CONTENT-CENTRIC AND SPECTRUM SHARING HETEROGENEOUS WIRELESS NETWORKS

by

Sebahat Sinem Kafıloğlu

B.S., Computer Engineering, Boğaziçi University, 2013

M.S., Computer Engineering, Boğaziçi University, 2016

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Graduate Program in
Boğaziçi University
2021

ACKNOWLEDGEMENTS

I would like to thank my thesis supervisor Prof. Fatih Alagöz for many enlightening suggestions and guidance during the development of this thesis, and helpful comments on the thesis text. Without his good will, help, and support, it would not be possible to finish this thesis.

I am grateful to all members of my thesis jury, Prof. Tuna Tuğcu, Assoc. Prof. Berk Canberk, Prof. Cem Ersoy and Prof. Sema Oktuğ for their insightful and constructive comments to improve my thesis.

I would like to offer my deepest thanks to Dr. Gürkan Gür for his mentoring, friendship, kind help, and patience.

Last, I would like to thank my family members for their patience and profound support through all the phases of my education.

This thesis was supported by the Scientific and Technical Research Council of Turkey (TUBITAK) under grant numbers 116E245 and 117E165.

ABSTRACT

CONTENT-CENTRIC AND SPECTRUM SHARING HETEROGENEOUS WIRELESS NETWORKS

The surging multimedia content demand has put wireless network systems under heavy energy consumption burden in an unprecedented manner. Besides, wireless networks are urged to the limits of service capacity due to the adopted enriched multimedia services. According to these facts, in this thesis we analyze content-centric wireless networks from trade-off natured energy and service capacity perspectives and specifically elaborate on the spectrum sharing heterogeneous networking paradigm. First, we provide a compound spectrum sharing heterogeneous network model enriched with satellite networking, cognitive radio and device-to-device paradigms and investigate this model rigorously in all aspects. Owing to the fact that multimedia content is the building block of our research, we propose a new content model. We propose caching methods based on our new content model due to the fact that in-network caching is an instrumental method to improve the performance of content transmissions in terms of both system capacity and energy consumption. We enhance our caching evaluation in two direction i) cooperation, ii) optimization. We profoundly analyze all caching verticals with respect to different system dynamics and compare our proposals to alternative techniques.

ÖZET

İÇERİK MERKEZLİ VE SPEKTRUM PAYLAŞAN HETEROJEN KABLOSUZ AĞLAR

Artan çoğul ortam içerik talebi, kablosuz ağ sistemlerini daha önce benzeri görülmemiş bir biçimde ağır bir enerji tüketim yükü altında bırakır. Ayrıca, kablosuz ağlar, kullanılan zengin çoğul ortam hizmetleri nedeniyle servis kapasitesi sınırlarına zorlanırlar. Tüm bu gerçeklere göre bu tezde içerik merkezli kablosuz ağları, dengeleme özellikli enerji ve servis kapasitesi perspektiflerinden analiz etmekte ve özellikle spektrum paylaşımlı heterojen ağ yaklaşımı üzerine ayrıntılı olarak inceleme yapmaktayız. İlk olarak uydu ağı, bilişsel radyo ve cihazdan cihaza iletim yaklaşımları ile zenginleştirilmiş bileşik bir spektrum paylaşımlı heterojen ağ modeli sağlamakta ve bu modeli tüm yönleriyle titizlikle incelemekteyiz. Çoğul ortam içeriğinin araştırmamızın yapıtaşı olması sebebiyle, yeni bir içerik modeli önermekteyiz. Ağ içi önbellekleme hem sistem kapasitesi hem de enerji tüketimi açısından içerik aktarımlarının performansını geliştirmede etkili bir yöntem olduğundan, yeni içerik modelimize dayanan önbellekleme yöntemleri önermekteyiz. Önbellek değerlendirmemizi i) işbirliği, ii) en iyileme yönlerinde genişletmekteyiz. Tüm önbellekleme çalışmalarımızı farklı sistem dinamikleri açısından derinlemesine analiz etmekte ve önerdiğimiz teknikleri alternatifleri ile karşılaştırmaktayız.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$Avg(c)$	The average frame size difference of content c
ACT_{ALL}	The mean of random assignments with all modes active
$\overrightarrow{b_i}$	The basis vector
B	The terrestrial channel operation bandwidth
$\mathbb{B}_c$	B frame set of content c
c	The speed of the light
c_i	The i^{th} content in the content set
c_k	The content unit with lowest $E_{all}^{(u)}$ value
c_x	The newly requested content unit
C	The cache capacity of a device
$\mathbb{C}$	The local device cache set
C_{BS}	The expected service capacity of content unit retrievals from the BS cache
C_{BS}^{cache}	The BS cache capacity
$C_{BS}(n)$	The channel capacity between the BS and the n^{th} device
$C_{BS(U)}$	The expected service capacity of content unit retrievals from the universal source to the BS cache
C_{D2D}	The D2D channel capacity for service
$C_{D2D}(n,m)$	The D2D channel capacity between the n^{th} and m^{th} devices
C_{Dev}^{cache}	The device cache capacity
C_{HU}^{BS}	The average channel capacity for HUs between the BS and the requester
$C_{HU}^{BS(u)}$	The average channel capacity for HUs between the BS and the universal source
C_{HU}^{D}	The average channel capacity for HUs in D2D mode
C_{HU}^{sat}	The average channel capacity for HUs between the satellite and the requester
$C_{HU}^{sat(u)}$	The average channel capacity for HUs between the satellite and the universal source
C_{loc}	The local service capacity for a content unit

C_{PU}^{ter}	The average channel capacity over the terrestrial link for PUs
C_{Sat}^{cache}	The satellite cache capacity
$Comp$	The set of requests for a content where all the base chunks are transmitted successfully (service completed successfully)
$Contents$	The content set
CAP	The assignment according to link capacities
CAP_{EX}	The extended version of CAP with D2D mechanism
d	The distance between receiver and transmitter device
d_0	The reference distance of device antenna
d_{BS}	The average distance of a PU and/or HU to the BS
d_{D2D}	The average distance between receiver and sender HUs in D2D mode
d_{sat}	The distance from LEO satellite to the earth
$d_{t:r(BS)}$	The distance between the transmitter and receiver in the BS mode
$d_{t:r(D2D)}$	The distance between the transmitter and receiver in the D2D mode
D_{max}	The maximum number of concurrent D2D operations allowed by the network
E_{all}	The total energy consumption of the system
$E_{all}^{(u)}$	The prospective energy consumption of content unit u
E_{BS}	The total transmission energy of the BS for direct services in the system
$E_{BS}^{(u)}$	The expected BS transmission energy consumption for direct transmission of content unit u
$E_{BS(U)}$	The total reception and transmission energy of the BS for services from the universal across the BS cache to requesters
$E_{BS(U)}^{(u)}$	The expected energy consumption for transmission from universal source across the BS of content unit u
E_{block}	The total blocking energy consumption in the system
$E_{block}(s_u)$	The activation energy of a device from the sleeping to the idling state for requesting a content unit u
E_{cum}	The cumulative prospective energy consumption of in-cache content units

E_{D2D}	The total transmission energy of D2D operating devices in the system
$E_{D2D}^{(u)}$	The expected D2D transmission energy consumption of content unit u
E_{loc}	The total energy consumed for local content unit hits in the system
$E_{loc}^{(u)}$	The expected local hit energy consumption of content unit u
E_{total}	The total energy consumption of the system
EE	The energy efficiency
EPB_{HU}	The total energy consumed per successfully transmitted bits
EQ_{ALL}	The assignment where all modes are equal
f	The frequency of the terrestrial link
f_{sat}	The frequency of the satellite link
f_{ter}	Th frequency of the terrestrial link
F	The total number of frames of a content
F_c	The total number of frames of a content c
$Fail$	The set of requests for content units that have failed
$\mathbb{F}_{BS}^{fit}$	The normalization function that outputs BS storage capability
$\mathbb{F}_{loc}^{fit}$	The normalization function that outputs device storage capability
G_{all}	The total network goodput
G_{D2D}	The total D2D goodput provided by the network
G_{loc}	The total local service rate provided in the network
G_{HU}	The total system goodput of HUs
G_{HU}^{local}	The local service rate of HUs
$G_{HU}^{threshold}$	The minimum total system goodput required by the network
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
$i_{HU}^{D(f_1)}$	The number of concurrent D2D HU transmissions used for content retrieval via terrestrial frequency f_1

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
$i_{PU}^{ter(f_1)}$	The indicator for terrestrial frequency f_1 if it is used by PU or not
$idle_s(x)$	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
$\mathbb{I}_c$	I frame set of content c
k	The Weibull distribution shape parameter
LQ	The link quality based assignment
LQ_{EX}	The extended version of LQ with the satellite
m_i^x	The mesh point of r_x with the addition of pattern vector p_i
n_{BS}	The path loss exponent of BS transmission
n_{D2D}	The path loss exponent of D2D transmission
N	The total number of contents
N_0	The noise power density
N_c	The number of contents in a device cache
N_{ch}^{max}	The maximum number of chunks of a content
N_D	The total number of devices located in the cell
$N_{f_{sat}}$	The total number of satellite frequencies
$N_{f_{ter}}$	The total number of terrestrial frequencies
N_{frame}	The total number of frames
N_l	The number of content layers
N_{ngh}	The number of neighbouring devices at most R_{D2D} away from a requester
N_α	The number of random initial mode weight vector points for PSA in only one mode off scenarios and in all modes on scenarios

