A Markovian Model for Satellite Integrated Cognitive and D2D HetNets

S. Sinem Kafıloğlu[a,*], Gürkan Gür[b], Fatih Alagöz[a]

[a] *Department of Computer Engineering, Bogazici University, Istanbul, Turkey*
[b] *Zurich University of Applied Sciences (ZHAW), Winterthur, Switzerland*

Abstract

Next-generation wireless systems are expected to provide bandwidth-hungry services in a cost-efficient and ubiquitous manner. D2D communications, spectrum sharing and heterogeneous network architectures (HetNets) are touted as crucial enablers to attain these goals. Moreover, the shifting characteristics of network traffic towards content consumption necessitate content-centric architectures and protocols. In this work, we propose a comprehensive analytical model for a content-oriented heterogeneous wireless network with cognitive capability. We model our HetNet architecture with a Continuous Time Markov Chain (CTMC) and characterize the trade-off between energy efficiency and system goodput. We elaborate on novel elements in our model, namely the integration of *universal source* concept (modeling the content retrieval operation from external networks), caching and overlaying in D2D mode. Besides, our investigation on network mode selection provides further insight on how resource allocation and performance are intertwined.

Keywords: HetNets, D2D communications, content-centric networks, cognitive radio, resource allocation, energy consumption

1. Introduction

Future Internet is envisaged to serve multimedia heavy traffic with x1000 capacity, much lower delays and massive connectivity compared to current wireless systems [1]. Moreover, the global network traffic is shifting to a content consumption driven mode leading to proposals of various content-oriented architectures and protocols for next-generation wireless networks [2]. At the same time, a drastic reduction of energy consumption per transmitted bit in communication networks is pursued for cost efficiency and minimal environmental impact [3]. A complex amalgamation of technical paradigms including Device-to-Device (D2D) communications, spectrum sharing, caching and heterogeneous network architectures (HetNets) are expected to be instrumental to address these challenges.

Although there is a wide range of works in the literature considering each of these topics, the intersection of satellite and cellular networks with D2D and cognitive extension is yet to be explored in detail. In that regard, we focus on a specific heterogeneous network in this work — a hybrid satellite-terrestrial network with D2D and cognitive communications having content consumption as the main usage mode. Such systems are going to be crucial for Future Internet ecosystem such as 5G networks. Particularly, hybrid satellite networks are regarded as an efficient and cost-effective facilitator for fulfilling 5G requirements in content-centric[1] operation mode [2, 4]. In the same vein, D2D communications and spectrum sharing via cognitive radios (CRs) are key technologies to improve resource efficiency and alleviate the emerging capacity crunch. In our previous works [5, 6], energy efficiency (EE) is inspected for the resource allocation (RA) for content delivery in a hybrid satellite-terrestrial network. In this work, we integrate universal source concept, model in-network caching mechanism and enable overlaid D2D operation in a content-centric HetNet. We aim to alleviate the incomplete treatment of hybrid satellite-terrestrial networks entailing overlaid D2D communication and in-network caching. With a more realistic network model, we perform rigorous analysis and investigate the impact of mode selection as a crucial question of RA.

In our Continuous Time Markov Chains(CTMC) model we have hybrid users (HUs) that are in primary mode at the satellite link. HUs use dynamic spectrum access mechanism at the terrestrial link and can get service from the BS (BS mode) or operate in D2D mode. In the following sections, we first construct our network model starting with caching followed by the resource allocation component. Accordingly, the CTMC state transitions are determined. Then, we define the performance metrics according to the system parameters and carry out numerical exper-

*Corresponding author
Email addresses: sinem.kafiloglu@boun.edu.tr (S. Sinem Kafıloğlu), gueu@zhaw.ch (Gürkan Gür), fatih.alagoz@boun.edu.tr (Fatih Alagöz)

[1] "Content-centric" refers to a network architecture where content determines the protocols and operation in a content dissemination and retrieval context (content/service resolution.) rather than point-to-point communication and exchange paradigm (conventional IP networks with machine resolution).

iments for performance investigation. The key contributions of our work can be summarized as follows:

- We propose a detailed analytical model for a heterogeneous network with D2D, cognitive communications and content-oriented operation.
- In resource allocation, overlaying in D2D is enabled which allows the usage of the same frequency by different services with controlled mutual interference. Overlaying boosts network capacity and improves EE. The consequent gains are investigated.
- For the sake of a more complete model, we tackle the case where some contents cannot be available in the caches of system units in our zone of interest. Such contents are to be retrieved from external content repositories. Hence, our model is extended with the *universal source* construct for a more realistic analysis.
- We extend our analytical model for caching and integrate a baseline popularity - driven caching policy to have a more complete system representation.

2. Content-Centric Networking and Caching in Integrated Satellite-Terrestrial HetNets

The major difference of content-centric networking in contrast to the classical IP-based Internet architecture is that it enables network nodes to communicate based on content resolution rather than machine/host resolution. Based on this paradigm, various content-centric networking architectures such as Data Oriented Network Architecture (DONA) and Content-Centric Networking (CCN) have been proposed to realize Future Internet concept [7]. In these systems, the loose coupling between content and its originator provides opportunities to facilitate mechanisms for many of the prevalent issues with the current network architecture such as multicast, multipath routing, and mobility [8]. Caching is a fundamental enabler function in these networks. The utilization of extensive in-network caching reduces delay and network traffic while improving content-based services. For instance, in mobile environments, "cache till you can transfer" leads to higher efficiency and availability for delay tolerant content services [2].

The advantages of content-centric design anticipated for Future Internet are also valid for 5G satellite-terrestrial HetNets considering the traffic characteristics driven by multimedia-heavy services. In content-centric networks, deployment of in-network caching at different points in a network is essential. However, conventional caching solutions are inadequate for hybrid satellite-terrestrial networks [9]. One key challenge is the cost of such infrastructure. Nonetheless, the development of more advanced terrestrial elements, especially user devices, and the "democratization" of satellite construction, deployment and

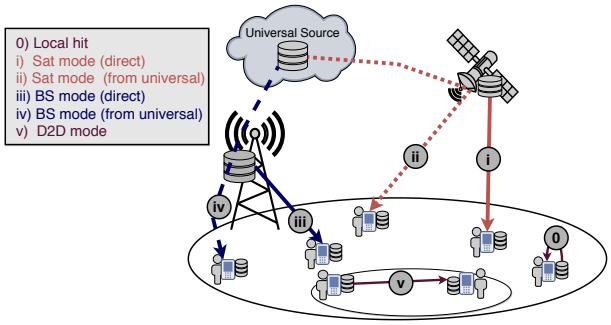


Figure 1: System model.

launch services are envisaged to drastically improve cost effectiveness. New hardware with more capable architectures, complex microelectromechanical systems and signal processing/PHY algorithms for adaptive operation are also opening new possibilities for hybrid satellite networks and content-driven services.

For addressing the peculiarities of 5G hybrid satellite networks where caching resides even on satellites, a holistic caching approach integrating terrestrial elements is crucial. The cache placement and management algorithms should consider the broadcast nature, single-hop access and large visibility of users in satellite networks (i.e. satellite-aware and adaptive cache management). Besides, a cache management framework should extensively exploit empirical information such as interest locality which refers to the correlation of requests and proximity of requesters.

3. System Model

Our modeled network is a content-centric HetNet with D2D and cognitive communications as shown in Fig. 1. We have network users in our model that can operate in both of two distinct frequency ranges as in [10, 11]: (i) satellite (ii) terrestrial. We assume that there exists a low-earth orbit (LEO) space network serving the users in addition to a terrestrial counterpart and we focus on the coverage area of one BS embedded within one LEO satellite's coverage.

We have hybrid users (HUs) that can fetch content (unless it finds in its local cache) from *i)* the satellite, *ii)* BS, or *iii)* some HU device. The HUs are native users of the satellite link. Hence, their satellite link access are in primary user (PU) mode. Solar-powered satellite is promising for alleviating energy consumption. However, the satellite bands typically have more challenging channel conditions compared to the terrestrial bands [5]. As a remedy to relatively low communication capacity of the satellite link, those users additionally utilize the terrestrial bands. In our construct, we assume that the terrestrial frequencies are already allocated for commercial use to some other legitimate users (i.e., PUs). Hence, our HU devices can only access the terrestrial bands opportunistically as SUs (cognitive mode) for capacity expansion. This multi-mode nature of our users builds on the rationale of utilizing energy efficient nature of the satellite while improving the

network capacity with more degrees of freedom via cognitive operation. Furthermore, content retrievals from the satellite or the BS can be *direct* or *indirect*. The direct retrievals occur from the satellite or BS cache to the requester HU device (req-HU). The indirect retrievals occur first from the universal source to the satellite or the BS cache and then from there to the req-HU.

In our model, we focus on the edge segment of a heterogeneous wireless network. During content consumption, a content not present in the local caches is supposed to be fetched from external network elements and servers located in the Internet. This phenomenon is very important for accurate content fetch modeling in the overall system and has an impact on how EE and throughputs materialize during experiments. Therefore, we rely on the "universal source" concept which is a logical shorthand representation for content stores/servers in the rest of Internet outside of our network-in-focus. The mode selection mechanism for content retrieval is discussed in Section 5.3. The analysis parameters are listed in Table 1.

In the following sections, we begin the construction of our analytical model with the caching aspect. Subsequently, Markov modeling of RA is done where the CTMC technique is used (Sec. 5). For that step, we first define our state space and then develop state transitions in separate parts describing PU transitions, D2D operation mode and HU transitions. We specifically look at PU transition as our users operate as SUs at the terrestrial link. Accordingly, we define PU arrival and departure transition rates (Sec. 5.1). Next, overlaying is considered in the D2D operation mode and we calculate content availability probability for D2D operations in overlaying regarding a controlled mutual interference regime (Sec. 5.2). Finally, we focus on HU transitions (Sec. 5.3).

4. Cache Model: Popularity-Driven Caching

Content consumption (e.g. video services) is the key use case in our system as observed with network traffic trends and envisaged future network characteristics [2]. Thereof, we analytically investigate content-oriented operation in our system. In such HetNet architectures, pervasive caching is a promising approach to tackle performance and cost challenges [12]. Motivated with this, we integrate a caching scheme into our investigated network architecture for content-centric operation. Caching relies on the rationale of exploitation of content access characteristics to reduce access cost (e.g., energy, bandwidth). Thus, efficient caching management alleviates resource requirements (e.g. bandwidth and server load) while improving QoS in a network [2]. Naturally, popularity-aware or -driven caching policies constitute the key caching approach for content-centric networks (e.g. see [13, 14]). Thus, they provide the baseline for constructing a comprehensive analytical model for a content-oriented HetNet. Accordingly, we make use of a popularity-driven caching (PDC) policy, and model and incorporate it into our network. From the

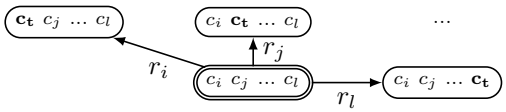


Figure 2: Cache update of a content-retrieval unit.

modeling perspective, the advantage of the PDC policy is the intuitive integration into our analytical Markov model.

