+\ldots+x_j(T^j(t))}{T^j(t)}$;
      $\text{index}^j(t) = \bar{X}^j(t) + m_j\sqrt{\frac{L_j \ln(t)}{T^j(t)}}$;
      Play the RAT with $\arg\max_j\{\text{index}^j(t)\}$
    end
  else
    Continue playing the RAT played in the previous timeslot $t-1$
  end
end

C. Upper Bound on the Total Regret

Using existing policies [5], it is possible to achieve $O(\ln(t))$ sampling regret for maximizing throughput for Markovian HetNets. However, with the addition of switching costs between RATs, these policies can incur arbitrarily high total regret due to unconstrained switching. We show that our mHS policy in Alg. 2 still has a total regret of $O(\ln(t))$ for mmWave HetNets with a $o(\ln(t))$ switching regret. The upper-bound on total regret in Theorem 1 holds for L_j sufficiently large, which depends on the smallest eigenvalue gap $\varepsilon(Q_j)$ over all RATs j in the HetNet for $Q_j = P_j^* P_j$, where P_j is the transition matrix for the jth RAT. To calculate L_j, both a limit on the absolute deviation $|x_j - E[x_j]| \leq m_j$, and the eigenvalue gap can be obtained from each accessible RAT (e.g. estimation, historical values). Note that the choice of L_j does not change the logarithmic behavior of regret: however, picking a larger L_j will result in more frequent exploratory actions, and change both the constant $O(1)$ and the coefficient of $O(\ln(t))$. By picking smaller values of L_j, the index policy emphasizes the sample mean $\bar{X}^j(t)$: in the limit where $L_j = 0$, the policy simply selects the RAT with the largest sample mean throughput at all times, resulting in very fast (but potentially suboptimal performance due to very few samples from alternate RATs); conversely, larger values of L_j emphasize the exploration term, which results in slower convergence times.

However, in our main result, we show that lower-bounding the exploration parameter L_j guarantees an upper-bound on the total regret uniformly over time, with regret coefficient dependent only on (m_j, L_j). The lower-bound of $L_j(m_j, \varepsilon_{min}) \geq \left(\frac{48+180/m_j)}{\varepsilon_{min}}\right)^2$ with constants from the Chernoff bound in [6], and smallest eigenvalue gap $\varepsilon_{min} = \min_j\{\varepsilon(P_j^* P_j)\}$ gives us our result with $C'(m_j, L_j) = \frac{4m_j^2 L_j}{(\mu^* - \mu^j)^2}$ and $C(m_j, L_j) = \frac{4m_j^2 L_j}{\mu^* - \mu^j}$.

We present our main result, which upper-bounds the total regret for the mHS algorithm in mmWave HetNets:

Theorem 1. *Under the mHS RAT selection algorithm, for all $j \in M$ such that $\mu(\theta_j) < \mu^*$,*
The expected number of samples of a suboptimal RAT j is upper-bounded by

$$i) E_\alpha[T^j(t)] \leq \quad [C'(m_j, L_j) + o(1)] \ln(t), \tag{8}$$

the expected number of switches to and from RAT j is upper-bounded by

$$ii) E_\alpha[s_d^j(t)] \leq \quad E_\alpha[s_a^j(t)] \leq o(\ln(t)), \tag{9}$$

and the total regret over all M RATs is upper-bounded by

$$iii) R_\alpha(t) \leq \quad \left[\sum_{j:\mu^j<\mu^*} C(m_j, L_j) + O(1)\right] \ln(t) \tag{10}$$

Proof (Sketch): In order to bound the total regret, first we bound the expected number of times a user samples a suboptimal RAT, $T^j(t)$, at the end of a frame d in our Switching Scheduler Algorithm. We choose a parameter L_j sufficiently large to ensure convergence of the expectation $E_\alpha[T^j(t)]$ according to the Chernoff bound in [6]. We extend this result to all time in the frame, which gives us (8). Next, we bound the number of times a user switches to and from ($s_a^j(t)$ and $s_d^j(t)$ respectively) a RAT at time $N_{d-1} < t < N_d$: taking the expectation and using the bounds on N_d in [11], we obtain the upper bound on $E_\alpha[s_a^j(t)]$ and $E_\alpha[s_d^j(t)]$ (9). Finally, we combine (8) and (9) using (5) to obtain the $O(\ln(t))$ bound (10). See Appendix A for the full proof. ■

Theorem 1 shows that, for the more general case of non-reversible Markovian behavior of throughput of a given RAT (e.g., mmWave RATs), the mHS algorithm is able to achieve the optimal $O(\ln(t))$ total regret as in the i.i.d. case with switching costs in [11].