$p_{BS}^{(u)}$	The availability probability in the BS cache for content unit u
$p_{c_i}^{BS}$	The steady state content c_i availability probability in the BS cache
$p_{c_i}^{lo}$	The steady state content c_i availability probability in the local device cache
$p_{c_i}^{sat}$	The steady state content c_i availability probability in the satellite cache
$p_{c_i}(s)$	The request probability for based on the Zipf distribution
$p_{c_i}^{D(f_1)}(x)$	The D2D content c_i availability probability of channel state x
$p_{D2D}^{(u)}$	The availability probability in at least one neighbour device for content unit u
p_{drop}^{BS}	The dropping probability of HUs in BS mode
p_{drop}^{D2D}	The dropping probability of HUs in D2D mode
p_{HQ}	The ratio of high quality consumers
p_i	The probability of content i being requested
$\overrightarrow{p_i}$	The pattern vector in the direction of basis vector b_i
p_j	The probability of content chunk j being requested
p_k	The probability of content layer k being requested
$p_{loc}^{(u)}$	The availability probability in the local cache for content unit u
p_{local}	The probability of an HU getting service from its local cache
$p_{(c_i)}$	The probability of local device cache storing only content c_i
$p_{(c_i\ c_j)}$	The probability of local device cache storing two distinct contents c_i and c_j
$p_{(empty)}$	The probability of local device cache storing no content
P	The pattern vector set
P_{all}	The total consumed power in the system
P_{BS}^{ch}	The transmission power of the BS per channel
P_{BS}^{tx}	The transmission power consumption of the BS
$P_{BS(u)}^{rec+tx}$	The BS power consumption for BS mode (from universal) HU services either completed or dropped
P_{BS}^{rec}	The reception power consumption of the BS

P_{dev}^{tx}	The transmission power of an HU device
P_{D2D}^{rec}	The reception power consumption of a device in D2D mode
P_{D2D}^{tx}	The transmission power consumption of a device in D2D mode
P_{loc}	The power consumption of local services for HUs
P_{loc}^{u}	The power consumption of a local content unit retrieval
$P_r(d)$	The received power of a device that is distance d away from the transmitter
P_{sat}^{ch}	The transmission power of the satellite per channel
$\mathbb{P}_c$	P frame set of content c
$\mathbb{P}^{c_i}_{(x_1,x_2)}$	The probability of content c_i available at system units x_1 and x_2
r_{BS}	The weight of the BS mode
r_{dev}	The weight of the D2D mode
r_i	The selection rate of content c_i for eviction from the cache
r_{PSA}	The assignment of mode weight vector by PSA
r_{sat}	The weight of the satellite mode
r_x	An instance of the assignment mode weight vector sequence
r	The assignment mode weight vector
req_u	The request for the content unit u
R_{BS}	The radius of the BS cell
$R_{BS}(x)$	The aggregate BS mode weight function
$R_c(B, E)$	The ratio of base layer size over enhancement of content c
R_{D2D}	The radius of the total reception range in D2D mode
$R_{D2D}^{\not I}$	The radius of interference free D2D transmission zone
$R_{D2D}(x)$	The aggregate D2D mode weight function
R_{Int}	The radius of an HU device transmission range that causes interference to active HU receivers at the terrestrial frequency f_1.
R_n	The radius of the circle from requester to the outer shell of the $n+1^{th}$ area portion
$R_{sat}(x)$	The aggregate satellite mode weight function
R_{mesh}^{x}	The mesh points set for the sequence vector r_x

R_z	The radius of the investigation zone
$\mathbb{R}_{D2D}$	The total reception area
$\mathbb{R}_n$	The $n+1^{th}$ reception area
s_b	The average base layer size
s_{ch}	The average chunk partition size
s_e	The average enhancement layer size
s_f	The size of the frame f
s_{in}	The initial mesh size
s_{mesh}	The mesh size
s_u	The size of the content unit u
$\mathrm{s}_{(x)}$	A channel state definition
$s(\hat{v_b})$	The mean content size
$\overline{s_{HQ}}$	The average size of HQ videos
$\overline{s_{SQ}}$	The average size of SQ videos
S_c	The set of content units in the cache
S_U	The set of content units identifiable by content, chunk and layer id uniquely
S_U^D	The set of content units in the requester device cache
S_{sort}	The content units sorted in descending order on $E_{all}^{(u)}$ values
$S_{(n)}^{BS}$	The set of services from the BS to n^{th} device
$S_{(n)}^{BS(U)}$	The set of services from the universal content repository to the n^{th} device across the BS
$S_{(n,m)}$	The set of services from the n^{th} device to m^{th} one
$\overline{S_c}$	The set of content units decided to remain in cache
T_{const}	The tolerance on constraint function G_{HU}
T_d	The BS to D2D mode offloading factor
T_{mesh}	The tolerance on mesh size
T_{obj}	The tolerance on objective function EPB_{HU}
T_{sim}	The simulation duration
Th_{HU}^{BS}	The BS mode (direct) HU throughput
$Th_{HU}^{BS(u)}$	The BS mode (from universal) HU throughput
Th_{HU}^{D}	The D2D mode HU throughput

Th_{HU}^{sat}	The satellite mode (direct) HU throughput
$Th_{HU}^{sat(u)}$	The satellite mode (from universal) HU throughput
u	The unique *content unit identifier* such that each $\{i, j, k\} \mapsto u$
U	The sum of the probabilities of being at channel states causing forcible BS mode HU drops
$Var(c)$	The variance of the frame size differences for content c
W	The residual cache capacity after the new content unit is inserted
W_{sat}	The bandwidth of the satellite link
W_{ter}	The bandwidth of the terrestrial link
x_i	The utility function of content c_i for eviction from the cache
X_n	The size of the n^{th} frame
X_n^b	The size of the n^{th} frame as a base layer
X_n^e	The size of the n^{th} frame as an enhancement layer
$\overline{X^b}$	The mean video frame size of the base layer
$\overline{X^e}$	The mean video frame size of the enhancement layer
α	The Zipf distribution parameter
α_c	The contraction factor for mesh size
$\alpha_{density}$	The interplay multiplier for changing device density in the cell
α_e	The expansion factor for mesh size
α_{enh}	The interplay multiplier for changing average enhancement layer size
α_P	The maximum priority class identifier in the D2D reception range (starts from 0)
β	The number of dedicated terrestrial frequencies for D2D mode
δ_{BS}	The incremental units for the BS cache
δ_{Dev}	The incremental units for the device cache
$\Delta_{HU}^{BS(u)}$	The average total service duration from the universal source across the BS to requester
$\Delta_{HU}^{sat(u)}$	The average total service duration from the universal source across the satellite to requester
γ_{BS}^{i}	A random number in (0,1) range

$\gamma_{HU}^{\theta}(i,\{x_1,x_2\})$	The HU transition rate for the retrieval of content c_i when it is available at system units x_1 and x_2 and retrieved in θ mode
γ_{sat}^{i}	A random number in (0,1) range
Γ_{HU}^{θ}	The expected arrival rate of θ mode HUs into the network
$\Gamma_{HU}^{\theta}(i)$	The total transition rate of θ mode HU service request for the retrieval of content c_i
λ	The Weibull distribution scale parameter
$\lambda_{c(size)}$	The content size distribution parameter
$\lambda_{eff(HU)}^{BS}$	The BS mode (direct) HU effective arrival rate
$\lambda_{eff(HU)}^{BS(u)}$	The BS mode (from universal) HU effective arrival rate
$\lambda_{eff(HU)}^{D2D}$	The D2D mode HU effective arrival rate
$\lambda_{eff(HU)}^{sat}$	The satellite mode (direct) HU effective arrival rate
$\lambda_{eff(HU)}^{sat(u)}$	The satellite mode (from universal) HU effective arrival rate
λ_{HU}	The mean content request rate of HUs
$\lambda_{HU}^{c_i}$	The mean request rate of HUs for content c_i
$\lambda_{N_{HU}}$	The mean density of HU devices located in the BS cell
λ_{PU}^{ter}	The average arrival rate of PUs at terrestrial link
λ_u	The content request rate of user devices
λ_{users}	The mean density of users located in a cell according to Poisson Point Process
μ_{HU}^{BS}	The average service rate of HUs in direct BS mode
$\mu_{HU}^{BS(u)}$	The average service rate of HUs from the universal source across the BS to requester
μ_{HU}^{D}	The average service rate of HUs in D2D mode
μ_{HU}^{sat}	The average service rate of HUs in direct satellite mode
$\mu_{HU}^{sat(u)}$	The average service rate of HUs from the universal source across the satellite to requester
μ_{PU}^{ter}	The average service rate of PUs over the terrestrial link
π_x	The steady state probability of being at channel state x
Π_{c_i}	The probability of the content c_i being retrievable over the terrestrial frequency f_1.
$\Pi_{rec}(x)$	The probability of the receiver HU not being interfered by other D2D operations

$\Pi_{tx}(x)$	The probability of the transmitter HU not causing interference to active D2D operations.
ρ	The Pearson correlation coefficient
θ	A random number selected uniformly in [0,1]
θ_{BS}	The parameter per channel BS reception power consumption (Per channel reception power of the BS is $\frac{P_{BS}^{ch}}{\theta_{BS}}$)
θ_{loc}	The parameter for device local hit power consumption (The local hit power consumption of a device is $\frac{P_{dev}^{tx}}{\theta_{loc}}$)
$\{i, j, k\}$	The unique *content unit tuple* (The i^{th} content's chunk j of layer k)

LIST OF ACRONYMS/ABBREVIATIONS

5G	Fifth Generation
6G	Sixth Generation
ACT	Active
AI	Artificial Intelligence
AP	Access Point
AR	Augmented Reality
AWGN	Additive White Gaussian Noise
BS	Base Station
CAP	Capacity
CC	Chunk based Caching
CCN	Content-Centric Network
$COOP_A$	Availability based Cooperation
$COOP_D$	Distance based Cooperation
CPPC	Chunk Prioritized Popularity Based Caching
$CPPC(E)$	Threshold based Energy Management Integrated Chunk Prioritized Popularity Based Caching
CR	Cognitive Radio
CTMC	Continuous Time Markov Chain
D2D	Device-to-Device
DCT	Discrete Cosine Transform
EE	Energy Efficiency
$EPDC$	Energy Prioritized D2D Caching
EQ	Equal
FIFO	First In First Out
f-RAN	Fog Radio Access Network
GoP	Group of Picture
HetNet	Heterogeneous Network
HQ	High Quality
HU	Hybrid User