In our setting, contents C=$\{c_1, c_2, ... c_N\}$ are chunks to be consumed by users. The content size distribution is exponential with mean $s(\hat{v_b})$=25 Mbits [5]. The request probability $p_{c_i}(s)$ for each content c_i is assigned based on the Zipf distribution with parameter s. The Poisson arrival processes are commonly used for multimedia traffic modeling [15, 16]. Thus, we take the content request rate of HUs as a Poisson process with mean λ_{HU}. Each content c_i has a request rate $\lambda_{HU}^{c_i}$=$p_{c_i}(s)\lambda_{HU}$ proportional to its popularity distribution. We utilize this request model and develop a PDC strategy that tries to keep popular contents in system unit caches with a higher probability. The pseudocode of the PDC algorithm used for HU device cache is available in [17]. Note that [17] is a preprint of this paper that serves to provide some technical details and derivations elaborately for the sake of completeness of analytical treatment. In the RA phase, we make use of content availability probabilities at system units. Thereof, we derive these probabilities at local HU device, satellite and BS caches in this section. The content availability for D2D operation is investigated in Section 5.2.

We construct the Markov chain for tracking content-retrieval unit (satellite, BS, or some HU device) states. In PDC policy, more popular contents are less likely to be preempted. For illustrating the cache update of a content-retrieval unit, an example cache is shown in Fig. 2. With probability $p_{c_t}(s)$ (if c_t is a popular content, $p_{c_t}(s)$ has higher value), c_t will be cached. If c_t cannot fit in the cache due to exceeded capacity, one of the c_i, c_j, ..., c_l is replaced considering their popularities. For instance, the least popular c_i is preempted for the sake of new comer c_t with the greatest probability, i.e. with the highest rate r_i:=$p_{c_t}(s)(\frac{x_i}{\sum_{\theta\in\{i,j,...,l\}} x_\theta})$ among all r_α's, $\alpha\in\{i,j,...,l\}$. r_i and x_γ are given as follows:

$$r_i := p_{c_t}(s)\left(\frac{x_i}{\sum_{\theta\in\{i,j,...,l\}} x_\theta}\right) \quad (1)$$

$$x_\gamma := \left[\prod_{x\in\{i,j,...,l\}} p_{c_x}(s)\right]/[p_{c_\gamma}(s)] \text{ where } \gamma\in\{i,j,...,l\} \quad (2)$$

The availability probabilities $p_{c_i}^{lo}$, $p_{c_i}^{BS}$ and $p_{c_i}^{sat}$ for any content c_i (in Table 1) is calculated with the help of Markov chains as further explained in [17].

5. Markovian Model of Resource Allocation

For the resource allocation (RA) problem in our HetNet, we perform a rigorous analysis on the channel usage of HUs for content retrievals. In that regard, our key assumptions are as follows: We do not have control over baseline users. Still, we have knowledge about their traffic characteristics. This can be actualized by central

Table 1: System Analysis Parameters.

Par.	Explanation
HU	Hybrid user (our user)
λ_{HU}	The mean arrival rate of HUs for content request
N	The total number of contents
$s(\hat{v_b})$	The mean content size requested by a HU
s	The Zipf parameter
c_i	The i^{th} content in the content set
$p_{c_i}(s)$	The request probability for content c_i based on Zipf distribution
$\lambda_{HU}^{c_i}$	The mean request rate of content c_i by HUs
$p_{c_i}^{lo}$	The probability of local availability for content c_i
$p_{c_i}^{BS}$	The probability of BS availability for content c_i
$p_{c_i}^{sat}$	The probability of satellite availability for content c_i
μ_{HU}^{sat}	The service rate for HUs getting content from the satellite
μ_{HU}^{BS}	The service rate for HUs getting content from the BS
μ_{HU}^{D}	The service rate for HUs getting content in D2D mode
$\mu_{HU}^{sat(u)}$	The service rate of HUs that fetch content from the universal source across the satellite
$\mu_{HU}^{BS(u)}$	The service rate of HUs that fetch content from the universal source across the BS
λ_{PU}^{ter}	The mean arrival rate of primary users at terrestrial link
μ_{PU}^{ter}	The service rate of primary users at terrestrial link
$N_{f_{sat}}$	The total number of satellite frequencies
$N_{f_{ter}}$	The total number of terrestrial frequencies
x	The channel state
$idle_s(x)$	The number of idle frequencies at the satellite link segment at channel state x
$idle_{t,\overline{f_1}}(x)$	The number of idle frequencies at the terrestrial link except for the frequency f_1 at channel state x
$\lambda_{N_{HU}}$	The mean density of HUs located in the BS cell
D_{max}	The maximum number of concurrent D2D operations allowed by the network
R_{BS}	The radius of the BS cell
R_{Int}	The HU device transmission range radius that causes interference to active HU receivers
$p_{c_i}^{D(f_1)}(x)$	The D2D content availability probability of content c_i for channel state x
r_{sat}	The weight of the satellite mode
r_{BS}	The weight of the BS mode
r_{dev}	The weight of the D2D mode

capabilities of the BS as a facilitator of cognitive operation [6]. BS performs the centralized RA function referring to mode selection for content retrieval over network links. Content requests are taken as arrivals in our network following Poisson distribution as commonly done in the literature [15, 16]. Besides, multimedia traffic completions are modeled by exponential distributions [5, 16]. In that regard, we model the content retrieval completions as exponentially distributed departures. Furthermore, Markov chains (MCs) are widely used to model multimedia traffic with a compact network view [18, 19, 16]. Thus, we also model the RA for HU content retrieval in a non-time slotted manner as a continuous time MC (CTMC). Our system analysis parameters are provided in Table 1.

PUs have priority over SU-mode HUs.After interruption due to PU appearance, the Markov property is satisfied by HUs via continuing content fetch from the same system unit (if some idle frequency exists). In our previous work [5], all frequencies are used in a non-overlay setting. In this work, we have a more advanced system model in that regard: the network has a non-overlay setting in satellite and BS modes but it operates in overlay setting in

Table 2: State definitions.

Part	Definition
i_{HU}^{sat}	The number of satellite frequencies where HUs retrieve contents directly from the satellite cache
$i_{HU}^{sat(u)}$	The number of satellite frequencies where HUs retrieve contents across the satellite from the universal source
$i_{PU}^{ter(\overline{f_1})}$	The number of terrestrial frequencies used by PUs except for terrestrial frequency f_1
i_{HU}^{BS}	The number of terrestrial frequencies where HUs retrieve contents directly from the BS cache
$i_{HU}^{BS(u)}$	The number of terrestrial frequencies where HUs retrieve contents across the BS from the universal source
$i_{PU}^{ter(f_1)}$	The indicator for terrestrial frequency f_1 if it is used by PU or not
$i_{HU}^{D(f_1)}$	The number of concurrent D2D HU transmissions used for content retrieval via terrestrial frequency f_1

D2D mode. To enable overlaying, at least one terrestrial frequency needs to be considered in D2D communications. However, with each additional frequency operating in D2D mode, the cognitive operation complexity increases. For keeping the analytical model compact and tractable without sacrificing its essence, one terrestrial frequency is used for HUs in the overlay-enabled D2D mode.

In our Markov model, for the calculation of mean service completion transitions (content retrieval completions), first we need to calculate channel capacities for content fetching. We calculate these capacities by Shannon's capacity formula under Additive White Gaussian Noise (AWGN) according to free space path model. The service rate for HUs that get the requested content over different system units such as the satellite, the BS or in D2D mode is $\mu_{HU}^{x} := \frac{C_{HU}^{x}}{s(\hat{v_b})}$ $x \in \{sat, BS, D\}$. The service rate for PUs at terrestrial link is $\mu_{PU}^{ter} := \frac{C_{PU}^{ter}}{s(\hat{v_b})}$.

The integration of the *universal source* concept into our analytical Markov model for such a HetNet with D2D+ cognitive communications and content dissemination is an important contribution. It is needed for a more realistic construction. The reason is some of the contents may not be available in the caches in our zone of interest and they need to be fetched from external repositories. When the universal source is used, the content transmissions take longer amount of time. The average channel capacity between the satellite and universal source $C_{HU}^{sat(u)}$ is listed in Table 5. $\Delta_{HU}^{x(u)} := s(\hat{v_b})/C_{HU}^{x(u)} + s(\hat{v_b})/C_{HU}^{x}$ is the mean aggregate service duration of a content fetch from the universal source across the satellite or BS to the req-HU where $x \in \{sat, BS\}$. Then by taking the reciprocal $\mu_{HU}^{x(u)} := \frac{1}{\Delta_{HU}^{x(u)}}$, we get the service rates of HUs from the universal source.

$$i_{HU}^{sat}, i_{HU}^{sat(u)}, i_{PU}^{ter(\overline{f_1})}, i_{HU}^{BS}, i_{HU}^{BS(u)}, i_{PU}^{ter(f_1)}, i_{HU}^{D(f_1)}$$

Figure 3: Channel state.

The single terrestrial frequency f_1 is used for D2D mode while the other terrestrial frequencies operate in BS mode. A state consists of seven components as shown in

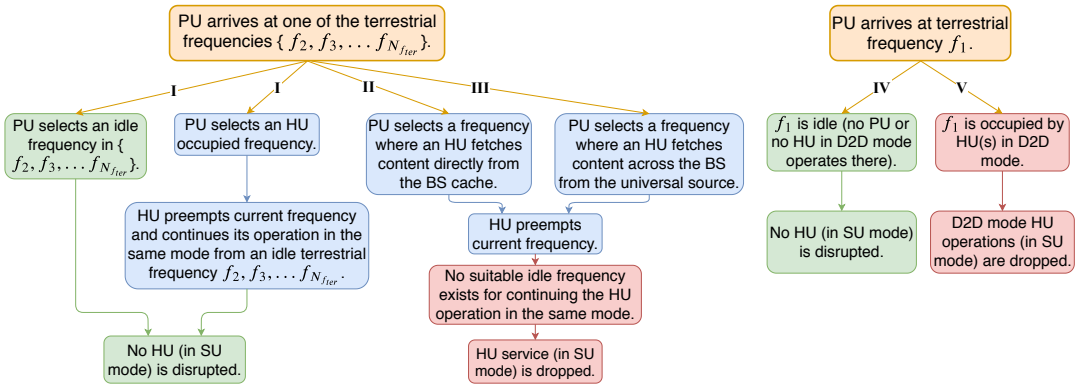


Figure 4: PU arrival layout (green:no drop, red:drop, blue:preemption, orange:case selection).

Table 3: Transitions originating at a generic state s_0 due to PU arrivals.