Remark 1. *Policies like mHS that achieve $O(\ln(t))$ regret achieve the optimal time-averaged total obtained throughput when compared with an "ideal best single RAT" policy. This can be seen in the asymptotic lower bound*

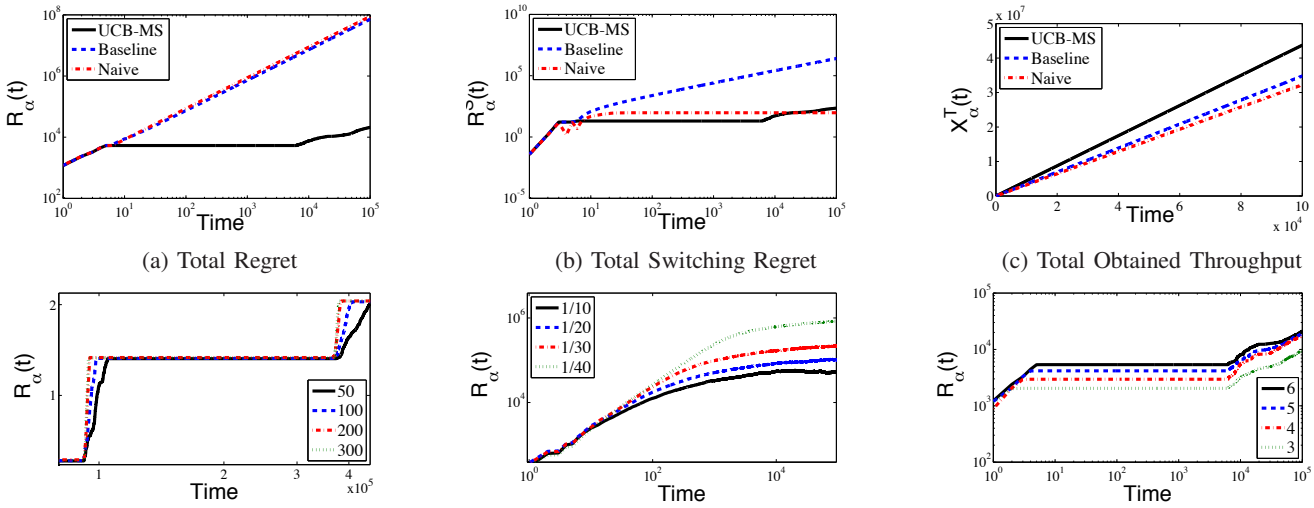
$$\frac{R_\alpha(t)}{t} \to 0 \implies \frac{E_\alpha[\textit{total reward}]}{t} \to \mu^* \tag{11}$$

In general, any policy that has sub-linear regret growth over time achieves the optimal time-averaged throughput.

It is well known that $O(\ln(t))$ sampling regret is order-optimal asymptotically for a policy that is uniformly good for any Markovian throughput (reward) distribution [7], [5]. However, mHS is order-optimal *uniformly* over time for total regret, rather than only asymptotically as in [7] and [11].

IV. Performance Evaluation

We compare the performance of mHS with the "Baseline" UCB policy in [5], a UCB-based algorithm for the Markovian MABP that does not consider switching costs, and a "Naive" policy that selects the RAT with the best last-observed throughput. Fig. 5 show simulation results for 6 RATs in a mmWave HetNet averaged over 100 runs, with parameters given in Table III and both association/dissociation switching costs of 25% of the minimum PHY throughput on each RAT. The mmWave transition matrix P is valued as in [4].



(a) Total Regret

(b) Total Switching Regret

(c) Total Obtained Throughput

(d) Total Regret with Increasing Number of Users

(e) Total Regret with Increasing mmWave Throughput

(f) Total Regret with Increasing Number of accessible RATs

Fig. 5: mHS achieves (a) lower total regret over time, (b) lower switching regret, and (c) greater total obtained throughput than Baseline and Naive policies. (d) Increasing the number of users increases regret more quickly during exploration; (e) Greater throughput gap between optimal and other RATs result in lower regret; (f) Increasing number of accessible RATs increases regret.