IoT	Internet of Things
ITU	International Telecommunication Union
LEO	Low-earth Orbit
LPPC	Layer Prioritized Popularity Based Caching
LQ	Link Quality
$LPPC(E)$	Threshold based Energy Management Integrated Layer Prioritized Popularity Based Caching
LRU	Least Recently Used
LTE	Long-Term Evolution
LTE-A	Long-Term Evolution Advanced
M2M	Machine-to-Machine
MAC	Medium Access Control
MC	Markov Chain
MIN-ACC	Minimum-Access
NLP	Non-linear Programming
NP	Non-deterministic Polynomial
OPT	Optimal
P2P	Peer-to-Peer
PDC	Popularity-Driven Caching
PPP	Poisson Point Process
Prob	Probability Caching
PSA	Pattern Search Algorithm
PU	Primary User
QCIF	Quarter Common Intermediate Format
QoE	Quality of Experience
QoS	Quality of Service
RA	Resource Allocation
RBC	Rate-based Control
rec-HU	Receiver Hybrid User
req-HU	Requester Hybrid User
RP	Random Point

SAT	Satellite
SINR	Signal-to-Interference-Plus-Noise Ratio
SQ	Standard Quality
SQM	Sum Queue Minimization
SU	Secondary User
SVC	Scalable Video Coding
SXO	Size*Order
tx-HU	Transmitter Hybrid User
umMTC	Ultra-massive Machine Type Communications
VR	Virtual Reality

1. INTRODUCTION

The multimedia traffic in wireless networks is propelled by the exploding content centric services that are becoming pervasively utilized. According to the CISCO forecast report, the machine-to-machine (M2M) connections consisting of video surveillance, healthcare monitoring, smart meters, etc. will reach to 14.7 billion connections in 2023 while smartphones and connected TVs are also growing at a large pace in terms of the number of connections [7]. Video streaming services in platforms such as Youtube, Netflix are highly popular recently. As of May 2019, Youtube had 2 billion logged-in monthly users and in the second quarter of 2020, Netflix had around 193 million paying global subscribers respectively [8, 9]. Furthermore, due to the Covid-19 pandemic, many countries announced lock-downs for several weeks and in these periods the major communication option was the live chat applications that further amplified the multimedia request on wireless networks.

Due to the increased demand on the multimedia, resources in wireless networks are put under a heavy burden and to solve this problem, spectrum sharing techniques are handy. In that regard, device-to-device (D2D) communication technique draws significant interest [10,11]. In this technique, devices communicate not as conventional cellular users across the base station (BS) but form direct links to transmit content from one device to the other. A frequency can be utilized by the cellular mode and also with meeting necessary interference requirements, it can be utilized by several D2D links simultaneously as well and thereby, the spectrum resource efficiency and capacity boosting are enabled. The reduced delay and transmission power consumption of D2D communications [12] are driving forces to make use of this technique. Cognitive Radio (CR) is also a broadly utilized technique to boost the spectral efficiency [13]. This is achieved by allowing cognitive radio users to operate at idle spectrum bands without harming licensed user activities [14]. CR paradigm serves for improving the service capacity [15] and throughput [16] as well. However, the additional energy for the sensing and channel switching in CR approach needs to be rigorously handled for meeting network-wise energy efficiency in such networks [17]. In this context, the power

minimization in CR networks is broadly studied [18, 19].

Fifth Generation (5G) standardization is employed in a vast majority of wireless network systems such as smart cities, autonomous cars, connected health, video streaming and more. It is an actuator of multimedia services that need to realize stringent Quality-of-Service (QoS) requirements and in that regard 5G communications require reaching high service rates, bandwidth, coverage and they also need to meet reduced latency and energy [20]. For instance, 5G enabled Internet of Things (IoT) has requirements to reach 25 Mbps data rates and support billions of low-power devices [21]. According to the CISCO Forecast Report [22], globally almost 80% of the mobile data traffic will be multimedia by 2022. The requirements of future multimedia applications such as 8K and virtual reality (VR) streaming, and cloud gaming will have intense burden in 5G networks and accordingly different architectural components are realized and investigated in 5G systems. Heterogeneous networks (HetNets) is one these components utilized by 5G paradigm. They contain cell types of different sizes and deliver increased service capacity for 5G systems [20]. Furthermore, energy efficiency is actualized in D2D-Enabled HetNets [23]. Satellite integration is an another dimension for boosting service capacity [24] in 5G communication networks. Besides, satellite terrestrial 5G systems provide 80% improvement in terms of energy efficiency in ultra-dense scenarios [25]. However, 5G paradigm is not the final destination for enrichment in multimedia services.

In the future, sixth Generation (6G) standard will arise with the envisioned usage scenarios enabling super-high-definition enriched multimedia services and providing ultra-low latency [26]. In [27], it is envisioned that 6G will reach 100-1000 times higher data rates than 5G systems. Moreover, 10-100 times greater EE for 6G over 5G is another expected requirement [26, 28]. Another usage scenario for the 6G is the ultra-massive machine type communications (umMTC) [26]. Dense D2D deployment is a variant of this. Accordingly, enriched multimedia systems meeting stringent service requirements are expected to be realized in the future. 6G is envisioned to develop on highly dynamic heterogeneous network systems with one of the key components umMTC. The usage of UAV systems both in civil and military applications will rise

and hence such systems will further increase the heterogeneity in the next-generation networks. Moreover, the deployment of mega sky constellations consisting of more than 10000 Low-earth-orbit (LEO) satellites elevates the heterogeneity of envisioned 6G systems. The Starlink project of the SpaceX is planned as a constellation of 12000 LEOs. In [29], it has been shown that Starlink like constellations achieve decreased latency over the fiber optic by using the ground stations as relays. In that regard, such constellations have promising results. SpaceX filed to the International Telecommunication Union (ITU) for 30000 more LEOs as well [30]. The manufacturing speed for Starlink satellites is also considerably large with 120 satellites per month [31]. Amazon also has a satellite constellation project with 3236 satellites that is filed with ITU as Kuiper systems [32]. Hongyan and GW are other broadband LEO constellation projects. In terms of the deployement these projects will continue up until 2030 when the 6G standardization will be completed. As of 2030, 6G standard requirements will be actualized by such mega satellite constellations, UAV systems and dense D2D deployed environments.

The aforementioned research directions (CR, D2D, satellite, HetNets, etc.) in wireless networks for multimedia based operations are arched over two trade-off objectives: *i)* service capacity and *ii)* energy consumption. Most important of all energy efficiency (EE) meta-objective is observed rigorously for the profound analysis of our spectrum sharing heterogeneous wireless network system. It is an instrumental objective for the wireless network analysis regarding the emergence of green networking paradigm and the reduction in the system energy cost from operational aspect.

1.1. Contributions and Thesis Outline

In this thesis work, we elaborate on the spectrum sharing heterogeneous wireless networks in terms of content consumption aspect. The tremendous demand on multimedia services results in a rigorous analysis on system, model and algorithms regarding multimedia transmission in wireless network structures. In that regard, the contributions of this thesis are listed as follows:

(i) Model for satellite integrated cognitive and D2D HetNets (Chapter 3): As a novelty, we built a compound analytical model of a satellite integrated cognitive and D2D HetNets for multimedia transmission. For the sake of completeness, D2D in overlaid form is utilized that enables service co-existence in the same frequency with controlled interference. Additionally, content fetches from repositories are realized and a baseline popularity caching technique is integrated into our model for a realistic network system view. We made a complete analysis and simulations on our model in terms of energy efficiency and goodput. Finally, we laid out various results and discussed the trade-off factors of the system. Another contribution is the formulation of an EE maximizing optimization constrained by the capacity for the mentioned HetNet in terms of mode assignment. In that regard, we developed a sub-optimal assignment technique as a novel endeavor and finally showed the corresponding EE improvement.

(ii) Content modeling and caching algorithms (Chapter 4): As an instrumental contribution joint *popularity*, *chunking* and *layering* attributes packed multi-dimensional content model is developed owing to the fact that multimedia content consumption is our main use-case. Caching is one of the facilitators for improving both system capacity and reducing energy consumption. In that regard, considering our content model we proposed dimension-prioritized cache replacement algorithms in a D2D network. For a rigorous analysis, our dimension-prioritized caching algorithms are studied in a D2D simulation environment in terms of energy consumption, system goodput and EE. In terms of caching, cooperation adds further improvement. Hence, we also propose cooperative cache replacement algorithms in a D2D network regarding the cache profiles in the neighborhood considering availability/proximity for the selection of cache replacement units as a key contribution. As another technical contribution, we investigate the video scene change dynamics on our layered content model.

(iii) Energy minimizing caching algorithms in cellular D2D networks (Chapter 5): Caching is an instrumental method for alleviating the burden on the energy depleting resources in a wireless network. Accordingly, we study the edge caching in cellular D2D networks. First, we introduce an energy consumption model for

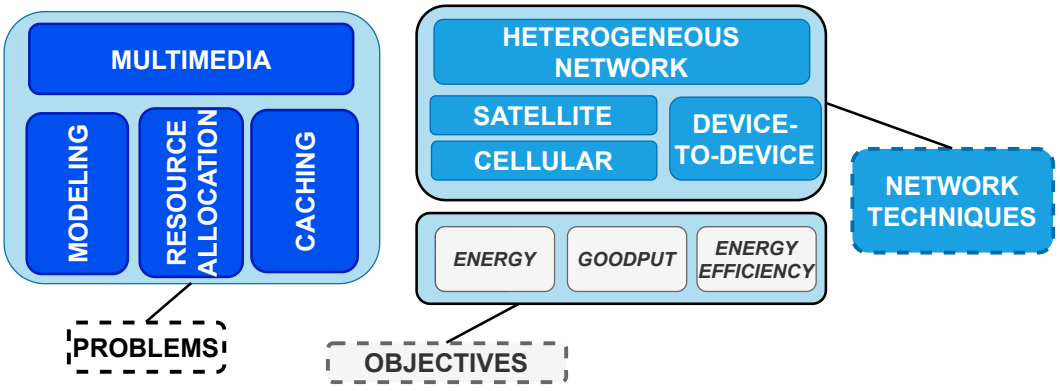

Figure 1.1. The problems (modeling, resource allocation, caching) and objectives studied, techniques integrated in thesis work.

content transmissions across a cellular D2D network considering different service-modes. As a key contribution, energy minimization is managed optimally for the cache replacement in such networks and due to the feasibility problem we propose our energy-cost regarding cache replacement algorithm, namely *Energy Prioritized D2D Caching (EPDC)*. Finally, we run optimal and energy-cost regarding *EPDC* algorithms as a simulation to evaluate different service-mode performances in terms of service rate and energy consumption.