	Dest. State	Transition Rate
I	$s_{(i_{PU}^{ter(\overline{f_1})}+1)}$	$\frac{(N_{fter}-1)\lambda_{PU}^{ter}}{N_{fter}} 1_{(idle_{t,\overline{f_1}}(s_0)>0)}$
II	$s_{(i_{PU}^{ter(\overline{f_1})}+1, i_{HU}^{BS}-1)}$	$\frac{(N_{fter}-1)\lambda_{PU}^{ter}}{N_{fter}} \cdot \frac{i_{HU}^{BS}(s_0)}{(N_{fter}-1)-i_{PU}^{ter(\overline{f_1})}(s_0)} \cdot [1_{((idle_{t,\overline{f_1}}(s_0)==0)\wedge((N_{fter}-1)-i_{PU}^{ter(\overline{f_1})}(s_0)>0))}]$
III	$s_{(i_{PU}^{ter(\overline{f_1})}+1, i_{HU}^{BS(u)}-1)}$	$\frac{(N_{fter}-1)\lambda_{PU}^{ter}}{N_{fter}} \cdot \frac{i_{HU}^{BS(u)}(s_0)}{(N_{fter}-1)-i_{PU}^{ter(\overline{f_1})}(s_0)} \cdot [1_{((idle_{t,\overline{f_1}}(s_0)==0)\wedge((N_{fter}-1)-i_{PU}^{ter(\overline{f_1})}(s_0)>0))}]$
IV	$s_{(i_{PU}^{ter(f_1)}+1)}$	$\frac{\lambda_{PU}^{ter}}{N_{fter}} 1_{((i_{PU}^{ter(f_1)}(s_0)==0)\wedge(i_{HU}^{D(f_1)}(s_0)==0))}$
V	$s_{(i_{PU}^{ter(f_1)}+1, i_{HU}^{D(f_1)}=0)}$	$\frac{\lambda_{PU}^{ter}}{N_{fter}} 1_{(i_{HU}^{D(f_1)}(s_0)>0)}$

Fig. 3. Their definitions are given in Table 2.

A channel state transition occurs due to PU/HU arrival/departure. If a user arrives, we increment the corresponding type of user in the channel state. After a content is completely retrieved, the user departs the channel. During RA leading to mode selection, we first check the content availability at different system units, and for choosing among them, we consider the channel states: for each available frequency, we assign mode weight and decide on the channel access according to the output of our RA function. By tuning these weights, we investigate how EE and overall system goodput are affected.

5.1. PU Transitions

Our HUs access terrestrial link opportunistically.Therefore, we investigate how the terrestrial PU activities impact the behaviour of HUs. HUs operate **in D2D mode** at the terrestrial frequency $\boldsymbol{f_1}$. For the **other terrestrial frequencies**, HUs operate in **BS mode** as described in Sec. 5.3. Hence, the HU mode characteristics are different between terrestrial frequency f_1 and others. For processing the preemptions of HUs in D2D or BS modes, we define (i) $i_{PU}^{ter(f_1)}$, (ii) $i_{PU}^{ter(\overline{f_1})}$. We denote the state in Fig. 3 as s_0 and elaborate on PU arrival cases originating at s_0 in Table 3. Compared to the generic state s_0, the incremented parts (arrivals) and/or decremented parts (departures) are represented with x in any destination state $s_{(x)}$. We also define some utility functions, $idle_s(x) := N_{f_{sat}} - i_{PU}^{sat}(x) - i_{HU}^{sat}(x) - i_{HU}^{sat(u)}(x)$ and $idle_{t,\overline{f_1}}(x) := (N_{f_{ter}}-1) - i_{PU}^{ter(\overline{f_1})}(x) - i_{HU}^{BS}(x) - i_{HU}^{BS(u)}(x)$ as explained in Table 1. $1_{(\theta)}$ is the indicator function defined as 1 if θ is true, 0 otherwise (these functions are also used in Sec. 5.3). We give a detailed layout in Fig. 4 for each PU arrival case listed in Table 3. Besides, PU service completion transitions and corresponding rates are available in [17].

5.2. D2D Operation Mode

Due to mobility in a wireless network, device locations show stochastic behaviour. For modeling such dynamic situation, a common technique is to employ a spatial model with devices distributed by Poisson Point Process (PPP) in the spatial domain [20, 21]. In our analysis, HUs are randomly located in the BS cell following a PPP with mean density $\lambda_{N_{HU}}$ in a similar setting. D2D operations are handled at the terrestrial frequency f_1 with **overlaying**. For any content c_i, the D2D content availability probability $p_{c_i}^{D(f_1)}(x)$ of channel state x is calculated in (3). D_{max} is the maximum number of concurrent D2D operations allowed in the network. The D2D content availability probability $p_{c_i}^{D(f_1)}(x)$ is zero if the number of concurrent D2D operations had reached D_{max}. Otherwise, it is the multiplication of the following probabilities:

- $\Pi_{rec}(x)$ in (4) : the receiver HU (rec-HU) is not being interfered by other D2D operations.
- $\Pi_{tx}(x)$ in (5) : the transmitter HU (tx-HU) will not cause interference to active D2D operations.
- Π_{c_i} in (6) : content c_i is retrievable over the terrestrial frequency f_1.

$$p_{c_i}^{D(f_1)}(x) := 1_{(0<=i_{HU}^{D(f_1)}(x)<D_{max})} \Pi_{rec}(x)\Pi_{tx}(x)\Pi_{c_i} \quad (3)$$

$$\Pi_{rec}(x) := \frac{max\{0,(\pi R_{BS}^2)-(i_{HU}^{D(f_1)}(x)\pi R_{Int}^2)\}}{\pi R_{BS}^2} \quad (4)$$

$$\Pi_{tx}(x) := \frac{max\{0,(\pi R_{BS}^2)-(i_{HU}^{D(f_1)}(x)\pi (2R_{Int})^2)\}}{\pi R_{BS}^2} \quad (5)$$

$$\Pi_{c_i} := 1-(1-p_{c_i}^{lo})^{\lfloor \lambda_{N_{HU}} \pi R_{Int}^2 \rfloor} \quad (6)$$

For modeling interference in [22], *interference range* is defined as the minimum distance to avoid concurrent transmissions interfering with each other. Similarly, in

our work, we define R_{Int} as the radius of the transmission range of an HU device that causes interference to active rec-HUs at the terrestrial frequency f_1. The interference to a D2D transmission at the terrestrial frequency f_1 out of this range is assumed to be negligible.

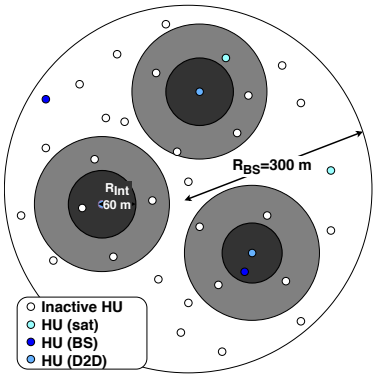


Figure 5: D2D spatial stochastic model.

$\Pi_{rec}(x)$ is calculated by subtracting the interference ranges of active tx-HU devices in D2D mode (dark shaded areas in Fig. 5 $\rightarrow i_{HU}^{D(f_1)}(x)\pi R_{Int}^2$) from the cell area ($\pi R_{BS}^2$) and then then dividing over πR_{BS}^2. The `max` function is used to assure probability $\Pi_{rec}(x)$ is non-negative.

A rec-HU in D2D mode is at most R_{Int} away from its tx-HU. If simultaneously another HU at most R_{Int} away from rec-HU also actively transmits then this may lead to collision at the rec-HU. To avoid such collisions, candidate tx-HUs are prohibited from concurrently operating in D2D mode in the $\pi(2R_{Int})^2$ area for each active rec-HU. We aggregate these ranges for all rec-HUs as a prohibition zone for new tx-HU candidates by $i_{HU}^{D(f_1)}(x)(\pi 2R_{Int})^2$. $\Pi_{tx}(x)$ is calculated by subtracting this aggregated prohibition zone for tx-HU candidates (dark shaded area+light shaded area in Fig. 5$\rightarrow i_{HU}^{D(f_1)}(x)(\pi 2R_{Int})^2$) from the cell area ($\pi R_{BS}^2$) and dividing over πR_{BS}^2. Again, `max` function is used.

For the content reception, requested c_i should be at some HU in the reception range of the requester HU (req-HU). The multiplication of $\lambda_{N_{HU}}$ with the area of this range (πR_{Int}^2) gives the average number of HU devices storing c_i in this area. $(1-p_{c_i}^{lo})^{\lfloor \lambda_{N_{HU}} \pi R_{Int}^2 \rfloor}$ is the probability that no HU device has content c_i in the reception range of req-HU. By taking its complement, Π_{c_i} in (6) gives the probability of finding at least one HU device storing content c_i in the reception range of req-HU.

The interference range of active tx-HUs in D2D mode (dark shaded areas) can intersect with the BS cell boundary. Besides, further occurrences explained in [17] lead to $p_{c_i}^{D(f_1)}(x)$ serving as a lower-bound for D2D content availability probability.

5.3. HU Transitions

The core component of our system is the mode selection. Our mode selection scheme for HU content requests considers caches of system units (caches of the satellite, BS, HU devices), channel state and mode weights (r_{sat}, r_{BS}, r_{dev}). These weights are configurable system parameters where $r_{sat}+r_{BS}+r_{dev}=1$. They control how likely a mode is selected for content dissemination. For the rigorous analysis, first we define some basic functions used in this context. We utilize aggregate mode weight functions $R_{sat}(x)$, $R_{BS}(x)$, $R_{D2D}(x)$ of a channel state x, defined in (7) - (9), for mode selection.

$$R_{sat}(x) := r_{sat} \cdot idle_s(x) \tag{7}$$

$$R_{BS}(x) := r_{BS} \cdot idle_{t,\overline{f_1}}(x) \tag{8}$$

$$R_{D2D}(x) := r_{dev} \cdot [1_{(0<i_{HU}^{D(f_1)}(x)<D_{max})} + 1_{((i_{HU}^{D(f_1)}(x)==0)\wedge(i_{PU}^{ter(f_1)}(x)==0))}] \tag{9}$$

We give an example scenario where our aggregate mode weight functions are useful in terms of mode selection. Consider at channel state x, an HU requests content c_i available in the BS and in some HU device in the reception range. $R_{BS}(x)$, the aggregate BS mode weight function, assigns BS mode weight (r_{BS}) for each idle terrestrial frequency among f_2,f_3,...$f_{N_{f_{ter}}}$. Similarly, the aggregate weight of D2D mode $R_{D2D}(x)$ is determined by r_{dev} for the terrestrial frequency f_1 if it is idle or used by less than D_{max} concurrent D2D operations. Then $\frac{R_{BS}(x)}{R_{BS}(x)+R_{D2D}(x)}$ is the probability of retrieving c_i in BS mode and $\frac{R_{D2D}(x)}{R_{BS}(x)+R_{D2D}(x)}$ in D2D mode.