RAT	RAT	Rate Vector (Mbps)
1	LTE [12] (@10MHz BW)	[.153,.234,.377,.602,.877, 1.16,1.477,1.914,2.406,2.73, 3.32,3.9,4.52,5.12,5.56]
2	802.11g	[6,9,12,18,24,36,48,54]
3	802.11n	[60,120,180,240,360,480,540,600]
4	802.11g	[6,9,12,18,24,36,48,54]
5	802.11n	[60,120,180,240,360,480,540,600]
6	mmWave-NLOS	[20,40,60,80,100]
	mmWave-LOS [1]	[1000,1100,1200,1300,1400]

TABLE III: Parameters for PHY Layer Throughput Distribution

A. Simulation Results

Fig. 5a shows the total regret over time: the mHS algorithm has a slower growth of regret compared to the other policies. Fig. 5b shows that the switching costs incurred by mHS is significantly less than both comparison policies: it is only after time $t = 5350$ for these experiments that the switching regret exceeds the Naive policy, because the Naive policy eventually locks onto a single (possibly incorrect) RAT and no longer incurs switching regret. Fig. 5c completes the comparison by examining the total throughput of the three policies, and our method obtains over 24% more total obtained throughput [5].

Note that in these figures, the mHS algorithm undergoes a long period up to about $t = 5350$ on average in our simulations, where total regret is constant. The high throughput available on the mmWave RAT results in a large m_j, causing users to remain on mmWave instead of exploring other RATs until after the transient period. The flat regret curve is a consequence of the policy selecting the "optimal" (max mean throughput) RAT, and therefore incurring no regret from selecting the optimal RAT. However, these periods of nonincreasing regret are transient: by design, the exploration terms in mHS for other RATs grow until they force a period of exploration, and the UE switches to different RATs to refine its knowledge of the RATs' average throughputs.

In Fig. 5d, we plot the average total regret of users over time in a HetNet system with a variable number of other users randomly distributed on the M RATs, which move to and from RATs under a Poisson probability with mean interarrival/departure times of 10. Throughputs on each RAT are shared in a time-fair manner (see Appendix B). By increasing the number of users, regret increases more quickly over time when the UE explore alternative RATs: this is caused by users obtaining less throughput per RAT on average, so the exploration term causes users to explore more frequently to obtain better throughputs.

We simulate the performance of our algorithm when we scale the potential throughputs on the mmWave RAT by multiplying by $\{1/40, 1/30, 1/20, 1/10\}$ in Fig. 5e. By reducing peak mmWave throughputs to values comparable to other RATs, total regret incurred by mHS increases. When a user has to learn the statistics of similar-throughput-RATs, more regret is incurred as it takes longer to differentiate between them: our algorithm functions better when there is a larger gap between the available throughputs on different RATs. In Fig. 5f, the number of RATs is increased causing regret to grow over a longer time period, as well as obtain a higher total regret (indicative of the extra time to learn each RAT).

For a UE with access to highly-variable-throughput RATs, such as mmWave, in HetNets, mHS achieves an optimal trade-off between maximizing expected throughput and minimizing switching costs, to minimize the total regret for a user, as well as significantly outperforming other techniques in terms of the total data downloaded. In practical terms, for time step durations on the order of channel coherence times for mmWave (1 ms [3]), users can identify and lock onto the optimal mmWave RAT in only a few ms to minimize regret. In general, our technique is most applicable when mmWave RATs

allow extremely high throughputs, while allowing the user to occasionally explore alternatives. It obtains $O(\ln(t))$ regret even if the throughput advantage of mmWave is significantly degraded to below that of LTE/802.11n.

V. Related Work

RAT Selection in HetNets: Optimal selection of wireless access networks in HetNets has been studied in various contexts. Many approaches invoke optimal association of users to RATs [13], or use RAT-centric techniques to guide user decisions towards an optimal system state [14], [15]. Alternatively, user-centric approaches [16] seek to shape user perspectives such that the outcome of local client decisions approaches a global optimum. However, these strategies generally assume thatv channel statistics change slower than algorithm convergence. In our work, we consider both stochastic behavior for HetNet channels and optimality from the user perspective in terms of the average regret observed during RAT selection.