In this thesis, we focus on the spectrum sharing heterogeneous wireless networks from multimedia transmission perspective. In Chapter 2, we present the related work to form the foundational background on the above mentioned wireless networking paradigms. In the following Chapter, we introduce and analyze our Markovian model for satellite integrated cognitive and D2D HetNets with respect to goodput and EE in a great detail and have an extension on mode assignment optimization. In Chapter 4, we propose our multimedia content model and also propose caching algorithms built upon the attributes of this model with a compound performance evaluation. Additionally, we study the cooperative caching and evaluate its performance rigorously. Chapter 5 elaborates on the energy minimization in cellular D2D networks and accordingly we solve the cache replacement problem optimally. Besides, we propose an energy-cost based heuristic algorithm and inspect several service-modes in terms of service rate and energy consumption in a great detail. Finally, we draw a conclusion in Chapter 6.

2. RELATED WORK

In this chapter, we provide the related background and work on each subject in this thesis that we focus on.

2.1. Heterogeneous Network Analysis

Heterogeneous networks (HetNets) consist of different network types (e.g. satellite, cellular, personal-area, device-to-device (D2D) networks) that co-exist with the aim of improving user experience and network efficiency. They enable flexible services and connectivity according to dynamic system needs. D2D networking technique is profoundly useful for enabling energy efficiency with proper management [33]. Satellite integration provides energy efficiency to heterogeneous systems as well. Cognitive radio (CR) provides network-wide capacity improvement by enabling opportunistic access to network [34]. Regarding all these factors, we build a model of satellite and terrestrial HetNet together with D2D and CR capabilities. The cache management and resource allocation problem are two basic blocks in our thesis. Thereby, we discuss some studies for the caching and resource allocation problems in satellite and terrestrial HetNets, D2D paradigm and CR techniques. We start with the investigation of caching studies regarding HetNets and then we focus on the resource allocation regarding different combination of the aforementioned aspects. We present the difference of these studies to our detailed HetNet analysis in Chapter 3.

There is a large body of literature studies on caching in HetNets especially from energy efficiency (EE) and/or Quality of service (QoS) perspective. Yao et al. propose a method taking into account the energy-delay tradeoff by applying sleep control and power matching method for single base station (BS) scenario [35]. They have devised a detailed sleep and active operation modes enabled power model in contrast to our study. In spite of this, in Chapter 3 we integrate satellite into our system and our devices operate opportunistically over the terrestrial link, thereby enlarging the capabilities of our system. Different from their performance metric delay, we focus on the system

goodput aspect for revealing network performance contributing to the differentiation from their study [35]. In [36], EE related to content in cache-enabled D2D network is formulated and the optimal caching strategy for maximizing EE is investigated. According to their results, the optimal caching is coupled with the transmission power of content cacher users. Different from [36], our proposal in Chapter 3 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. Besides, they do not construct a Markov chain for analyzing their cache-enabled D2D network system. Zhang et al. consider a software-defined network and propose a satellite-terrestrial HetNet [25]. The proposed scheme improves EE for sparse and ultra-dense networks compared to Long-Term Evolution (LTE) and additionally improves coverage capability. Even though we both deal with satellite-terrestrial heterogeneous networks, the approach perspectives are remarkably divergent. In [25], the EE exploration is tuned on from software defined network perspective while we analyze the EE problem with a content based Markov model.

In [37], Li et al. present a survey on caching in cache-enabled cellular networks with a taxonomy of macro-cellular, heterogeneous, D2D, cloud-radio access, and fog-radio access network architectures. They provide a broad analysis on cache placement, delivery or hybrid perspectives looking at metrics throughput, backhaul cost, power consumption, network delay and hit rate. In [38], EE boosting caching methods have been proposed in a cellular and D2D hybrid network regarding user request preferences with user collaboration and non-collaboration schemes. Based on their assumption, diverse preference profiles of different users exist for same contents. As we have a more complex network system with satellite extension, we keep the content preference simple for the sake of reduced analysis complexity in the cache and resource management. Therefore, we utilize a more general preference layout with all content preferences distributed according to the Zipf distribution. In [39], Xu and Liu focus on content transmission elaborating on acceptable QoS guarantee in cellular network together with D2D. They propose a caching mechanism with the aim of QoS improvement by reducing cache overflow and storing in devices sufficient contents. Additionally, they present their resource allocation (RA) algorithm improving EE constrained by

acceptable delays. Similar to our study, they elaborate on the content management both in terms of caching and resource allocation for D2D cellular networks with spatial reuse. However, we have extended our HetNet architecture with satellite extension and our devices have cognitive capability.

Next, we look at the plethora of work on the RA problem from EE and/or QoS aspects. A large body of works exists in satellite system resource managements. According to [40], the satellite system EE improvement opportunities basically stem from the features of the satellite such as solar power usage, adaptation of energy-efficient hardware and/or usage of energy efficient protocols to the satellite. In [41], Brückner et al. propose a dependency-aware reservation technique in mobile satellite communication systems that utilizes power to signal path dependencies at the resource management phase of the satellite network. In contrast to dependency unaware schemes, their proposal reduces waste of resource. In our work, we mainly focus on distinct mode management and cognitive operation management in our complex satellite integrated D2D architecture rather than reactive link formation as in [41]. Besides, we perform a more rigorous EE analysis for our HetNet architecture.

D2D is a powerful networking technique serving for EE and capacity expansion goal in HetNets. In [33], Xu et al. study on D2D cellular networks and propose a contract-based approach to select the D2D transmitters and corresponding rewards assigned by the BS while keeping the BS pay-off low. Next, with the aim of energy reduction random and optimal matching algorithms are applied for D2D link establishments. In our HetNet system, we elaborate on content-based services as opposed to [33]. In D2D networks resource management with clustering is another broadly utilized approach. The dispersion of the burden on the resource depletion to a variety of devices allows improved orchestration of resources and improves the energy efficiency. A cluster head rotation technique in D2D systems is proposed in [42] and is shown to be energy efficient. However, in our system we do not use clustering in D2D link management since we deal with a more compound heterogeneous network with satellite integration in Chapter 3. Furthermore, we show that the system goodput is as a key metric to demonstrate the system performance in accordance with the trade-off factor

EE. In [43], Wang et al. propose a distributed content download method based on expected available content durations. They present the superiority of their algorithm in terms of the amount of D2D offloading download. They utilize *i)* social influence between users, *ii)* user wait tolerances and *iii)* connectivity for decision strategies. Instead, we employ resource allocation strategy in our HetNet based on cache states (i.e. content availabilities calculated according to content popularities), channel availabilities and mode weights. In [44], optimal RA for the capacity of D2D users is proposed where they select cellular D2D users based on signal-to-interference-plus-noise ratio (SINR) requirement and then optimize the user transmission powers with Langrange Multiplier. In our study we focus on the mode selection analysis rather than the D2D link power management. Besides, we model a more complex HetNet with satellite integration and our devices access the channel opportunistically in D2D mode. There also exist studies presenting the trade-off factors EE and QoS in D2D networks. Schmidt et al. investigate EE in D2D cellular networks [45]. According to their results, promoting D2D decreases power consumption and still preserves aggregate throughput for large but not-fully loaded networks. Similarly, EE and goodput trade-off factors are considered in Chapter 3. Despite this similarity in comparison, we have a more compound HetNet system with solar-powered satellite extension and our main use-case is the content service in this HetNet that [45] is short of.

For improving networking capabilities for heterogeneous networks Long-Term Evolution Advanced (LTE-A) and dynamic spectrum access technology are also promising paradigms. In [46], an RA algorithm has been proposed that outperforms graph-based RA, joint RA, rate-based control (RBC) and sum queue minimization (SQM) techniques with respect to mean packet loss and end-to-end delay in LTE-A D2D cellular network. In their proposal, they define a target level with no user allowed above this interference threshold. As opposed to this, our interference management in Chapter 3 is tuned according to operation modes in our system. In particular, overlaying at the same frequency is allowed in D2D mode and therefore new requests are reactively checked against causing interference to active transmissions and/or harmed by them. Apart from the interference management difference, we develop a more advanced EE model as well. Cross-layer medium access control (MAC) scheme over CR networks

with two sensing algorithms are proposed in [47]. One of the sensing algorithms is random while the other one is negotiation-based. They have constructed a Markov chain model of type $M/G^Y/1$ with contention based access mechanism in a slotted system. On the contrary, we manage the content arrivals in a non-slotted manner as Poisson processes and transmission completions as exponentially distributed departures in Chapter 3. As opposed to our main video consumption scenario, their model is not examined accordingly. Besides, we present trade-off factors EE and goodput whereas they elaborate on delay and throughput instead.

The customizability of HetNets is driven from the component integration/ disintegration capability and mode selection mechanism. D2D paradigm is broadly employed for EE improvement [48,49] and in that regard we study the optimal connectivity mode management in heterogeneous D2D networks in Section 3.5. There are three service modes: *i) D2D*, *ii) cellular* and *iii) satellite*. The satellite service can essentially provide fifth generation (5G) services to locations under inadequate network support due to cost concerns [50]. Furthermore, the satellite has a significant caching-and-broadcasting gain [51]. In spite of these advantages, it is more prone to channel impairments like tropospheric effects, phase noise, nonlinearity [52] and has rapid capacity saturation issue.