For HU transition inspection, the state in Fig. 3 is denoted as h_0. First, we analyze the transitions originating at h_0 upon HU arrivals listed in Table 4 and visualized in Figure 6. When a requester-HU (req-HU) requests content c_i with rate $\lambda_{HU}^{c_i}$, our RA mechanism first analytically calculates the local content availability $p_{c_i}^{lo}$. With probability 1-$p_{c_i}^{lo}$, content is not found in the local cache and there are five different possible states for service mode. The system will calculate the possibility of choosing each of these modes (shown in Fig. 6) to serve the req-HU: **i) satellite mode (direct)**: c_i is fetched from the satellite cache to req-HU. **ii) satellite mode (from universal)**: first fetched from the universal source to the satellite cache and then from there to req-HU. **iii) BS mode (direct)**: c_i is fetched from the BS cache to req-HU. **iv) BS mode (from universal)**: first fetched from the universal source to the BS cache and then from there to req-HU. **v) D2D mode**: c_i fetched from the cache of some HU device in reception range of req-HU.

Next, for the retrieval of c_i, our RA mechanism analytically calculates the transition rates to each aforementioned modes. For instance, let us consider the transition rate of mode-i (*satellite mode direct*) as shown in Fig. 6 (1). While calculating the corresponding transition rate, we branch into each content availability combination. These branches i-a, i-b, i-c and i-d and corresponding transition rates are provided in Table 4. We sum over these rates to get the aggregate transition rate of the mode-i service request for the retrieval of content c_i in (10). By summing (10) for all c_i's, we get the expected arrival rate of mode-i

Table 4: Transitions originated at a generic state h_0 due to HU Arrivals.

Id	Content availability	Dest. State	Prob. of content availability	Transition Rate
i-a	satellite cache only	$s_{(i_{HU}^{sat}+1)}$	$\mathbb{P}^{c_i}_{(S)}=[1\text{-}p^{lo}_{c_i}]\cdot p^{sat}_{c_i}\cdot[1\text{-}p^{BS}_{c_i}]\cdot[1\text{-}p^{D(f_1)}_{c_i}(h_0)]$	$\gamma^{sat}_{HU}(i,\{sat\})=\lambda^{c_i}_{HU}\mathbb{P}^{c_i}_{(S)}1_{((idle_s(h_0)>0)\wedge(r_{sat}>0))}$
i-b	satellite and BS cache	$s_{(i_{HU}^{sat}+1)}$	$\mathbb{P}^{c_i}_{(S,B)}=[1\text{-}p^{lo}_{c_i}]\cdot p^{sat}_{c_i}\cdot p^{BS}_{c_i}\cdot[1\text{-}p^{D(f_1)}_{c_i}(h_0)]$	$\gamma^{sat}_{HU}(i,\{sat,BS\})=\lambda^{c_i}_{HU}\mathbb{P}^{c_i}_{(S,B)}[\frac{R_{sat}(h_0)}{R_{sat}(h_0)+R_{BS}(h_0)}]$ $1_{(R_{sat}(h_0)+R_{BS}(h_0)>0)}$
i-c	satellite cache and some HU device cache within the reception range of req-HU	$s_{(i_{HU}^{sat}+1)}$	$\mathbb{P}^{c_i}_{(S,D)}=[1\text{-}p^{lo}_{c_i}]\cdot p^{sat}_{c_i}[1\text{-}p^{BS}_{c_i}]\cdot p^{D(f_1)}_{c_i}(h_0)$	$\gamma^{sat}_{HU}(i,\{sat,Dev\})=\lambda^{c_i}_{HU}\mathbb{P}^{c_i}_{(S,D)}[\frac{R_{sat}(h_0)}{R_{sat}(h_0)+R_{D2D}(h_0)}]$ $1_{(R_{sat}(h_0)+R_{D2D}(h_0)>0)}$
i-d	satellite cache, BS cache, some HU device cache within the reception range of req-HU	$s_{(i_{HU}^{sat}+1)}$	$\mathbb{P}^{c_i}_{(S,B,D)}=[1\text{-}p^{lo}_{c_i}]\cdot p^{sat}_{c_i}\cdot p^{BS}_{c_i}\cdot p^{D(f_1)}_{c_i}(h_0)$	$\gamma^{sat}_{HU}(i,\{sat,BS,Dev\})=\lambda^{c_i}_{HU}\mathbb{P}^{c_i}_{(S,B,D)}[\frac{R_{sat}(h_0)}{R_{sat}(h_0)+R_{BS}(h_0)+R_{D2D}(h_0)}]$ $1_{(R_{sat}(h_0)+R_{BS}(h_0)+R_{D2D}(h_0)>0)}$
v-a	some HU device in the reception range of req-HU only	$s_{(i_{HU}^{D(f_1)}+1)}$	$\mathbb{P}^{c_i}_{(D)}=[1\text{-}p^{lo}_{c_i}]\cdot[1\text{-}p^{sat}_{c_i}]\cdot[1\text{-}p^{BS}_{c_i}]\cdot p^{D(f_1)}_{c_i}(h_0)$	$\gamma^{D(f_1)}_{HU}(i,\{Dev\})=\lambda^{c_i}_{HU}\mathbb{P}^{c_i}_{(D)}[1_{(r_{dev}>0)}]$ $[(1_{(0<i^{D(f_1)}_{HU}(h_0)<D_{max})})+(1_{((i^{D(f_1)}_{HU}(h_0)==0)\wedge(i^{ter(f_1)}_{PU}(h_0)==0))})]$
v-b	satellite cache and some HU device cache in the reception range of req-HU	$s_{(i_{HU}^{D(f_1)}+1)}$	$\mathbb{P}^{c_i}_{(S,D)}=[1\text{-}p^{lo}_{c_i}]\cdot p^{sat}_{c_i}\cdot[1\text{-}p^{BS}_{c_i}]\cdot p^{D(f_1)}_{c_i}(h_0)$	$\gamma^{D(f_1)}_{HU}(i,\{sat,Dev\})$
v-c	BS cache and some HU device cache in the reception range of req-HU	$s_{(i_{HU}^{D(f_1)}+1)}$	$\mathbb{P}^{c_i}_{(B,D)}=[1\text{-}p^{lo}_{c_i}]\cdot[1\text{-}p^{sat}_{c_i}]\cdot p^{BS}_{c_i}\cdot p^{D(f_1)}_{c_i}(h_0)$	$\gamma^{D(f_1)}_{HU}(i,\{BS,Dev\})$
v-d	satellite cache, BS cache and some HU device cache in the reception range of req-HU	$s_{(i_{HU}^{D(f_1)}+1)}$	$\mathbb{P}^{c_i}_{(S,B,D)}=[1\text{-}p^{lo}_{c_i}]\cdot p^{sat}_{c_i}\cdot p^{BS}_{c_i}\cdot p^{D(f_1)}_{c_i}(h_0)$	$\gamma^{D(f_1)}_{HU}(i,\{sat,BS,Dev\})$

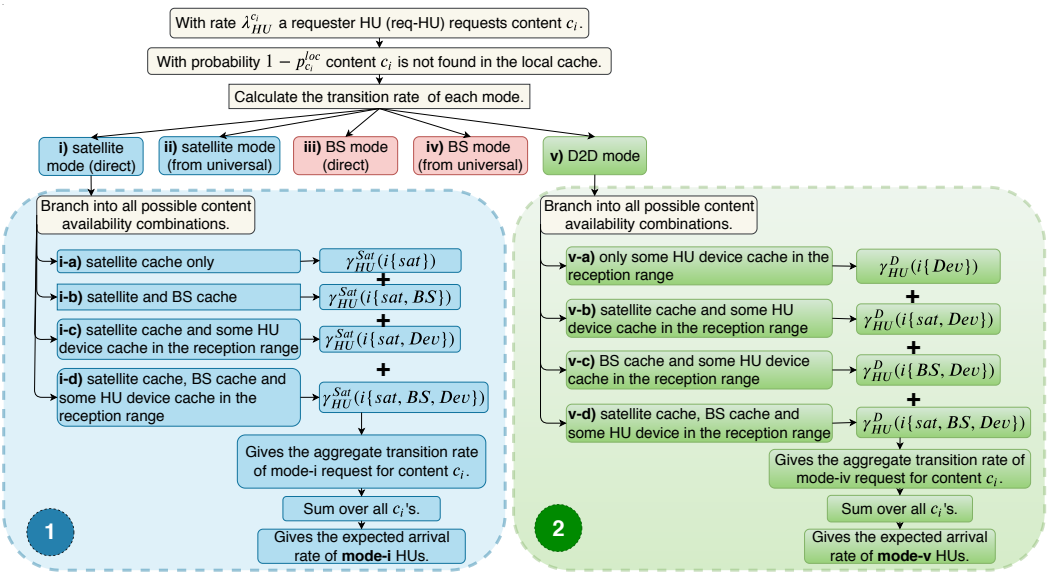


Figure 6: Scheme for HU arrival state transition calculation(blue:satellite, red:BS, green:D2D).

HUs ($\Gamma^{sat}_{HU}:=\sum_{i=1}^{N}\Gamma^{sat}_{HU}(i)$) into the network. This is the general scheme for the mode-i state transition calculation.

$$\Gamma^{sat}_{HU}(i):=\gamma^{sat}_{HU}(i,\{sat\})+\gamma^{sat}_{HU}(i,\{sat,BS\})+\gamma^{sat}_{HU}(i,\{sat,Dev\})+\gamma^{sat}_{HU}(i,\{sat,BS,Dev\}) \quad (10)$$

Now, let us explain some of the mode-i branch calculations. For instance, when we look at the branch (i-a) (in Table 4), the requested content c_i is only in the satellite. The corresponding probability $\mathbb{P}^{c_i}_{(S)}$ is given by $[1-p^{lo}_{c_i}]p^{sat}_{c_i}[1-p^{BS}_{c_i}][1-p^{D(f_1)}_{c_i}(h_0)]$. c_i is fetched from the satellite if the satellite link is available and the satellite mode weight is greater than zero ($1_{((idle_s(h_0)>0)\wedge(r_{sat}>0))}$). The corresponding rate of this branch is denoted as $\gamma^{sat}_{HU}(i,\{sat\})$. In branch (i-b), the requested content c_i is in the satellite and BS cache but nowhere else. Its probability is $\mathbb{P}^{c_i}_{(S,B)}$. For selecting between satellite and BS, $R_{sat}(h_0)$ and $R_{BS}(h_0)$ are utilized. With probability $\frac{R_{sat}(h_0)}{R_{sat}(h_0)+R_{BS}(h_0)}$, c_i retrieved from the satellite cache. The corresponding branch rate is denoted as $\gamma^{sat}_{HU}(i,\{sat,BS\})$. The transitions (i-c) and (i-d) are available in Table 4.

When we consider mode-v (*D2D mode*) transition, the aggregate transition rate of mode-v service request for the retrieval of content c_i is given in (11). When transitions over all contents are aggregated, we get the expected arrival rate of mode-v HUs into the network ($\Gamma^{D(f_1)}_{HU}:=\sum_{i=1}^{N}\Gamma^{D(f_1)}_{HU}(i)$). The general scheme for this calculation (2) is available in Fig. 6.

$$\Gamma^{D(f_1)}_{HU}(i):=\gamma^{D(f_1)}_{HU}(i,\{Dev\})+\gamma^{D(f_1)}_{HU}(i,\{sat,Dev\})+\gamma^{D(f_1)}_{HU}(i,\{BS,Dev\})+\gamma^{D(f_1)}_{HU}(i,\{sat,BS,Dev\}) \quad (11)$$

To exemplify a branch, when we look at (v-a) (in Table 4), the probability of finding content c_i *only* at some HU device in the reception range is $\mathbb{P}^{c_i}_{(D)}$. In this case c_i is retrieved only if $r_{dev}>0$. Besides, either the terrestrial frequency f_1 should be idle ($1_{((i^{D(f_1)}_{HU}(h_0)==0)\wedge(i^{ter(f_1)}_{PU}(h_0)==0))}$) or the number of maximum concurrent D2D transmission(s) has not been reached ($1_{(0<i^{D(f_1)}_{HU}(h_0)<D_{max})}$).