Millimeter Wave RATs: Studies in mmWave frequencies [1], [17], [18] show its potential as a possible high-throughput access network. However, there is a need for network selection algorithms that are able to function with its dynamic channel statistics, which may change during a single session. Specifically, mmWave RATs have been shown [19] to exhibit highly variable SNR and bitrates, which motivates the need for adaptive network selection algorithms that incorporate this behavior. Authors in [4] have modeled the mmWave RAT as a Markov chain with pre-determined PHY-layer rates and perform handovers assuming steady-state channel conditions and known state transition probabilities. In our work, we do not assume perfect knowledge of transition probabilities, allow for parameterized distributions of rates in each state, and do not assume steady-state channel conditions.

Multi-Armed Bandits: For the classical multi-armed bandit problem with $i.i.d$ rewards for the arms/actions, the seminal work of [7] shows that the asymptotic lower bound on the regret of any uniformly-good policy scales logarithmically in time ($\mathrm{O}(\ln(t))$). The authors in [11] provide a distribution-dependent block-based online policy to minimize switching costs for the MABP with stationary and $i.i.d$ reward distributions. The authors in [5] propose a distribution independent index based policy (similar to the UCB policy in [10]) for the MABP with Markovian rewards, which can incur a high switching cost. In our work, we consider the more general case of Markovian rewards under irreversibility characteristic of mmWave and provide a simpler online learning policy, which is distribution-independent and incurs only logarithmic regret with sub-logarithmic switching regret.

VI. Conclusion

In this work, we have addressed the problem of stochastic RAT selection in HetNets with mmWave RATs by modeling it as a UE-centric multi-armed bandit problem. In our model, mmWave RATs are realistically modeled as irreversible Markov chains and we also take into account the overhead incurred upon frequent switching between RATs. We developed the mHS policy and proved that it achieves the optimal $\mathrm{O}(\ln t)$ regret with a sub-logarithmic switching regret over time. Finally, we evaluated the performance of our policy through simulations and showed that it not only achieves a significantly better total regret and throughput performance, but also incurs very low switching overhead compared to existing baseline policies developed for MABP with Markovian rewards.

VII. Acknowledgements

This research was supported by a gift from Intel and by NSF grant CNS-1527513. The authors would like to thank Nageen Himayat, Sarabjot Singh, Shu-Ping Yeh, Shilpa Talwar and David Ott at Intel Labs for helpful discussions.

References

[1] M. R. Akdeniz, Y. Liu, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *CoRR*, vol. abs/1312.4921, 2013.

[2] F. Boccardi, R. Heath, A. Lozano, T. Marzetta, and P. Popovski, "Five disruptive technology directions for 5g," *Communications Magazine, IEEE*, vol. 52, pp. 74–80, February 2014.

[3] S. Rangan, T. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proceedings of the IEEE*, vol. 102, pp. 366–385, March 2014.

[4] M. Mezzavilla, A. Dhananjay, S. Panwar, S. Rangan, and M. Zorzi, "An mdp model for optimal handover decisions in mmwave cellular networks," *arXiv preprint arXiv:1507.00387*, 2015.

[5] C. Tekin and M. Liu, "Online algorithms for the multi-armed bandit problem with markovian rewards," in *Proc. of Allerton Conference on Communications, Control, and Computing*, 2010.

[6] P. Lezaud, "Chernoff-type bound for finite markov chains," *The Annals of Applied Probability*, vol. 8, no. 3, pp. pp. 849–867, 1998.

[7] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, pp. 4–22, Mar. 1985.

[8] K.-C. Huang and Z. Wang, *Millimeter wave communication systems*, vol. 29. John Wiley & Sons, 2011.

[9] FCC, "Millimeter wave propagation: Spectrum management implications," 1997.

[10] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, pp. 235–256, May 2002.

[11] R. Agrawal, M. Hedge, and D. Teneketzis, "Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost," *Automatic Control, IEEE Transactions on*, vol. 33, pp. 899–906, Oct 1988.

[12] 3GPP, "36.213: Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures."

[13] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User Association for Load Balancing in Heterogeneous Cellular Networks," *CoRR*, vol. abs/1205.2833, 2012.

[14] P. Coucheney, C. Touati, and B. Gaujal, "Fair and Efficient User-Network Association Algorithm for Multi-Technology Wireless Networks," in *INFOCOM 2009, IEEE*, pp. 2811–2815, Apr. 2009.

[15] S. Deb, K. Nagaraj, and V. Srinivasan, "MOTA: Engineering an Operator Agnostic Mobile Service," in *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*, MobiCom '11, (New York, NY, USA), pp. 133–144, ACM, 2011.