D2D mechanism has a potential to boost network service capacity [53] and is more energy-wise rewarding. However, devices have limited cache capacity with intermittent cooperation opportunities, and to overcome this issue we can make use of the cellular counterpart with greater cache capacity. Cellular links have longer transmission range than D2D links. They are also more stable due to one non-mobile transmission end-point as in D2D services. However, BS communications suffer more transmission energy consumption in contrast to the D2D approach due to large BS power and longer service durations [1]. Regarding all such contradicting factors, we analyze the optimal *mode selection* rigorously in Section 3.5 for the sake of a complete trade-off analysis in terms of the system service capacity and EE.

In a nutshell, our Markov model based contribution diverges from the literature with its compound portrait of the content-centric satellite and terrestrial HetNet extended by D2D and opportunistic access scheme through extensive EE and goodput investigation.

2.2. Content Modeling and Caching

There is a plethora of literature that makes use of any subset of *i) popularity, ii) chunking* and *iii) layering* dimensions for the content modeling. Firstly, the popularity dimension projects the video monitoring preferences of users. In that regard, [54, 55] utilize the popularity characteristic of contents. The Zipf distribution is a broadly utilized for the popularity projection in wireless networks [56]. Secondly, the chunking dimension portraits how videos are partitioned into smaller segments. This dimension also has the capability of improving system efficiency. In [57], Hwang et al. utilize the chunking dimension for improved bandwidth utilization. Finally, the layering dimension is utilized with the aim of providing scalability to the content dissemination in networks. [58, 59] utilize the layer concept in their content models. After the literature studies that contain only one-tuple of dimensions, we will look up the studies comprising dimension pairs for their content models. When we consider the pair of popularity and chunking dimensions, [60, 61] exploit them in their content model. In [62], Ramzan et al. utilize layering and chunking dimension pair for video streaming. Besides, [63, 64] are two example literature studies that make use of content popularities and focus on caching for layered contents as well.

In Chapter 4, we build our content model making use of all three aforementioned dimensions: *popularity, chunking, layering* as a novel endeavor. Further, we integrate our *multi-dimensional* content model to the caching mechanism in a wireless D2D network. To profoundly investigate the D2D caching based on the content modeling, we first elaborate on the in-network caching studies. We branch the in-network caching survey into two main aspects: *i) content-based caching* and *ii) D2D caching*.

In the literature, there is a vast amount of study elaborating on the content-based caching that primarily utilize the content features for the cache management [64–67]. Hong et al. have devised a request score and chunk based caching (CC) scheme for the multimedia streaming [65]. In their scheme, popular contents are rewarded by greater caching rate. Additionally, the chunking dimension utilization is actualized by the fact that subsequent chunks of a monitored video chunk are requested as well. In [66], they propose a caching policy where the number of cached chunks in a content router increases exponentially by increased access rate. Zhan et al. propose a heuristic caching policy for layered content dissemination in heterogeneous networks in terms of reduced latency [64]. In [67], Suksomboon et al. propose the PopCache technique in Content-Centric Networks (CCNs) that decreases the distance between storage routers for popular contents and their requesters and distributes contents along the routers for reducing the redundancy factor. However, these caching techniques lack D2D operation domain. On the contrary, in the literature there exists a large body of D2D caching studies utilizing content features [2, 63, 68–71]. Zhan et al. propose a hybrid caching and service orchestration strategy for layered content dissemination in D2D networks by utilizing the Zipf distribution-based content popularity model [63]. In [68], the caching and resource allocation in D2D network and small cells is studied where the popularity of contents are projected by the Zipf distribution. In another work, Chen et al. optimize the caching and resource allocation jointly in terms of maximizing the offloading of services to local hits or from neighbouring devices in D2D operation mode [69]. They utilize the Zipf distributed content probabilities for the calculation of offloading probability. Ji et al. study the caching paradigm for chunked content setting in multi-hop D2D networks [70]. For the worst-case investigation, they make use of an arbitrary content request pattern. In [71], a framework of content caching is proposed for ameliorating capacity and quality of experience (QoE) with a Zipf distribution fitted content popularity and chunk-oriented content data. However, none of these studies utilizes all three dimensions: *popularity*, *chunking* and *layering* in their content model for a profound analysis. As a novel endeavor, we build a content model utilizing all these three dimensions. Besides, we propose caching techniques based on this content model by the prioritization of different dimensions in Chapter 4.

For the caching orchestration *cooperation* is an important tool in terms of performance improvement. In that regard, we investigate cooperative caching studies in the literature. Zhang et al. propose a mobility-aware cooperative edge caching mechanism [72]. According to their simulation results, they achieve improvement in terms of service latency. In [73] a cooperative caching scheme on a peer-to-peer (P2P) mobile network system is proposed regarding improvement in terms of cache hit ratio, cache replacement time and power efficiency. Ghandeharizadeh et al. propose a cooperative cache replacement mechanism considering contents also present in cooperative group devices [74]. They decide on eviction according to the order of content values where order are calculated by request frequency over size. In [75], a collaborative cache orchestration protocol is studied in D2D networks. The controller collects information from the neighborhood and decides based on the difference between caching proportion and popularity. A reception does not necessarily entail caching of a content and they primarily focus on the content placement rather than replacement. In [76], cooperative caching is proposed to cache highly requested data at nodes along the transmission route and cache the path itself as well for devices closer to caching node. They utilize order and size information of content in the cache for the cache replacement. In [73,77], cooperative caching management for multimedia services are studied as well. Overall, there is a bulk of cooperative caching studies that decide on the eviction according to several different criteria.

Due to the remarkable significance of the cooperation in terms of caching, in Chapter 4, we study the content caching in D2D networks from that aspect as well. We elaborate on the cache replacement problem and propose cooperative cache replacement techniques that differ from the literature in terms of utilized eviction criteria. We utilize the cache profile of the neighborhood in terms of content replicas in the reception range and corresponding availability/ distance as eviction criteria. Furthermore, by using our proposed chunk prioritized popularity based caching ($CPPC$) [4] algorithm for tie breaking purpose, we integrate content popularity, layering and partitioning aspects into the eviction decision mechanism as well. By the complete usage of all these mentioned aspects for the eviction process our caching proposals differ from the literature body. As another contribution, the scene change dynamics on layered content

model is analyzed rigorously in Chapter 4.

2.3. Edge Caching in Cellular D2D Networks

Edge caching in D2D networks is a broadly investigated topic regarding the content consumption. The edge caching has a useful impact on energy efficient network design [78]. It is also utilized to improve successful delivery probability [79]. The D2D paradigm also improves the network service capacity while achieving energy reduction purposes. Content consumption is our main use-case and hence we throughly look up on the literature for edge caching in D2D networks from content aspect. We portray the difference of our caching study in cellular D2D network to the existing literature in Chapter 5.

Initially, we elaborate on content-based D2D cache management studies in the literature. Chen et al. manage caching according to learned user preferences and achieve a caching gain over content popularity utilizing technique [80]. They elaborate on the offloading whereas we study the energy consumption with respect to varying service modes. In [81], they have shown the practical usability of the Zipf distribution for content modeling and the superiority of the D2D cache network over the BS unicast system. In accordance with their modeling outcome, we also utilize the Zipf distribution for the content popularity in Chapter 5. However, we elaborate on the cache replacement management in terms of energy and rather than a comparison to BS unicast, we compare our replacement proposal to conventional techniques. Li et al. propose caching in Fog Radio Access Network (f-RAN) enhanced with D2D [82]. Social preferences dynamically determine access point (AP) communities and these communities cooperatively cache. D2D mode is also utilized for further decrease in delay with the selection of the most appropriate device in each AP and largely requested contents for caching. We have a general network layout without specifically building on top of a f-RAN. Furthermore, we diverge from their study in terms of investigated metrics. They monitor cache hit and delay performance while we interrogate the network energy consumption. Above all they had cache placement for one central device only in each AP. On the contrary, our cache replacement algorithm operates in all devices in our

cellular D2D network.

We elaborate on cache management studies from the energy consumption aspect. 5G networking with tremendous demand on multimedia and requirement for low latency challenges network greening [83]. In-network caching is widely employed to overcome this energy burden. Yang et al. propose a self-optimizing in-network caching algorithm for improving energy-efficiency in 5G networks [84]. However, the D2D communication mechanism is not utilized in their study. As D2D paradigm is a major component in our system, we particularly focus on the D2D network caching. Vu et al. study coded and uncoded caching in content networks with provider integration [85]. Uncoded version achieves improved energy efficiency with small device capacity. Differently, we utilize D2D communication as a major technique in our system and then reveal the performance results in Chapter 5. Lin et al. propose a sub-optimal caching strategy regarding energy ratio to improve energy efficiency [86]. The superiority of their proposal over the equal probability random, most popular random and cut-off random strategies are demonstrated. They assume restrictive cases without universal source and energy consumption for local hits whereas we consider them in Chapter 5. Further, they define EE as the ratio of the energy in caching network over no-caching one while our definition is the total energy consumption over the successfully received data considering not only the caching but specifically regarding content transmission. Chen et al. propose a content placement method with mobility-awareness to optimize cache hit and tune transmission powers of small base stations and devices with the aim of minimizing energy cost [87]. They investigate cache placement but we rather elaborate on the replacement in our system. To sum up, there is a wide variety of D2D caching studies from energy efficiency perspective.

We elaborate on *optimization* in D2D caching network systems. Chai et al. propose a game-based partitioning for optimal cache space allocation and service delay minimizing formulation is done for content placement and user association in D2D HetNets [88]. Lagrangian partial relaxation and partitioning into sub-problems is applied for solving the problem and the proposal's superiority is validated by simulations. Despite the compound caching optimization study in D2D networks, it lacks realistic

content modeling aspects like chunk, layer and it does not look up for energy efficiency contrary to what we study in Chapter 5. In [69], offload maximizing optimal caching and resource allocation is presented limited to offloading probability inspection only. But in our complete D2D cellular network rather the service capacity and energy consumption are elaborated in great detail. In [89], they propose the overall observed content quality maximizing cache placement strategy and also a resource allocation algorithm for maximizing the content quality of users restricted by tolerable delay. They elaborate on the quality in [89]. However, our motivation is to improve the EE in our D2D network in Chapter 5.