In all modes, the transitions for all branches and de-

tailed calculations of state transitions are available in [17]. The HU service completion transitions and corresponding rates are also available in [17]. Together with them, we get the complete set of balance equations. Now, we can calculate the steady state probabilities π_x of being at channel state x. These π_x's are utilized in the definition of utility functions provided in the Performance Metrics section.

6. Performance Metrics

Our analytical model provides an apparatus to investigate our HetNet for its performance characteristics. We define two key system metrics, namely *goodput* and *energy efficiency*, based on the system parameters. As already explained, the network supports five modes: *i)* satellite mode (direct) *ii)* satellite mode (from universal), *iii)* BS mode (direct), *iv)* BS mode (from universal), *v)* D2D mode. Before defining system metrics, we define some utility functions.

The HU effective arrival rates for five different modes are defined in (12)-(16). Their detailed explanation and derivations are provided in [17].

$$\lambda^{sat}_{eff(HU)} := \sum_{x\in S} \Gamma^{sat}_{HU}(x)\pi_x \tag{12}$$

$$\lambda^{sat(u)}_{eff(HU)} := \sum_{x\in S} (\sum_{i=1}^{N} \gamma^{sat(u)}_{HU}(i,x))\pi_x \tag{13}$$

$$\lambda^{BS}_{eff(HU)} := \sum_{x\in S} \Gamma^{BS}_{HU}(x)\pi_x \tag{14}$$

$$\lambda^{BS(u)}_{eff(HU)} := \sum_{x\in S} (\sum_{i=1}^{N} \gamma^{BS(u)}_{HU}(i,x))\pi_x \tag{15}$$

$$\lambda^{D2D}_{eff(HU)} := \sum_{x\in S} \Gamma^{D(f_1)}_{HU}(x)\pi_x \tag{16}$$

The dropping probability of HUs in BS mode is p^{BS}_{drop} and in D2D mode is p^{D2D}_{drop}. The probability of an HU getting service from its local cache is p_{local}. Please refer to [17] for their definitions and explanations.

6.1. Goodput

We investigate the overall system goodput. To this end, we calculate the throughput (the rate HU content requests are served) through aforementioned network modes.

HUs in the satellite link are in PU mode, so the effective arrival rate $\lambda^{sat}_{eff(HU)}$ in (12) is equal to the effective service rate of mode-*i* HUs. Multiplying this with average content size $s(\hat{v}_b)$, we get the mode-*i* HU throughput Th^{sat}_{HU} (contents fetched directly from the satellite). The HUs throughput in mode-*ii* is calculated similarly as $Th^{sat(u)}_{HU} := \lambda^{sat(u)}_{eff(HU)} \cdot s(\hat{v}_b)$.

The mode-*iii* HU throughput Th^{BS}_{HU} is $\lambda^{BS}_{eff(HU)} \cdot (1-p^{BS}_{drop}) \cdot s(\hat{v}_b)$. For the effective service rate, dropped contents are excluded by 1-p^{BS}_{drop} since they do not contribute to successful transmissions. Multiplying mode-*iii* HUs arrival rate $\lambda^{BS}_{eff(HU)}$ in (14) with 1-p^{BS}_{drop} gives the effective service rate of mode-*iii* HUs. The mode-*iv* HUs throughput $Th^{BS(u)}_{HU} := \lambda^{BS(u)}_{eff(HU)} \cdot (1-p^{BS}_{drop}) \cdot s(\hat{v}_b)$ and mode-*v* HUs throughput (D2D mode) $Th^{D}_{HU} := \lambda^{D2D}_{eff(HU)} \cdot (1-p^{D2D}_{drop}) \cdot s(\hat{v}_b)$ are calculated similarly.

For local hits, we look at the G^{local}_{HU} value. The effective request rate of HUs over the local cache is equal to the request arrival rate λ_{HU} times the probability of an HU getting service locally p_{local}. For the calculation of service rate in bps, this effective request rate is multiplied by the average content size $s(\hat{v}_b)$.

$$G^{local}_{HU} := \lambda_{HU} \cdot p_{local} \cdot s(\hat{v}_b) \tag{17}$$

The overall system goodput of HUs is the summation of services taken without using network sources (requested content found in the local cache, G^{local}_{HU}) and the summation of services given over the network in bps:

$$G_{HU} := G^{local}_{HU} + Th^{sat}_{HU} + Th^{sat(u)}_{HU} + Th^{BS}_{HU} + Th^{BS(u)}_{HU} + Th^{D}_{HU} \tag{18}$$

6.2. Energy Efficiency

Energy consumption is a crucial criterion to evaluate the performance of mode selection and characterizing our model. Energy efficiency is defined as the consumed energy in Joule per successfully transmitted bits to HUs in (20). It is calculated by the division of the consumed overall power P_{all} in (19) over the overall system goodput of HUs in (18).

$$P_{all} := P_{BS} + P_{BS(u)} + P_{D2D} + P_{loc} \tag{19}$$

$$EPB_{HU} := \frac{P_{all}}{G_{HU}} \tag{20}$$

The satellite is solar powered, so the effective power consumption P_{all} does not include that. P_{all} consists of four components: *a)* P_{BS}, *b)* $P_{BS(u)}$, *c)* P_{D2D} and *d)* P_{loc}.

P_{BS} in (21) is the BS transmission power consumption for mode-*iii* HU services either **completed** or **dropped**. $\lambda^{BS}_{eff(HU)} \cdot (1-p^{BS}_{drop})$ is the effective service rate of completed mode-*iii* HUs while the BS consumes $P^{ch}_{BS}/\mu^{BS}_{HU}$ transmission energy per such service. Multiplying them, gives the BS transmission power for completed mode-*iii* HU services.

$\lambda^{BS}_{eff(HU)} \cdot p^{BS}_{drop}$ is the rate of dropped mode-*iii* HU services. Assuming no bias, they capture in average half of a complete service ($\frac{1}{2\cdot\mu^{BS}_{HU}}$s). So, $\frac{P^{ch}_{BS}}{2\cdot\mu^{BS}_{HU}}$ is the average BS transmission energy per each such incomplete HU service. Multiplying this with $\lambda^{BS}_{eff(HU)} \cdot p^{BS}_{drop}$ gives the transmission power of the BS for dropped mode-*iii* HU services.

$$P_{BS} := (\lambda^{BS}_{eff(HU)} \cdot (1-p^{BS}_{drop}) \cdot \frac{P^{ch}_{BS}}{\mu^{BS}_{HU}}) + (\lambda^{BS}_{eff(HU)} \cdot p^{BS}_{drop} \cdot \frac{P^{ch}_{BS}}{2\cdot\mu^{BS}_{HU}}) \tag{21}$$

$P_{BS(u)}$ in (22) is the BS power consumption for mode-*iv* HU services consisting of two service types: **i)** completed **ii)** dropped. While calculating $P_{BS(u)}$, we consider additional cost imposed by the universal source integration, namely the BS reception energy. P_{D2D} in (23) is the transmission power expenditure of HU devices operating in mode-*v*. The details of the $P_{BS(u)}$ and P_{D2D} are available in [17].

$$P_{BS(u)} := \{\lambda^{BS(u)}_{eff(HU)} \cdot (1-p^{BS}_{drop}) \cdot ([\frac{P^{ch}_{BS}}{\mu^{BS}_{HU}}] + [\frac{P^{ch}_{BS}/\theta_{BS}}{C^{BS(u)}_{HU}/s(\hat{v}_b)}])\} \tag{22}$$
$$+ \{\lambda^{BS(u)}_{eff(HU)} \cdot p^{BS}_{drop} \cdot ([(P^{ch}_{BS} \cdot (\frac{\Delta^{BS(u)}_{HU}}{2} - \frac{s(\hat{v}_b)}{C^{BS(u)}_{HU}})] + [\frac{P^{ch}_{BS}/\theta_{BS}}{C^{BS(u)}_{HU}/s(\hat{v}_b)}])\}$$

$$P_{D2D} := (\lambda^{D2D}_{eff(HU)} \cdot (1-p^{D2D}_{drop}) \cdot \frac{P^{tx}_{dev}}{\mu^{D}_{HU}}) + (\lambda^{D2D}_{eff(HU)} \cdot p^{D2D}_{drop} \cdot \frac{P^{tx}_{dev}}{2\cdot\mu^{D}_{HU}}) \tag{23}$$

Table 5: Simulation parameters and values.

Par.	Explanation	Value
P_{BS}^{ch}	Per channel transmission power of the BS	6 W
P_{dev}^{tx}	Transmission power of a hybrid user device	80 mW
d_{sat}	Distance from LEO satellite to earth	300 km
d_{BS}	Mean distance of a PU and/or HU to the BS	150 m
d_{D2D}	Mean distance between receiver and sender HUs	30 m
W_{ter}	Bandwidth of terrestrial link	2 MHz
W_{sat}	Bandwidth of satellite link	36 MHz
f_{sat}	Frequency of satellite link	20 GHz
f_{ter}	Frequency of terrestrial link	700 MHz
N	Total number of contents	20
s	Zipf parameter	1.2
λ_{HU}	Arrival rate of hybrid users for content request	2.4 $\frac{user}{sec}$
λ_{PU}^{ter}	Arrival rate of primary users at terrestrial link	0.03 $\frac{user}{sec}$
Cac_{sat}	Satellite cache size	125 Mbs
Cac_{BS}	Base station cache size	100 Mbs
Cac_{Dev}	HU cache size	50 Mbs
$C_{HU}^{sat(u)}$	The average channel capacity between the satellite and universal source	1 Mbps
$C_{HU}^{BS(u)}$	The average channel capacity between the BS and universal source	10 Mbps

Some HU requests are satisfied by the local caches with power consumption $P_{loc} := (\lambda_{HU} \cdot p_{local}) \cdot \frac{P_{dev}^{tx}}{\theta_{loc}} \cdot \frac{1}{\mu_{HU}^{D}}$. Here, $\lambda_{HU} \cdot p_{local}$ is the effective HU local service rate and $\frac{P_{dev}^{tx}}{\theta_{loc}} \cdot \frac{1}{\mu_{HU}^{D}}$ is the average energy consumed for each HU local service.