[16] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT selection games in HetNets," in *INFOCOM, 2013 Proceedings IEEE*, pp. 998–1006, Apr. 2013.

[17] S. Sun and T. Rappaport, "Wideband mmwave channels: Implications for design and implementation of adaptive beam antennas," in *Microwave Symposium (IMS), 2014 IEEE MTT-S International*, pp. 1–4, June 2014.

[18] T. Rappaport, F. Gutierrez, E. Ben-Dor, J. Murdock, Y. Qiao, and J. Tamir, "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," *Antennas and Propagation, IEEE Transactions on*, vol. 61, pp. 1850–1859, April 2013.

[19] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *CoRR*, vol. abs/1305.5216, 2013.

APPENDIX

A. Proof of Theorem 1

Proof of 1.i: First, we prove that for a time index at the end of a frame $d > 0$, $t = N_d$, Theorem 1.i holds:

$$E_\alpha[T^j(N_d)] \leq \left(\frac{4m_j^2 L_j}{(\mu^* - \mu^j)^2} + o(1) \right) \ln N_d \tag{12}$$

and then extend the result for all t.

Let $\bar{X}^j(T^j(t))$ denote the sample mean of the reward collected from arm j over the first t time indexes, and let $c_{t,s} = m_j \sqrt{L_j \ln(t)/s}$ and integer $l > 0$.

$$\begin{aligned}
T^j(t) &= \sum_{i=1}^{n} 1(\alpha(i) = j) = \sum_{f=1}^{d} b_f \sum_{k=1}^{k_f} 1(\alpha(t(f,k)) = j) \\
&\leq l + \sum_{f=1}^{d} b_f \sum_{k=1}^{k_f} 1(\phi^j(t(f,k), l), T^j(t(f,k)-1) \geq l) \\
&\leq l + \sum_{f=1}^{d} b_f \sum_{k=1}^{k_f} 1(\min_{0<s<t(f,k)} (\bar{X}^*(s) + c_{t(f,k)-1,s}) \\
&\qquad \leq \max_{l<s_j<t(f,k)} (\bar{X}^j(s_j) + c_{t(f,k)-1,s_j})) \\
&\leq l + \sum_{f=1}^{\infty} b_f \sum_{k=1}^{k_f} \sum_{s=1}^{t(f,k)-1} \sum_{s_j=l}^{t(f,k)-1} 1(\bar{X}^*(s) + c_{t(f,k),s} \leq \\
&\qquad \bar{X}^j(s_j) + c_{t(f,k),s_j})
\end{aligned} \tag{13}$$

with indicator $1(.)$ and $\phi^j(t(f,k), l)$ is the event when

$$\begin{aligned}
\bar{X}^*_{T^*(t(f,k)-1)} + c_{t(f,k)-1,T^*(t(f,k)-1)} \\
\leq \bar{X}^j_{T^j(t(f,k)-1)} + c_{t(f,k)-1,T^j(t(f,k)-1)}
\end{aligned}$$

Next, if $\bar{X}^*(s) + c_{t(f,k),s} \leq \bar{X}^j(s_j) + c_{t(f,k),s_j}$ holds, then one of the following must hold:

$$\bar{X}^*(s) \leq \mu^* - c_{t(f,k),s} \tag{14}$$

$$\bar{X}^j(s_j) \geq \mu^j + c_{t(f,k),s_j} \tag{15}$$

$$\mu^* < \mu^j + 2c_{t(f,k),s_j} \tag{16}$$

By choosing $s_j \geq \frac{4m_j^2 L_j \ln(t)}{(\mu^* - \mu^j)^2}$, we have that (16) is false. Choosing $l = \left\lceil \frac{4m_j^2 L_j \ln(t)}{(\mu^* - \mu^j)^2} \right\rceil$ and taking expectation for (13):

$$\begin{aligned}
E_\alpha[T^j(t)] &\leq \left(\frac{4m_j^2 L_j \ln(t)}{(\mu^* - \mu^j)^2} + 1 \right) \\
&+ \sum_{f=1}^{\infty} b_f \sum_{k=1}^{k_f} \sum_{s=1}^{t(f,k)-1} \sum_{s_j=l}^{t(f,k)-1} P(\bar{X}^j(s_j) \geq \mu^j + c_{t(f,k),s_j}) \\
&+ \sum_{f=1}^{\infty} b_f \sum_{k=1}^{k_f} \sum_{s=1}^{t(f,k)-1} \sum_{s_j=l}^{t(f,k)-1} P(\bar{X}^*(s) \leq \mu^* - c_{t(f,k),s})
\end{aligned} \tag{17}$$