Our core motivation is to optimize the cache replacement problem of popularity, partition and quality based multi-attributed multimedia contents for minimizing the energy burden of D2D edge networks. In a nutshell, the energy efficiency is not fully studied in the literature for D2D caching optimization problem and hence our study in Chapter 5 differs from the literature with the full and comprehensive EE investigation for the D2D edge caching optimization problem with its low-complexity heuristic algorithm addendum.

3. A MARKOVIAN MODEL FOR SATELLITE INTEGRATED COGNITIVE AND D2D HETNETS

In this chapter, we elaborate on a hybrid satellite-terrestrial network with D2D and cognitive communications from content consumption aspect. The proliferation of multimedia requests are handled with the advent of several wireless networking techniques. Regarding the spectrum sharing, CR and D2D approaches are handy to alleviate the capacity crunch [51]. The satellite terrestrial HetNets are also profoundly useful in terms of 5G system envision for the service capacity expansion. Hybrid satellite networks are considered as instrumental and efficient systems for 5G actualization in multimedia operations [24]. Despite the fact that there are many literature studies making use of these networking approaches, their intersection as a satellite and cellular network with D2D and cognitive extension is yet to be explored comprehensively, especially from the multimedia transmission aspect for prospective 5G networking. In this work, we model satellite and cellular HetNet architecture entailing CR paradigm and D2D operation. We include model factors such as the integration of *universal source* concept (modeling the content retrieval operation from external networks), caching and overlaying in D2D mode regarding a complete hybrid network treatment with a realistic view. In terms of performance proliferation, we elaborated on these factors rigorously and finally, we analyzed the effect of mode selection in such network systems. Our content-oriented model and its profound analysis is available in [1] and [2]. Besides, we investigated connectivity mode assignment for EE optimization. Due to feasibility concerns, we developed a sub-optimal technique. This connectivity mode assignment study is available in [3].

3.1. System Model

In this section, we model satellite integrated cognitive and D2D Hetnets and present the analytical system performance results and also verify them with simulations. Our modeled network is a content-centric HetNet with D2D and cognitive

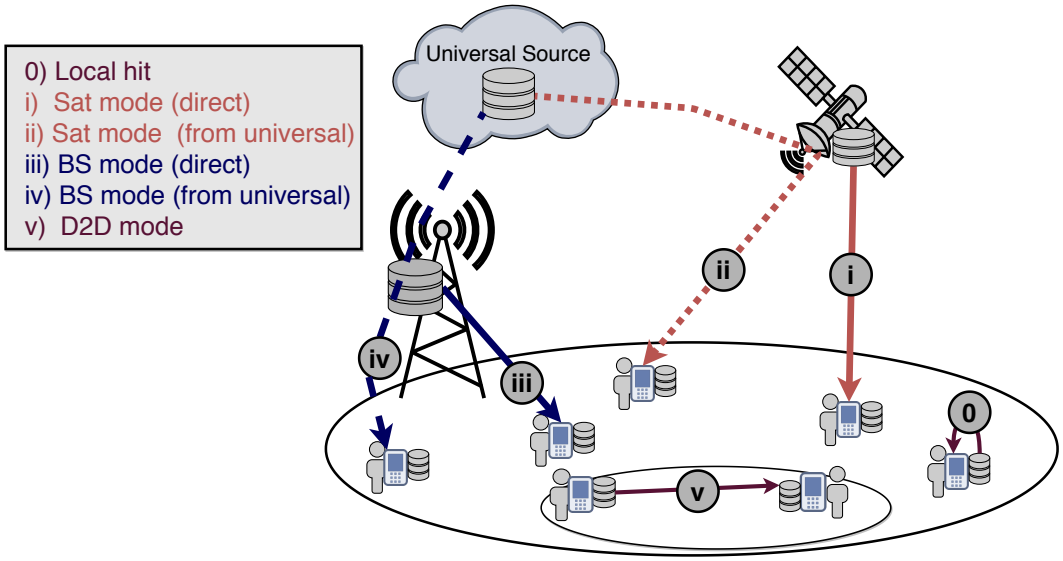

Figure 3.1. Multi-mode operating system model [1].

communications as shown in Figure 3.1. We have network users in our model that can operate in both of two distinct frequency ranges as in [90, 91]: (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 [51]. 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 throughput 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 Subsection 3.3.3. The analysis parameters are listed in Table 3.1.

In the following section, we begin the construction of our analytical model from the caching aspect. Subsequently, Markov modeling of the resource allocation (RA) is done where the continuous time Markov chain (CTMC) technique is used (Section 3.3). 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 transitions at the terrestrial link as our users operate as SUs at this type of link. Accordingly, we define PU arrival and departure transition rates (Subsection 3.3.1). 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 (Subsection 3.3.2). Finally, we focus on HU transitions (Subsection 3.3.3).

3.2. 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 [24]. 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 [92]. 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 [24]. Naturally, popularity-aware or -driven caching policies constitute the key caching approach for content-centric networks (e.g. see [67, 93, 94]). 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 modeling perspective, the advantage of the PDC policy is the intuitive integration into our analytical Markov model.

In our setting, $Contents$=$\{c_1, c_2, ... \ c_N\}$ are chunks to be consumed by users. The content size distribution is exponential with mean $s(\hat{v_b})$=25 Mbits [51]. The request probability $p_{c_i}(\alpha)$ for each content c_i is assigned based on the Zipf distribution with parameter α. These probabilities $p_{c_i}(\alpha)$'s for $i \in \{1, 2, ...N\}$ follow $\sum_{c_i \in Contents} p_{c_i}(\alpha) = 1$ and $\forall c_i \in Contents$, $p_{c_i}(\alpha) \geq 0$ with $p_{c_i}(\alpha) = \frac{\frac{1}{i^\alpha}}{\sum_{j=1}^{N}(\frac{1}{j^\alpha})}$. The Poisson arrival processes are commonly used for multimedia traffic modeling [95, 96]. 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 $\lambda_{HU}^{c_i}$=$p_{c_i}(\alpha)\lambda_{HU}$ 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 provided in Figure 3.2 and designated as a flow in Figure 3.3. In the RA phase, we make use of content availability probabilities at system units. Therefore, we derive these probabilities at local HU device, satellite and BS caches in this subsection. The content availability for D2D operation at some HU device within the reception range of the requester is investigated in Subsection 3.3.2.

When we look at the content size distribution parameter $\lambda_{c(size)}$, it is equal to $\frac{1}{s(\hat{v_b})}$ with $s(\hat{v_b}) = 25$ Mbits. The probability density function for content size is define

INPUTS

$\mathbb{C}$: The local cache set — C_{Dev}^{cache}: The device cache capacity

c_{new}: The newly requested content — N: The total number of contents

α: The Zipf distribution parameter

PDC($\mathbb{C}, C_{Dev}^{cache}, N, \alpha, c_{new}$){

if ($\mathbb{C} = \emptyset$ && $size(c_{new}) \leq C_{Dev}^{cache}$) **then**

  Cache c_{new};

else if ($c_i \in \mathbb{C}$ && $size(c_i) + size(c_{new}) \leq C_{Dev}^{cache}$) **then**

  Cache c_{new};

else

  %Local cache contains contents c_i and c_j.

  %Select a $\theta \in [0, 1]$ random uniformly

  %Calculate the request probabilities $p_{c_i}(\alpha)$ and $p_{c_j}(\alpha)$ based on the Zipf distribution.

  if ($\theta \leq \frac{p_{c_j(\alpha)}}{p_{c_i(\alpha)} + p_{c_j(\alpha)}}$) **then**

    if ($size(c_{new}) + size(c_j) \leq C_{Dev}^{cache}$) **then**

      Evict c_i;

      Cache c_{new};

  else

    if ($size(c_i) + size(c_{new}) \leq C_{Dev}^{cache}$) **then**

      Evict c_j;

      Cache c_{new};

  end

end

Figure 3.2. Popularity-driven caching (PDC) algorithm for two contents in the cache case [2].

as follows:

$$f_c(x; \lambda_{c(size)}) := \begin{cases} \lambda_{c(size)} e^{-\lambda_{c(size)} x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

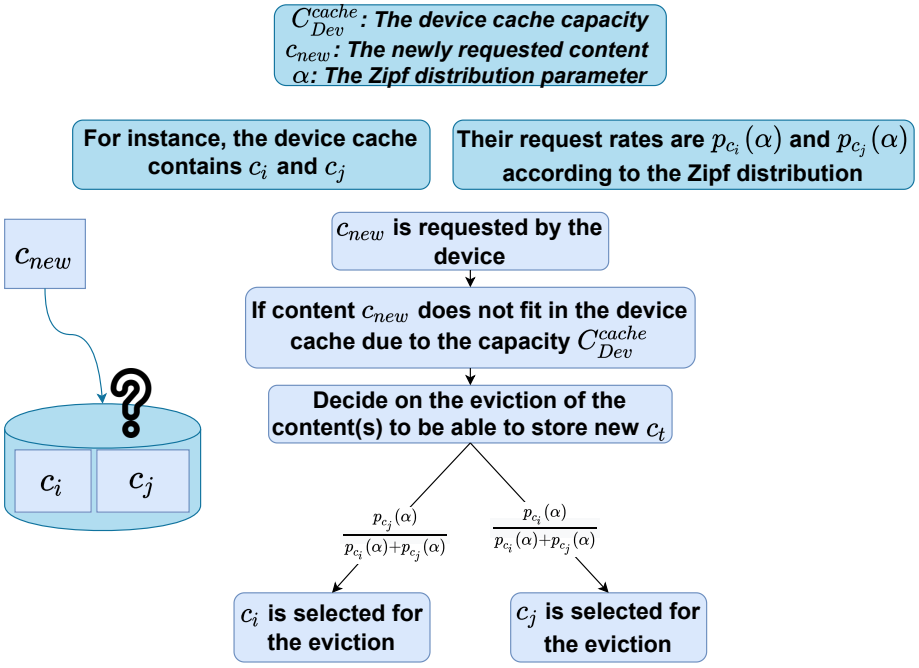

Figure 3.3. PDC algorithm for the scenario with two contents in the cache.