7. Performance Evaluation

In our study, we observe EPB_{HU} and G_{HU} as performance metrics. The objective of our performance investigation is two-fold: First of all, we perform the system simulations to compare their results with our analytical results for verifying our system model. Furthermore, we investigate the impact of different system capabilities/functions such as integration of satellite, D2D communications, cognitive operation and in-network caching on the performance characteristics. We implemented our simulator in Matlab. For each experiment case, we run the simulations 10 times, each for 1200 s. We have an event-based simulation approach. The simulator processes content request arrivals and service completions of PUs and HUs. The simulations are based on our analytical model. The PU arrivals and departures are handled as explained in the Subsection 5.1. For HUs, when a content request arrives to the system, first the local cache is checked. If the requested content is not available in the local cache, one of the service modes among *i)* satellite mode (direct) *ii)* satellite mode (from universal), *iii)* BS mode (direct), *iv)* BS mode (from universal), *v)* D2D mode is selected. This selection is done as follows: First, the content availability for system units and universal source *on/off* state at the request time are checked. Then, the aggregate mode weight functions in (7)-(9) of the selected units are calculated and one of them is selected in a random manner proportional to its weight for the content transmission. The service completions are handled by preempting the corresponding frequencies. With this event-driven scheme, we simulate our complex hybrid system. In the experimental setup, we use the parameters in Table 5. In this list, transmission power of system components (BS, HU device), mean distance of requesters to the system unit (e.g. satellite, BS, ...), channel parameters (bandwidth and frequencies), content parameters, user arrival rates, cache capacities and finally channel capacities for universal source extension are provided. In the following subsections, the EPB_{HU} and G_{HU} results of varying λ_{HU} are investigated to convey how the request density affects the system. In Subsection 7.3, the results for varying D2D mode weight are presented.

For large λ_{HU} values in Fig. 7, the simulation G_{HU} results are larger than analytical G_{HU} results for both random caching and PDC policies. The reason is for large λ_{HU}'s the impact of D2D mode services is greater and as explained in Sec. 5.2 in analytical traction $p_{c_i}^{D(f_1)}(x)$ is a lower bound for the probability of a content c_i being available in some HU device within some vicinity of requester while being retrievable. In simulations, for mode selection we are not restricted by lower bounds so we obtain more accurate and larger G_{HU} results. Thus, for large λ_{HU}'s analytical G_{HU} results are lower compared to simulation for both caching policies. Further studies on cache model and comparison to baseline caching techniques (e.g. LRU) are available in [17].

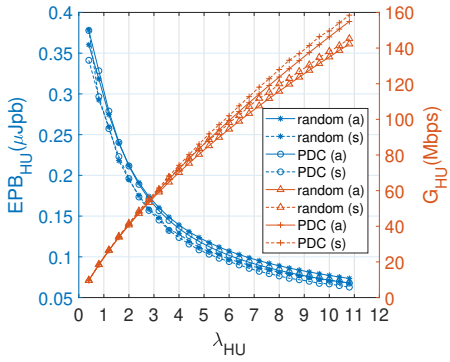
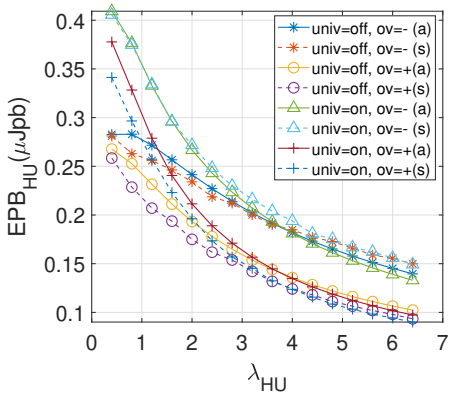


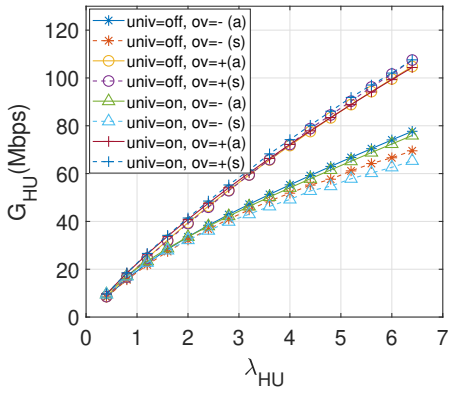
Figure 7: Caching EE and goodput results (a: analytical, s: sim.).

7.1. Integration of Universal Source and Overlaying Mechanism for D2D Operation Mode

In this part, we investigate how two key model elements affect the performance, namely universal source and overlaying mechanism for D2D operation. We tune D_{max} for enabling/disabling overlaying mechanism for D2D mode. Setting $D_{max} = 1$ means only one D2D operation is allowed which corresponds to disabled overlaying. For enabling it, we set it to five in these experiments. We consider a setup where PDC policy is used and all mode weights (r_{sat}, r_{BS}, r_{dev}) are assigned to 1/3. This way, we cancel out the effect of different system unit weighting (i.e., no favored transmission mode) to specifically focus



(a) EE results.



(b) Goodput results.

Figure 8: Universal, overlay scenarios (a: analytical, s: simulation, -: disabled, +: enabled, ov: overlay).

on universal source and overlaying mechanisms. Apparently, for all settings (universal source on/off, overlaying enabled/disabled) EPB_{HU} decreases (Fig. 8a) and G_{HU} increases (Fig. 8b) with increasing λ_{HU} rate. By introducing universal source to both D2D overlaying enabled and disabled scenarios, EPB_{HU} increases for λ_{HU} values lower than 3.2 $^{\text{user}}/_{\text{sec}}$ as shown in Fig. 8a. Unavailable contents are fetched over the universal source with extra reception energy cost at the BS and this leads to the reduction in the EE for these λ_{HU} values. For larger content request rates, the impact of D2D mode services increases and thereof the energy cost at the BS becomes a less dominant factor on EPB_{HU} metric. By enabling D2D overlaying in both universal source-on and -off scenarios, EPB_{HU} decreases for any λ_{HU} compared to the scenario without overlaying as shown in Fig. 8a, i.e., EE is improved.

With the introduction of universal source for both D2D overlaying scenarios, G_{HU} results do not change significantly for any λ_{HU} as shown in Fig. 8b. The universal source enables unavailable contents to be transmitted so previously unserved requests can then contribute to the goodput. But the services used by universal source are active for a larger amount of time, which in turn reduces the probability of these frequencies being idle. So this situation reduces transmission capacity for the corresponding frequencies and the capacity reduction affects the overall network goodput negatively. Overall, these effects roughly

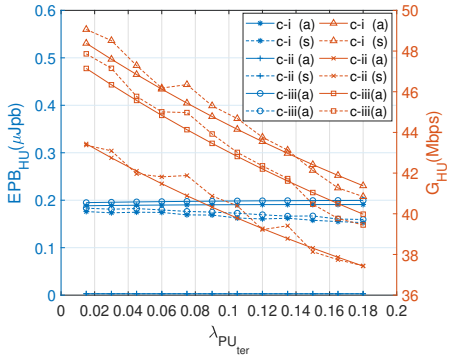


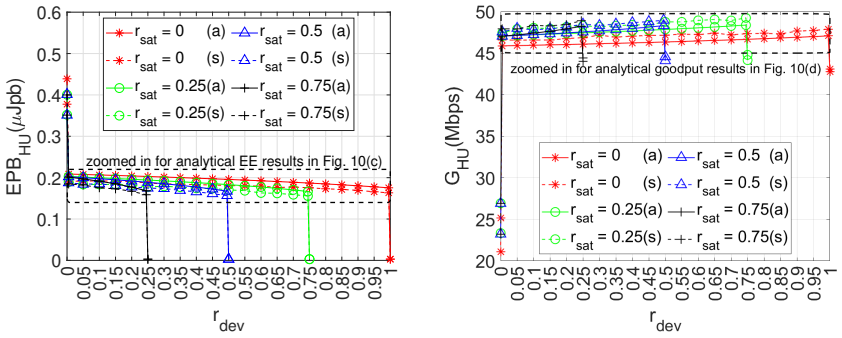
Figure 9: EE and goodput results for varying PU arrivals in terrestrial link (a: analytical, s: simulation, c: constellation).

cancel each other and hence the introduction of universal source does not significantly affect the G_{HU} results. However, with the introduction of overlaying for D2D mode, the goodput of HUs improves for both universal source on and off scenarios for any λ_{HU} in Fig. 8b.

For analysis of universal source integration, let us focus on two settings: (A) universal source on and D2D overlaying enabled (B) universal source off and D2D overlaying disabled. For $\lambda_{HU} \in (0.4, 1.2]$, the EPB_{HU} in setting (A) has larger values compared to setting (B). With the universal source integration, unavailable contents are retrieved with extra reception energy cost at the BS leading to larger EPB_{HU}. Enabling overlaying for D2D is useful for EE and is expected to reduce EPB_{HU} to alleviate the impact of universal source integration. However, the network is not in need of concurrent D2D transmissions since the low λ_{HU} value means less content requests and the request traffic is not dense enough to necessitate overlaying in D2D. Thereof, for $\lambda_{HU} \in (0.4, 1.2)$, universal source impact is dominant and setting (A) has larger EPB_{HU} than (B). In $\lambda_{HU} \in (1.2, 1.6)$ regime, EPB_{HU} of both settings intersect and for $\lambda_{HU} \in [1.6, 6.4)$ regime, setting (A) attains lower EPB_{HU} value than (B). With larger λ_{HU} the network becomes needy for concurrent D2D transmissions and hence in setting (A) with D2D overlaying, D2D services start to rectify the negative EE impact of universal source leading to lower EPB_{HU} (improved EE) compared to setting (B). Note that the simulation results follow the same trend with the analytical EPB_{HU} and G_{HU} results for all scenarios.