For the second term, the probability in the summation is bounded by

$$\begin{aligned}
&P(\bar{X}^j(s_j) \geq \mu^j + c_{t(f,k),s_j}) \\
&= P(\frac{1}{s_j}(\sum_{i=1}^{s_j} X^j(i) - \sum_{y \in S^j} \pi_y^j \mu_y^j) \geq c_{t,s_j}) \\
&\leq Z_q \exp\{-\frac{s_j \varepsilon(Q) c_{t,s_j}^2}{8b^2(1 + h(5c_{t,s_j}/b^2)}\} \qquad (18) \\
&\leq \frac{1}{\pi_{min}} \exp\{-\frac{\sqrt{L_j}\varepsilon(Q)\sqrt{\ln(t)}}{8 + \frac{30}{m_j}}\} \qquad (19) \\
&\leq \frac{1}{\pi_{min}} t^{-\frac{\varepsilon_{min}\sqrt{L_j}}{16 + \frac{60}{m_j}}}
\end{aligned}$$

where (18) follows from Theorem 3.3 of [6], which provides a Chernoff bound independent of transition matrix reversibility, and in (19), we bound $Z_{q^j} = ||\frac{q^j}{\pi}||_2$ as

$$Z_{q^j} = \left| \left(\frac{q_y^j}{\pi_y^j} \right), y \in S^j \right|_2 \leq \sum_{y \in S^j} \left| \frac{q_y^j}{\pi_y^j} \right|_2 \leq \frac{1}{\pi_{min}}$$

and (19) holds if $L_j > 1$ as $\ln(t) > 1, \forall t > M$. Next, the third term can be bounded same as the second term. Now, (17) can be written as:

$$\begin{aligned}
E_\alpha[T^j(t)] &\leq \frac{4m_j^2 L_j \ln(t)}{(\mu^* - \mu^j)^2} + 1 \\
&\quad + 2\sum_{f=1}^{\infty} b_f \sum_{k=1}^{k_f} \sum_{s=1}^{t(f,k)-1} \sum_{s_j=1}^{t(f,k)-1} \frac{1}{\pi_{min}} t^{-\frac{\varepsilon_{min}\sqrt{L_j}}{16+60/m_j}} \\
&\leq \frac{4m_j^2 L_j \ln(t)}{(\mu^* - \mu^j)^2} + 1 + \frac{2}{\pi_{min}} \sum_{f=1}^{\infty} b_f \sum_{k=1}^{k_f} t^{-\frac{\varepsilon_{min}\sqrt{L_j}}{16+60/m_j} - 2}
\end{aligned} \tag{20}$$

where (20) bounds the double summation over s, s_j with t^2. From the mHS algorithm, we use $b_f = f$, and $k_f \leq (e^{f^2} - e^{(f-1)^2} + f)M/f$:

$$\begin{aligned}
E_\alpha[T^j(t)] &\leq \frac{4m_j^2 L_j \ln(t)}{(\mu^* - \mu^j)^2} + 1 \\
&+ \frac{2M}{\pi_{min}} \sum_{f=1}^{\infty} (e^{f^2} - e^{(f-1)^2} + f) t^{-\frac{\varepsilon_{min}\sqrt{L_j}}{16+60/m_j} - 2}
\end{aligned}$$

Next, rewriting $t = N_d$ in terms of frames f:

$$N_d = M + \sum_{f=1}^{d} (N_f - N_{f-1}) \leq M(e^{d^2} + f^2/2 - f/2) \tag{21}$$