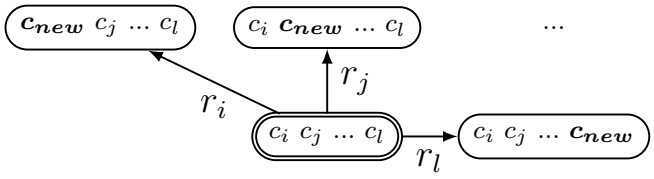

Figure 3.4. Cache update of a content-retrieval unit [1].

The cumulative distribution function for content size is utilized to analytically ensure that the cache capacity is not exceeded by cached contents.

$$F_c(x;\lambda_{c(size)}) := \begin{cases} 1 - e^{-\lambda_{c(size)}x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

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 Figure 3.4. With probability $p_{c_{new}}(\alpha)$ (if c_{new} is a popular content, $p_{c_{new}}(\alpha)$ has higher value), c_{new} will be cached. If c_{new} 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_{new} with the greatest probability, i.e. with the highest rate r_i:=$p_{c_{new}}(\alpha)(\frac{x_i}{\sum_{\theta\in\{i,j,...,l\}} x_\theta})$ among all r's for $\{i,j,...,l\}$. r_i and x_γ are given as follows:

$$r_i := p_{c_t}(\alpha)(\frac{x_i}{\sum_{\theta\in\{i,j,...,l\}} x_\theta}) \tag{3.3}$$

$$x_\gamma := [\prod_{x\in\{i,j,...,l\}} p_{c_x}(\alpha)]/[p_{c_\gamma}(\alpha)] \text{ where } \gamma \in \{i,j,...,l\} \tag{3.4}$$

In the tractable example of local cache analysis storing two contents, we cover all the states and then derive balance equations. By solving the equation set, we derive the probability $p_{(empty)}$ of being at empty cache state, the probability $p_{(c_i)}$ of one content storing states (c_i) and the probability $p_{(c_i\ c_j)}$ of two distinct content storing states $(c_i\ c_j)$ for i $\in \{1,2,...,N\}$ and j $\in \{1,2,...,N\}$ with $j \neq i$. Based on these probabilities, we finally reach to the steady state content availability probabilities $p_{c_i}^{lo}$ provided in Equation 3.5.

$$p_{c_i}^{lo} := p_{(c_i)} + \sum_{j=1,j\neq i}^{N} p_{(c_i\ c_j)} \tag{3.5}$$

The availability probabilities $p_{c_i}^{BS}$ and $p_{c_i}^{sat}$ for any content c_i (in Table 3.1) are calculated with the help of Markov chains as well. Further explanation is provided in our technical pre-print paper [2].

3.3. 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 capabilities of the BS as a facilitator of cognitive operation [97]. BS performs the centralized

Table 3.1. System analysis parameters for the markovian model.

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
α	The Zipf parameter
c_i	The i^{th} content in the content set
$p_{c_i}(\alpha)$	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

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 [95, 96]. Besides, multimedia traffic completions are modeled by exponential distributions [51, 96]. 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 [96, 98, 99]. 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 3.1.

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 [51], 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 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 PUs over the terrestrial link is calculated as $\mu_{PU}^{ter} := \frac{C_{PU}^{ter}}{s(\hat{v_b})}$. 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 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 3.13 for the performance evaluation section in Chapter 3.

A requested content is transmitted from the universal source to the satellite with mean transmission duration $\frac{s(\hat{v_b})}{C_{HU}^{sat(u)}}$. Next, with duration $\frac{s(\hat{v_b})}{C_{HU}^{sat}}$ it is transmitted from the satellite to the req-HU. By adding these durations as shown in (3.6), we get the mean total service duration of a content fetch from the universal source across the satellite to the requester HU (req-HU). The mean total service duration of some content fetch from the universal source across the BS to the req-HU is calculated similarly given in Equation 3.6 where $x \in \{sat, BS\}$.

$$\Delta_{HU}^{x(u)} := \frac{s(\hat{v_b})}{C_{HU}^{x(u)}} + \frac{s(\hat{v_b})}{C_{HU}^{x}} \tag{3.6}$$

The service rate of HUs that fetch content from the universal source across the satellite or the BS to the req-HU is the reciprocal of the mean aggregate service time as given in Equation 3.7.

$$\mu_{HU}^{x(u)} := \frac{1}{\Delta_{HU}^{x(u)}} \text{ where } x \in \{sat, BS\} \tag{3.7}$$

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

Figure 3.5. Channel state [1].

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 3.5. Their definitions are given in Table 3.2.

Table 3.2. 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

Table 3.3. Transitions originating at a generic state s_0 due to PU arrivals.

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

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.

3.3.1. 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*

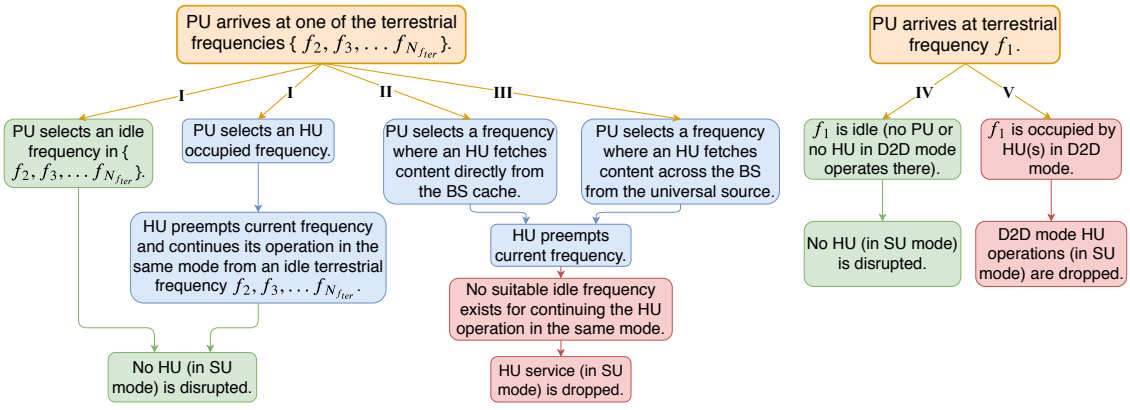

Figure 3.6. PU arrival layout (green:no drop, red:drop, blue:preemption, orange:case selection) [1].

Table 3.4. Transitions originating at a generic state s_0 due to PU departures.

	Destination State	**Transition Rate**
I	$s_{(i_{PU}^{ter(\overline{f_1})}-1)}$	$i_{PU}^{ter(\overline{f_1})}(s_0)\mu_{PU}^{ter}$
II	$s_{(i_{PU}^{ter(f_1)}-1)}$	$i_{PU}^{ter(f_1)}(s_0)\mu_{PU}^{ter}$

at the terrestrial frequency f_1. For the *other terrestrial frequencies*, HUs operate in *BS mode* as described in Subsection 3.3.3. 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 Figure 3.5 as s_0 and elaborate on PU arrival cases originating at s_0 in Table 3.3. We also elaborate on PU departures cases originated at s_0 in Table 3.4. Each row in these tables corresponds to a destination state from s_0 with the corresponding transition rate. 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 3.1. $1_{(\theta)}$ is the indicator function defined as 1 if θ is true, 0 otherwise (these functions are also used in Subsection 3.3.3). We give a detailed layout in Figure 3.6 for each PU arrival case listed in Table 3.3.

Until now, we investigated PU activities originated at s_0. For getting the complete set of balance equations, we also need the transitions destined to s_0. These transitions

Table 3.5. Transitions destined to a generic state s_0 due to PU arrivals.

Source State	**Transition Rate**
$s_1 := \mathrm{s}_{(i_{PU}^{ter(\overline{f_1})}-1)}$	$\lambda_{PU}^{ter} 1_{((idle_{t,\overline{f_1}}(s_1)>0)\wedge(i_{PU}^{ter(\overline{f_1})}(s_0)>0))}(\frac{N_{f_{ter}}-1}{N_{f_{ter}}})$
$s_2 := \mathrm{s}_{(i_{PU}^{ter(\overline{f_1})}-1,i_{HU}^{BS}+1)}$	$\lambda_{PU}^{ter} 1_{((i_{PU}^{ter(\overline{f_1})}(s_0)>0)\wedge(i_{HU}^{BS(u)}(s_0)<(N_{f_{ter}}-1)))}$ $[1_{((idle_{t,\overline{f_1}}(s_2)==0)\wedge(N_{f_{ter}}-i_{PU}^{ter(\overline{f_1})}(s_2)>0))}\times(\frac{i_{HU}^{BS}(s_2)}{N_{f_{ter}}-i_{PU}^{ter(\overline{f_1})}(s_2)})](\frac{N_{f_{ter}}-1}{N_{f_{ter}}})$
$s_3 := \mathrm{s}_{(i_{PU}^{ter(\overline{f_1})}-1,i_{HU}^{BS(u)}+1)}$	$\lambda_{PU}^{ter} 1_{((i_{PU}^{ter(\overline{f_1})}(s_0)>0)\wedge(i_{HU}^{BS(u)}(s_0)<(N_{f_{ter}}-1)))}$ $[1_{((idle_{t,\overline{f_1}}(s_3)==0)\wedge(N_{f_{ter}}-i_{PU}^{ter(\overline{f_1})}(s_3)>0))}\times(\frac{i_{HU}^{BS(u)}(s_3)}{N_{f_{ter}}-i_{PU}^{ter(\overline{f_1})}(s_3)})](\frac{N_{f_{ter}}-1}{N_{f_{ter}}})$
$s_4 := \mathrm{s}_{(i_{PU}^{ter(f_1)}-1)}$	$\lambda_{PU}^{ter} 1_{((i_{PU}^{ter(f_1)}(s_4)==0)\wedge(i_{HU}^{D(f_1)}(s_4)==0))}(\frac{1}{N_{f_{ter}}})$
$s_5 := \mathrm{s}_{(i_{PU}^{ter(f_1)}-1,i_{HU}^{D(f_1)}>0)}$	$\lambda_{PU}^{ter} 1_{((i_{HU}^{D(f_1)}(s_5)>0)\wedge(i_{PU}^{ter(f_1)}(s_0)==1)\wedge(i_{HU}^{D(f_1)}(s_0)==0))}(\frac{1}{N_{f_{ter}}})$

Table 3.6. Transitions destined to a generic state s_0 due to PU departures.