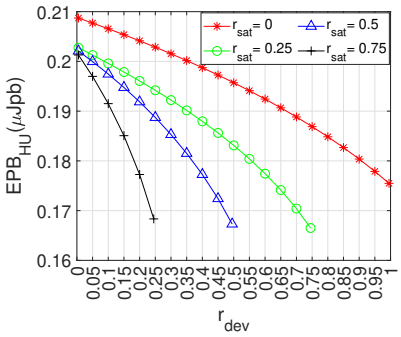
7.2. Impact of Primary User Activity in Terrestrial Frequencies

Another important research question is how our model behaves for different PU activity. EPB_{HU} and G_{HU} results for increasing λ_{PU}^{ter} are shown in Fig. 9. We look at varying λ_{PU}^{ter} as our HUs are in cognitive mode in the terrestrial link. We assume the universal source is on and D2D overlaying is enabled. The arrival rate of HU requests is λ_{HU} =2.4 $^{\text{user}}/_{\text{sec}}$. The λ_{PU}^{ter} range we investigate is [0.015, 0.18] $^{\text{user}}/_{\text{sec}}$ as HUs are the driving source of the traffic and thus we assume light PU traffic at the terrestrial link. We investigate three different mode weight constellations: (i)

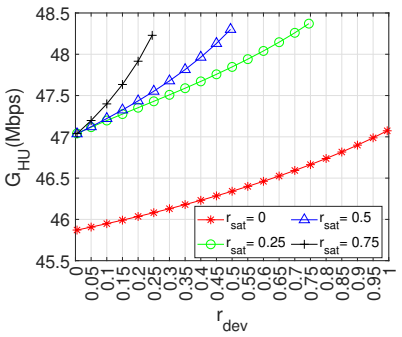


(a) EE results

(b) Goodput results



(c) Analytical EE results



(d) Analytical goodput results

Figure 10: Results for varying r_{dev} values where r_{sat} is fixed ($r_{dev} = 1 - r_{sat} - r_{BS}$, a: analytical, s: simulation).

all mode weights are equal (r_x=1/3, $x \in \{sat, BS, dev\}$) (*ii*) only D2D mode is on (r_{dev}=1) (*iii*) the satellite is off while BS and D2D are on with equal weights (r_{sat}=0, r_{BS}=1/2, r_{dev}=1/2). As shown in Fig. 9, we do not observe a significant change in EPB_{HU} with increasing λ_{PU}^{ter} in all constellations. Compared to other two constellations (*constellation-i (c-i)* and *c-iii*), for any λ_{PU}^{ter} value EPB_{HU} is lower in the *c-ii* where only D2D mode is on. This means EE is better for "only D2D mode on" scenario. However, as depicted in Fig. 9, *c-ii* has the lowest G_{HU} among three constellations for any λ_{PU}^{ter}. In all constellations, G_{HU} value decreases with increased λ_{PU}^{ter}. In *c-ii* and *c-iii*, with increased λ_{PU}^{ter} the probability of HU requests that are interrupted by PUs and that cannot continue retrieval from another idle terrestrial frequency increases. Moreover, the probability of HU requests that cannot be served upon their arrival due to the terrestrial channel being occupied by PUs and/or HUs increases. Thus, the overall network goodput decreases. In *c-i*, the service durations in the satellite link are longer and the satellite link gets saturated rapidly as observed in [6]. Thereof, the probability of finding the satellite link idle is low and the increase in the arrival rate of PUs to the terrestrial link λ_{PU}^{ter} decreases the network goodput G_{HU}.

After inspecting G_{HU} with increasing λ_{PU}^{ter} for all three constellations, we examine for any fixed λ_{PU}^{ter} how these constellations differ. In that case, *c-ii* constellation has the lowest G_{HU} value while *c-i* has the highest. The *c-ii* cannot take advantage of relatively large satellite and BS caches and this reduces the overall network goodput G_{HU}. On the contrary, *c-i* allows all system units to be used and the system can take advantage of caches of the satellite, BS and HU devices within some proximity of r-HUs. Besides, compared to *c-ii* and *c-iii* both the satellite and terrestrial links can be utilized for HU services in *c-i*. Thus, it attains highest G_{HU} value among all three constellations for any fixed λ_{PU}^{ter}. From *c-ii* to *c-iii* BS mode is activated, while from *c-iii* to *c-i* satellite mode is activated. Note that for any fixed λ_{PU}^{ter}, as satellite link saturates rapidly [6], with the activation of satellite mode from *c-iii* to *c-i* less improvement in G_{HU} is observed compared to the activation of BS mode from *c-ii* to *c-iii*. The simulation results follow the same trend with the analytical EPB_{HU} and G_{HU} results for all scenarios.

7.3. Impact of Mode Selection

For investigating the benefit of a heterogeneous architecture, it is crucial to inspect how different operation modes manifest themselves. This effort provides the initial ground to devise resource allocation schemes, which basically reveal themselves as which network mode (or link) is utilized for which device leading to efficient content delivery. We consider a setup where $N_{f_{sat}}$=2 and $N_{f_{ter}}$=3 with the universal source on and overlaying in D2D enabled. We examine several mode weight configurations and discuss how they affect EPB_{HU} and G_{HU} performance. Overall, the simulation results are consistent with the analytical EPB_{HU} and G_{HU} results. For each fixed $r_{sat} \in \{0, 0.25, 0.5, 0.75\}$, we inspect the change in EPB_{HU} (Fig. 10a) and G_{HU} (Fig. 10b) with respect to D2D mode weight r_{dev} ($r_{dev} = 1 - r_{sat} - r_{BS}$).

In $r_{sat}\in\{0, 0.25, 0.5, 0.75\}$ configurations, when D2D mode is off (r_{dev}=0) and BS mode is on, EPB_{HU} is high meaning poor EE performance (e.g. for r_{sat}=0.25, r_{BS}= 0.75, r_{dev}=0, EPB_{HU} attains 0.35 μJpb analytically.) as given in Fig. 10a. Besides, G_{HU} is low (e.g. for r_{sat}=0.25, r_{BS}=0.75, r_{dev}=0 G_{HU} is 26.7 Mbps analytically.) as shown in Fig. 10b. For the same r_{sat} configurations, when the BS mode is off and the D2D mode is on EPB_{HU} achieves low values (e.g. for r_{sat}=0.25, r_{BS}=0, r_{dev}=0.75 EPB_{HU} attains 0.003 μJpb analytically) which is EE favorable. Compared to the previous cases where D2D mode is off and the BS mode is on, "BS mode off - D2D mode on" scenarios are better in terms of G_{HU} values (e.g. for r_{sat}=0.25, r_{BS}=0, r_{dev}=0.75 G_{HU} attains 44.2 Mbps analytically). However, the overall system goodput attains even larger values for "both BS and D2D modes are on" scenarios as shown in Fig. 10b.

We also inspect more closely the network characteristics for "both BS and D2D modes on" case in terms of analytical EPB_{HU} and G_{HU} for $r_{sat} \in \{0, 0.25, 0.5, 0.75\}$ configurations. First, we inspect the EE performance. As shown in Fig. 10c, for any fixed D2D mode weight, EPB_{HU} increases with decreasing r_{sat} (e.g. for r_{dev}= 0.2 when r_{sat} decreases from 0.75 to 0, EPB_{HU} increases from 0.177 μJpb to 0.204 μJpb.). This is due to the increase in BS usage for smaller r_{sat}. The BS mode transmissions are costly in terms of energy leading to that degradation in EE. When we examine Fig. 10c again, for fixed $r_{sat}\in\{0, 0.25, 0.5, 0.75\}$ values, EPB_{HU} decrease (an improvement in EE) is observed with increased r_{dev} and simultaneously decreased r_{BS}. This observation is natural as HU devices consume less energy compared to BS for the transmission of the same content both due to lower power levels ($P^{tx}_{dev} < P^{ch}_{BS}$) and shorter service durations.

Next, we investigate the system goodput results. For some fixed r_{dev}, the utilization of the satellite decreases with decreasing r_{sat} and thus the advantage of large satellite cache is less exploited. That leads to decrease in the overall system goodput G_{HU} as depicted in Fig. 10d. An evident decrease in G_{HU} is noticed when the satellite mode is completely deactivated since the satellite cache and link are not utilized at all. For any fixed $r_{sat} \in \{0, 0.25, 0.5, 0.75\}$ configuration, an improvement in G_{HU} is monitored with increasing r_{dev} in Fig. 10d. The D2D services to HUs capture short amount of time. Thus, new HU requests can find the D2D terrestrial frequency in idle state with a greater probability. This way, we observe an improvement in the overall system goodput. However, HU devices have small cache capacities. Due to this limitation, finding a requested content is not always possible and the improvement in overall system goodput is bounded.

8. Related Work

We discuss some relevant studies for the caching and RA problems regarding satellite and terrestrial HetNets, D2D paradigm and CR techniques. They are examined concisely with the perspective how they employ these concepts (e.g. RA for CR in HetNets or caching in CR and D2D combination) and deal with energy efficiency (EE) and/or QoS aspects. The investigated metrics are revealed for each study. Our Markov model based contribution differs from the literature with its integrated portrait of the content-centric satellite and terrestrial HetNet extended by D2D and opportunistic access scheme with both EE and goodput investigation.

There is a plethora of technical works on caching in HetNets especially from EE and/or QoS perspective. In [23], EE related to content in cache-enabled D2D network is formulated and the optimal caching strategy for maximizing EE is investigated. Different from [23], our proposal focuses on the opportunistic access scheme in D2D mode. In our system, we keep the device transmission power level stable as opposed to their work. Yao et al. propose an algorithm that considers the energy-delay tradeoff by applying sleep control and power matching method for single BS scenario [24]. As distinct from their performance metric delay, we focus on the system goodput for revealing network performance. Besides, we introduced satellite into our system and our devices operate in cognitive mode for the terrestrial link.

In [25], Xu and Liu elaborate on content transmission focusing on acceptable QoS guarantee in cellular network together with D2D. They propose a caching algorithm to improve QoS by reducing overflow issue in caches and having sufficient contents cached at devices. Secondly, they come up with an RA algorithm that tries to improve EE constrained by acceptable delays. In contrast to our study, theirs does not support cognitive capabilities in devices and has no satellite extension. In [26], caching strategies for improving EE are proposed in a cellular and D2D hybrid network taking user request preferences into account. They assume that different users can have distinct preferences for same content. However, in our work a more general preference setup is used where all content preferences are distributed according to the Zipf distribution. As we have a more complex network with satellite extension, we keep the content preference simple for the sake of reduced complexity in the cache and resource management analysis.

After the investigation of caching studies, we look at the literature on the RA problem from EE and/or QoS aspects. A large body of works on RA in satellite systems exists in the literature. Brückner et al. propose a dependency-aware reservation approach for mobile satellite communications in [27]. This reservation mechanism utilizes power to signal path dependencies at the resource management phase of the satellite network. In our work, we mainly focus on distinct mode management and cognitive operation management in our complex satellite integrated D2D architecture rather than link formation as in [27]. Besides, we perform a more rigorous EE analysis for the network. Apart from satellite systems, dynamic spectrum access technology and Long-Term Evo-

lution Advanced (LTE-A) are promising paradigms for enhancing networking capabilities for heterogeneous networks. In [28], Su and Zhang propose a cross-layer medium access control (MAC) scheme over CR networks with two different sensing schemes. In their work, the trade-off between delay and throughput in a non-saturated network is demonstrated. Similar to our study, they have constructed a Markov chain to develop an analytical model. However, their model is not tuned for video consumption. It is of type $M/G^Y/1$ with a contention based access mechanism in a slotted system. On the contrary, we manage transmission completions as exponentially distributed departures. Besides, we have performed a detailed EE and goodput analysis while they studied delay and throughput. In [29], an LTE-A network consisting of cellular users and D2D users is investigated from the perspective of the RA problem. Their interference management scheme is different than ours. In ours, the interference management changes according to operation modes rather than a prespecified target interference level. For satellite and BS modes, we assume that incumbent interference management schemes mitigate any significant interference issues. Nonetheless, in D2D mode, new requests are reactively checked if they cause interference to active D2D transmissions and/or harmed by them within allowed vicinity. Thus, overlaying at the same frequency is allowed. Besides, we develop a more advanced EE model.