Using (21), we can upper bound $E_\alpha[T^j(t)]$ as:

$$\begin{aligned}
E_\alpha[T^j(t)] &\leq \frac{4m_j^2 L_j \ln(t)}{(\mu^* - \mu^j)^2} + 1 \\
&\quad + \frac{2M}{\pi_{min}} \sum_{f=1}^{\infty} (M(e^{f^2} + f^2/2 - f/2))^{-(\frac{\varepsilon_{min}\sqrt{L_j}}{16+60/m_j} - 3)} \\
&\leq \frac{4m_j^2 L_j \ln(t)}{(\mu^* - \mu^j)^2} + 1 + \frac{2M^{(\frac{\varepsilon_{min}\sqrt{L_j}}{16+60/m_j} - 2)}}{\pi_{min}} \sum_{f=1}^{\infty} e^{-f(\frac{\varepsilon_{min}\sqrt{L_j}}{16+60/m_j} - 3)}
\end{aligned}$$

which converges if L_j is chosen such that $L_j \geq \left(\frac{3(16+60/m_j)}{\varepsilon_{min}}\right)^2$. Thus,

$$E_\alpha[T^j(t)] \leq \frac{4m_j^2 L_j \ln(t)}{(\mu^* - \mu^j)^2} + 1 + \text{DB; where}$$

$$D = \frac{2}{\pi_{min}} M^{(\frac{\varepsilon_{min}\sqrt{L_j}}{16+60/m_j}-2)}, \quad B = \sum_{f=1}^{\infty} e^{-f(\frac{\varepsilon_{min}\sqrt{L_j}}{16+60/m_j}-3)}$$

Extending this result for general t. Let $d : N_{d-1} < t \leq N_d$. Then,

$$\begin{aligned}\frac{E_\alpha[T^j(t)]}{\ln(t)} \leq & \frac{E_\alpha[T^j(N_d)]}{\ln N_{d-1}} \leq \frac{4m_j^2 L_j}{(\mu^*-\mu^j)^2}\frac{\ln N_d}{\ln N_{d-1}} + \frac{DB}{\ln N_{d-1}} \\ \leq & \frac{4m_j^2 L_j}{(\mu^*-\mu^j)^2} + o(1)\end{aligned}$$

which completes our proof of Theorem 1.i. *Proof of 1.ii*: Next, we show that the switching regret is bounded by o($\ln(t)$):

$$\begin{aligned}R_\alpha^S(t) &= \sum_{j=1}^{t} c_a(j)E_\alpha[S_a^j(t)] + \sum_{j=1}^{t} c_d(j)E_\alpha[S_d^j(t)], \\ \text{where } S_a^j(t) &= \sum_{i=2}^{t} 1(\alpha(t)=j, \alpha(t-1)\neq j); \\ S_d^j(t) &= \sum_{i=2}^{t} 1(\alpha(t)\neq j, \alpha(t-1)=j)\end{aligned}$$

with constant arrival and departure costs $c_a(j), c_d(j)$ for each RAT j. Letting $N_{d-1} < t \leq N_d$,

$$\begin{aligned}S_a^j(t) \leq & 1 + \textstyle\sum_{f=1}^{d} \frac{T^j(N_f) - T^j(N_{f-1})}{f} \\ = & \frac{T^j(N_d)}{d} + \textstyle\sum_{f=1}^{d-1} T^j(N_{f-1})(\frac{1}{f} - \frac{1}{f+1}) \\ \leq & \frac{T^j(N_d)}{d} + \textstyle\sum_{f=1}^{d-1} T^j(N_f)/f^2\end{aligned}$$

Taking the expectation,

$$E_\alpha[S_a^j(t)] \leq \frac{E_\alpha[T^j(N_d)]}{d} + \sum_{f=1}^{d-1} E_\alpha[T^j(N_f)]/f^2 \qquad (22)$$

From (8) and (4.2, 4.3) of [11], which upper- and lower-bounds N_d in terms of M and d,

$$\begin{aligned}E_\alpha[T^j(N_d)] & \leq \frac{4m_j^2 L_j \ln N_d}{(\mu^*-\mu^j)^2} + o(1)\ln(N_d) \\ & = \left[\frac{4m_j^2 L_j}{(\mu^*-\mu^j)^2} + \zeta\right]\ln(Me^{d^2} + Md^2) \\ & \leq K(\zeta)d^2 \text{for some } K(\zeta)\end{aligned}$$

For some time $t > N_{f_0}$

$$\begin{aligned}E_\alpha[s_a^j(t)] \leq & \frac{E_\alpha[T^j(N_d)]}{d} + \textstyle\sum_{f=1}^{d-1} \frac{E_\alpha[T^j(N_f)]}{f^2} \\ \leq & K(\zeta)d + \textstyle\sum_{f=f_0}^{d-1} K(\zeta) + \sum_{f=1}^{f_0-1} \frac{E_\alpha[T^j(N_f)]}{f^2} \\ \leq & 2K(\zeta)d + M(\zeta)\end{aligned}$$

where $M(\zeta) = \sum_{f=1}^{f_0-1} \frac{E_\alpha[T^j(N_f)]}{f^2}$. Comparing with $\ln(t)$,

$$\frac{E_\alpha[s_a^j(t)]}{\ln(t)} \leq \frac{E_\alpha[s_a^j(t)]}{\ln N_{d-1}} \leq \frac{2K(\zeta)d + M(\zeta)}{(d-1)^2} = o(1)$$

which concludes the proof of $E_\alpha[s_a^j(t)] = o(1)\ln(t)$.

For $E_\alpha[s_d^j(t)]$, recognize that there is a symmetry in associating with and dissociating from RATs: in order to switch to a new RAT at time t, a user must depart the previous RAT at the same time. This means that $s_d^j(t) \leq s_a^j(t)$ must hold true, with $s_d^j(t) = s_a^j(t)$ if $\alpha(t) \neq j$ and $s_d^j(t) = s_a^j(t) - 1$ if $\alpha(t) = j$. This means that $E_\alpha[s_d^j(t)] \leq E_\alpha[s_a^j(t)] = o(1)\ln(t)$, completing the proof of Theorem 1.ii.

Proof of 1.iii: From (8),

$$R_\alpha^T(t) = \textstyle\sum_{j:\mu^j<\mu^*} \left[\frac{4m_j^2 L_j}{(\mu^*-\mu^j)} + o(1)\right]\ln(t) \qquad (23)$$

Similarly,

$$\begin{aligned}R_\alpha^S(t) \leq & \textstyle\sum_{j=1}^{M} c_\Sigma^j E_\alpha[s_a^j(t)] \\ = & \textstyle\sum_{j\neq j^*}(c_\Sigma^j)E_\alpha[s_a^j(t)] + E_\alpha[s_a^*(t)]c_\Sigma^* \\ \leq & \textstyle\sum_{j\neq j^*}(c_\Sigma^j)E_\alpha[s_a^j(t)] + \sum_{j\neq j^*}(c_\Sigma^j)(E_\alpha[s_a^j(t)]+1) \\ \leq & \textstyle\sum_{j\neq j^*}(c_\Sigma^j + c_\Sigma^*)E_\alpha[s_a^j(t)] + (M-1)c_\Sigma^* = o(\ln(t))\end{aligned}$$

where $c_\Sigma^j = c_a^j + c_d^j$. Combining the two for the total regret,

$$\begin{aligned}R_\alpha(t) &= R_\alpha^T(t) + R_\alpha^S(t) \\ &\leq \sum_{j:\mu^u<\mu^*}(\mu^* - \mu^j)E_\alpha[T^j(t)] \\ &\quad + \sum_{j\neq j^*}\left[E_\alpha[s_a^j(t)](c_\Sigma^j + c_\Sigma^*)\right] + (M-1)c_\Sigma^* \\ &\leq \sum_{j:\mu^j<\mu^*}\frac{4m_j^2 L_j \ln(t)}{(\mu^*-\mu^j)} + \sum_{j:\mu^j<\mu^*}(\mu^*-\mu^j)(1+DB) \\ &\quad + \sum_{j\neq j^*}\left[E_\alpha[s_a^j(t)](c_\Sigma^j + c_\Sigma^*)\right] + (M-1)c_\Sigma^*\end{aligned}$$

Hence, Theorem 1.iii follows. ■

B. Throughput Model for Simulations

In our simulations, we model the downlink MAC-layer throughput of a user on BS j, denoted x_j, as a function of n_j, and the instantaneous physical (PHY) layer rate $R_j(t)$ of a user on BS j, where $x_j = R_j/n_j$. Examples of RATs with this throughput-sharing function are Time-Fair TDMA-type MACs like single-user MIMO mmWave deployments, and Proportional-Fair schedulers (PFS) like those used in interference-limited OFDMA/LTE systems and 3G. For LTE and 802.11 networks, we assume time invariant channel conditions; for mmWave networks, we consider a Markovian model described in Section II-A on the sub-second timescales of RAT selection. The throughput achieved by each user is dependent upon the instantaneous physical (PHY) layer rate $R_i(t)$ of user i on BS j, and depends on the modulation and coding scheme (MCS) and instantaneous channel conditions at time t, and we ignore the impact of mobility on instantaneous throughput.