D2D is a powerful networking technique for improving EE and capacity in HetNets. Initially, we look at the technical works inspecting the EE objective. In [30], Xu et. al. propose a contract-based approach to select the devices willing to transmit in D2D mode and the rewards given to these devices by the BS while keeping the BS payoff small. Then, random and optimal matching algorithms are applied to establish D2D links with the objective of energy reduction. In contrast to theirs, we specifically focus on the content-based services in the HetNets. In general, D2D communications is not solely used for EE. It also serves for capacity expansion goal in HetNets. In [31], an optimal RA algorithm for the capacity of D2D users is proposed. First, they determine cellular D2D users based on SINR requirement of D2D pairs. Then, they optimize the transmission powers of users with Langrange Multiplier technique. In our study, we do not specifically focus on the power management for D2D links but focus on the mode selection analysis particularly. Furthermore, we integrate a satellite into the network (HetNet perspective) and our devices can access the channel opportunistically in D2D mode. Wang et. al propose and evaluate a distributed algorithm for content download based on expected available duration of contents [32]. Their decision mechanism considers connectivity, social influence between users and wait tolerance levels of users. In our system, we consider cache states (i.e. content availabilities calculated according to content popularities), channel availabilities and mode weights at the resource allocation phase.

CR technology and D2D communications can also coexist in wireless networks. In [33], the optimal power allocation in CR based D2D network is studied. The D2D mechanism operates in the spectrum in an opportunistic manner. An optimal power allocation scheme is proposed for the maximization of utility and also attaining requested service quality of devices in D2D mode while not exceeding interference limits to primary users. Similar to their construct, our devices access the channel opportunistically in D2D mode. Hence, we also utilize the cohabitation of CR and D2D in a network. Beyond this trait, we integrate a satellite into the network and tune onto the mode selection study from the content transmission perspective.

Overall, the literature lacks a complete treatment and an analytical model of content-oriented hybrid satellite and cellular networks with D2D and spectrum sharing from EE and QoS perspectives. Typically a single or a tuple of these aspects are studied. In this work, we jointly elaborate on these aspects and provide a holistic analysis.

9. Conclusions

In this paper, we model a HetNet of satellite and terrestrial elements (BS and end-user devices) with D2D and cognitive communications as a Continuous Time Markov Chain. We integrate the universal source concept and in-network caching into our content-centric system. We also enable overlaying in D2D operation. Our users operate as PUs over the satellite link and as secondary users over the terrestrial link. Although the universal source integration degrades EE for light HU request load, this degradation is negligible for increased load. On the other side, it does not affect overall system goodput negatively while allowing access to content outside the network boundaries (i.e. improves availability). Furthermore, the enabling of overlaying in D2D mode improves both EE and overall system goodput. Our users operate in SU mode in the terrestrial link. Thereof, we inspect how our system is affected by increasing PU arrival rate λ_{PU}^{ter} to the terrestrial link. In that case, EE is apparently not affected. However, the overall system goodput decreases. According to our mode selection experiments, turning D2D mode off ($r_{dev} = 0$) deteriorates EE and decreases system goodput at the same time. Generally, increasing r_{dev} improves EE for different cases. But for all r_{dev}'s, when satellite mode is inactive, the satellite cache is not utilized and overall system goodput degrades.

In a nutshell, our analytical model and mode selection investigation render the intrinsic trade-offs and interactions among system constituents in a complex HetNet with D2D and cognitive communications.

Acknowledgments

This work was supported by the Scientific and Technical Research Council of Turkey (TUBITAK) under grant number 116E245.

References

[1] I. F. Akyildiz, S. Nie, S.-C. Lin, M. Chandrasekaran, 5G roadmap: 10 key enabling technologies, Computer Networks 106 (2016) 17 – 48 (2016).

[2] G. Gur, Spectrum sharing and content-centric operation for 5G hybrid satellite networks: Prospects and challenges for space-terrestrial system integration, IEEE Vehicular Technology Magazine (2019) 0–0 (2019).

[3] F. Alagoz, G. Gur, Energy efficiency and satellite networking: A holistic overview, Proceedings of the IEEE 99 (11) (2011) 1954–1979 (Nov 2011).

[4] 3GPP, Study on scenarios and requirements for next-generation access technologies, TR 38.913, v.14.3.0 Release 14 (Mar. 2017).

[5] S. Kafiloglu, G. Gür, F. Alagöz, Modeling and analysis of content delivery over satellite integrated cognitive radio networks, in: 2016 14th Int. Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2016, pp. 1–8 (May 2016).

[6] G. Gür, S. Kafıloğlu, Layered content delivery over satellite integrated cognitive radio networks, IEEE Wireless Communications Letters 6 (3) (2017) 390–393 (June 2017).

[7] B. A. Alzahrani, M. J. Reed, J. Riihijärvi, V. G. Vassilakis, Scalability of information centric networking using mediated topology management, Journal of Network and Computer Applications 50 (2015) 126 – 133 (2015).

[8] V. Jacobson, D. K. Smetters, J. D. Thornton, M. F. Plass, N. H. Briggs, R. L. Braynard, Networking named content, in: Proceedings of the 5th International Conference on Emerging Networking Experiments and Technologies, CoNEXT '09, ACM, New York, NY, USA, 2009, pp. 1–12 (2009).

[9] L. Galluccio, G. Morabito, S. Palazzo, Caching in information-centric satellite networks, in: 2012 IEEE International Conference on Communications (ICC), 2012, pp. 3306–3310 (June 2012).

[10] Y. Kawamoto, Z. M. Fadlullah, H. Nishiyama, N. Kato, M. Toyoshima, Prospects and challenges of context-aware multimedia content delivery in cooperative satellite and terrestrial networks, IEEE Communications Magazine 52 (6) (2014) 55–61 (June 2014).

[11] T. Aman, T. Yamazato, M. Katayama, Traffic prediction scheme for resource assignment of satellite/terrestrial frequency sharing mobile communication system, in: 2009 International Workshop on Satellite and Space Communications, 2009, pp. 40–44 (Sep. 2009).

[12] C. Yang, Z. Chen, Y. Yao, B. Xia, Performance analysis of wireless heterogeneous networks with pushing and caching, in: 2015 IEEE International Conference on Communications (ICC), 2015, pp. 2190–2195 (June 2015).

[13] K. Suksomboon, S. Tarnoi, Y. Ji, M. Koibuchi, K. Fukuda, S. Abe, N. Motonori, M. Aoki, S. Urushidani, S. Yamada, Popcache: Cache more or less based on content popularity for information-centric networking, in: 38th Annual IEEE Conference on Local Computer Networks, 2013, pp. 236–243 (Oct 2013).

[14] E. B. Abdelkrim, M. A. Salahuddin, H. Elbiaze, R. Glitho, A hybrid regression model for video popularity-based cache replacement in content delivery networks, in: 2016 IEEE Global Communications Conference (GLOBECOM), 2016, pp. 1–7 (Dec 2016).

[15] W. Miao, G. Min, Y. Wu, H. Wang, J. Hu, Performance modelling and analysis of software-defined networking under bursty multimedia traffic, ACM Trans. Multimedia Comput. Commun. Appl. 12 (5s) (2016) 77:1–77:19 (Sep. 2016).

[16] T. Jiang, H. Wang, A. V. Vasilakos, QoE-driven channel allocation schemes for multimedia transmission of priority-based secondary users over cognitive radio networks, IEEE Journal on Selected Areas in Communications 30 (7) (2012) 1215–1224 (August 2012).

[17] S. S. Kafıloğlu, G. Gür, F. Alagöz, Analysis of content-oriented heterogeneous networks with D2D and cognitive communications, CoRR abs/1808.01021 (2018). arXiv:1808.01021.

[18] T. Jiang, H. Wang, Y. Zhang, Modeling channel allocation for multimedia transmission over infrastructure based cognitive radio networks, IEEE Systems Journal 5 (3) (2011) 417–426 (Sep. 2011).

[19] N. Vo, T. Q. Duong, H. Zepernick, M. Fiedler, A cross-layer optimized scheme and its application in mobile multimedia networks with QoS provision, IEEE Systems Journal 10 (2) (2016) 817–830 (2016).

[20] H. J. Kang, C. G. Kang, Mobile device-to-device (D2D) content delivery networking: A design and optimization framework, Journal of Communications and Networks 16 (5) (2014) 568–577 (Oct 2014).

[21] C. Liu, B. Natarajan, Average achievable throughput in D2D underlay networks, in: 2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2016, pp. 118–123 (April 2016).

[22] C. Güven, S. Bayhan, G. Gür, S. Eryigit, Optimal resource allocation for content delivery in D2D communications, in: 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), 2017, pp. 1–5 (Oct 2017).

[23] Y. Long, Y. Cai, D. Wu, L. Qiao, Content-related energy efficiency analysis in cache-enabled device-to-device network, in: 2016 8th International Conference on Wireless Communications Signal Processing (WCSP), 2016, pp. 1–5 (Oct 2016).

[24] H. Yao, C. Fang, C. Qiu, C. Zhao, Y. Liu, A novel energy efficiency algorithm in green mobile networks with cache, EURASIP J. on Wireless Communications and Networking 2015 (1) (2015) 139 (May 2015).

[25] Y. Xu, F. Liu, QoS provisionings for device-to-device content delivery in cellular networks, IEEE Transactions on Multimedia 19 (11) (2017) 2597–2608 (Nov 2017).

[26] M. C. Lee, A. F. Molisch, Individual preference aware caching policy design for energy-efficient wireless D2D communications, in: GLOBECOM 2017 - 2017 IEEE Global Communications Conference, 2017, pp. 1–7 (Dec 2017).

[27] M. Brückner, P. Drieß, M. Osdoba, A. Mitschele-Thiel, A dependency-aware QoS system for mobile satellite communication, in: 2016 IEEE Wireless Communications and Networking Conference, 2016, pp. 1–6 (April 2016).

[28] H. Su, X. Zhang, Cross-layer based opportunistic MAC protocols for QoS provisionings over cognitive radio wireless networks, IEEE Journal on Selected Areas in Communications 26 (1) (2008) 118–129 (Jan 2008).

[29] A. Asheralieva, Y. Miyanaga, QoS-oriented mode, spectrum, and power allocation for D2D communication underlaying LTE-A network, IEEE Transactions on Vehicular Technology 65 (12) (2016) 9787–9800 (Dec 2016).

[30] L. Xu, C. Jiang, Y. Shen, T. Q. S. Quek, Z. Han, Y. Ren, Energy efficient D2D communications: A perspective of mechanism design, IEEE Transactions on Wireless Communications 15 (11) (2016) 7272–7285 (2016).

[31] B. Yu, Q. Zhu, A QoS-based resource allocation algorithm for D2D communication underlaying cellular networks, in: 2016 Sixth International Conference on Information Science and Technology (ICIST), 2016, pp. 406–410 (May 2016).

[32] Z. Wang, H. Shah-Mansouri, V. W. S. Wong, How to download more data from neighbors?A metric for D2D data offloading opportunity, IEEE Transactions on Mobile Computing 16 (6) (2017) 1658–1675 (2017).

[33] H. Yao, T. Huang, C. Zhao, X. Kang, Z. Liu, Optimal power allocation in cognitive radio based machine-to-machine network, EURASIP J. on Wireless Communications and Networking 2014 (1) (2014) 82 (2014).