Source State	**Transition Rate**
$s_6 := \mathrm{s}_{(i_{PU}^{ter(\overline{f_1})}+1)}$	$i_{PU}^{ter(\overline{f_1})}(s_6)\mu_{PU}^{ter} 1_{((idle_{t,\overline{f_1}}(s_0)>0)\wedge(i_{PU}^{ter(\overline{f_1})}(s_0)<N_{f_{ter}}-1))}$
$s_7 := \mathrm{s}_{(i_{PU}^{ter(f_1)}+1)}$	$i_{PU}^{ter(f_1)}(s_7)\mu_{PU}^{ter} 1_{((i_{PU}^{ter(f_1)}(s_0)==0)\wedge(i_{HU}^{D(f_1)}(s_0)==0))}$

of PU arrival type are illustrated in Table 3.5 and PU departure type in Table 3.6.

3.3.2. 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 [100, 101]. 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*. D_{max} is the maximum number of concurrent D2D operations allowed by the network that is bounded as shown in Equation 3.8. We subtract the maximum number of non-D2D mode (satellite and BS mode) active HU devices $(N_{f_{sat}} + (N_{f_{ter}} - 1))$ from the average number of HU devices in the cell $(\lambda_{N_{HU}}\pi R_{BS}^2)$ to be able to guarantee that even if all non-D2D operations are active there is still room for D2D. Two HUs are actively used (one as a transmitter, the other as a receiver) and blocked for new operation in D2D mode and hence we divide the

maximum number of HUs in D2D mode by two to determine the upper bound for D_{max}.

$$D_{max} \leq \llcorner \frac{(\lambda_{N_{HU}} \pi R_{BS}^2) - N_{f_{sat}} - (N_{f_{ter}} - 1)}{2} \lrcorner \tag{3.8}$$

For any content c_i, the D2D content availability probability $p_{c_i}^{D(f_1)}(x)$ of channel state x is calculated in Equation 3.9 that is designated in great detail in Figure 3.7. $p_{c_i}^{D(f_1)}(x)$ 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 Equation 3.10 : the receiver HU (rec-HU) is not being interfered by other D2D operations.
- $\Pi_{tx}(x)$ in Equation 3.11 : the transmitter HU (tx-HU) will not cause interference to active D2D operations.
- Π_{c_i} in Equation 3.12 : content c_i is retrievable over the terrestrial frequency f_1.

$$p_{c_i}^{D(f_1)}(x) := 1_{(0<=i_{HU}^{D(f_1)}(x)<D_{max})} \Pi_{rec}(x) \Pi_{tx}(x) \Pi_{c_i} \tag{3.9}$$

$$\Pi_{rec}(x) := \frac{max\{0, (\pi R_{BS}^2) - (i_{HU}^{D(f_1)}(x) \pi R_{Int}^2)\}}{\pi R_{BS}^2} \tag{3.10}$$

$$\Pi_{tx}(x) := \frac{max\{0, (\pi R_{BS}^2) - (i_{HU}^{D(f_1)}(x) \pi (2R_{Int})^2)\}}{\pi R_{BS}^2} \tag{3.11}$$

$$\Pi_{c_i} := 1 - (1 - p_{c_i}^{lo})^{\lfloor \lambda_{N_{HU}} \pi R_{Int}^2 \rfloor} \tag{3.12}$$

For modeling interference in [56], *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

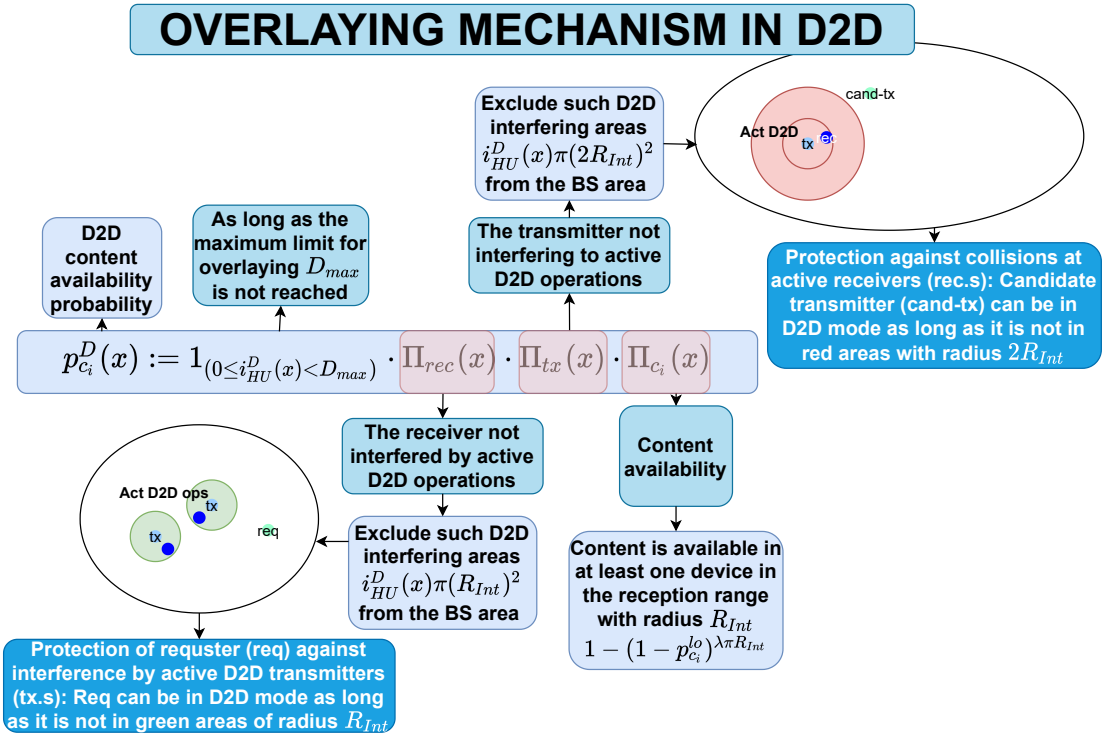

Figure 3.7. Overlaying mechanism in D2D mode for the requester.

be negligible.

$\Pi_{rec}(x)$ is calculated by subtracting the interference ranges of active tx-HU devices in D2D mode (dark shaded areas in Figure 3.8 $\rightarrow i^{D(f_1)}_{HU}(x)\pi R^2_{Int}$) from the cell area ($\pi R^2_{BS}$) and then then dividing over πR^2_{BS}. 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 $i^{D(f_1)}_{HU}(x)(\pi 2R_{Int})^2$. $\Pi_{tx}(x)$ is calculated by subtracting this aggregated prohibition zone for tx-HU candidates (dark shaded area+light shaded area in Figure 3.8$\rightarrow$ $i^{D(f_1)}_{HU}(x)(\pi 2R_{Int})^2$) from the cell area ($\pi R^2_{BS}$) and dividing over πR^2_{BS}. Again, `max` function is used.

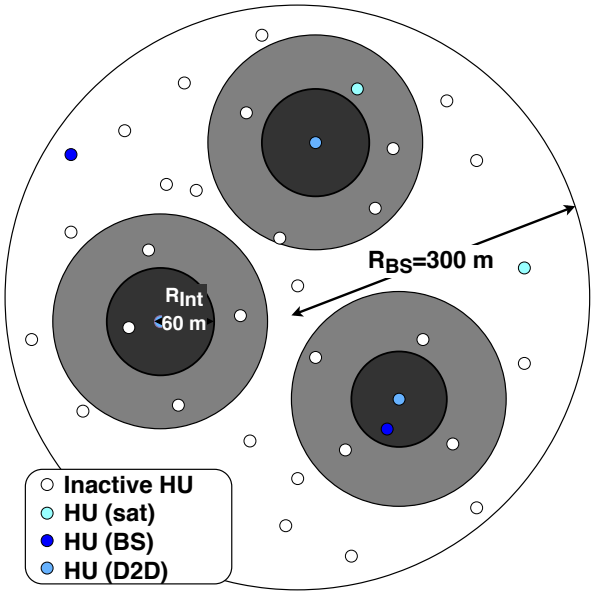

Figure 3.8. D2D spatial stochastic model [1, 2].

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 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 Equation 3.12 gives the probability of finding at least one HU device storing content c_i in the reception range of req-HU.

The prohibited areas for a new HU transmitter candidate in D2D mode (dark shaded areas+light shaded areas in Figure 3.8) can intersect with each other and thus the prohibited area can have values lower than $(i_{HU}^{D(f_1)}(x)\pi(2R_{Int})^2)$ resulting in $\Pi_{tx}(x)$ serve as a lower bound. Besides, the interference range of active tx-HUs in D2D mode (dark shaded areas in Figure 3.8) or the prohibited areas for a new HU transmitter candidate in D2D mode (dark shaded areas+light shaded areas in Figure 3.8) can intersect with the BS cell boundary. So, $\Pi_{rec}(x)$ and $\Pi_{tx}(x)$ serve as lower bounds for the corresponding probabilities and hence this leads to $p_{c_i}^{D(f_1)}(x)$ serving as a lower-bound for D2D content availability probability.

3.3.3. 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 Equations 3.13-3.15, for mode selection.

$$R_{sat}(x) := r_{sat} \cdot idle_s(x) \tag{3.13}$$

$$R_{BS}(x) := r_{BS} \cdot idle_{t,\overline{f_1}}(x) \tag{3.14}$$

$$R_{D2D}(x) := r_{dev} \cdot [1_{(0<i_{HU}^{D(f_1)}(x)<D_{max})} + 1_{((i_{HU}^{D(f_1)}(x)==0)\wedge(i_{PU}^{ter(f_1)}(x)==0))}] \tag{3.15}$$

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 Figure 3.5 is denoted as h_0. First, we analyze the transitions originating at h_0 upon HU arrivals visualized in Figure 3.9. 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

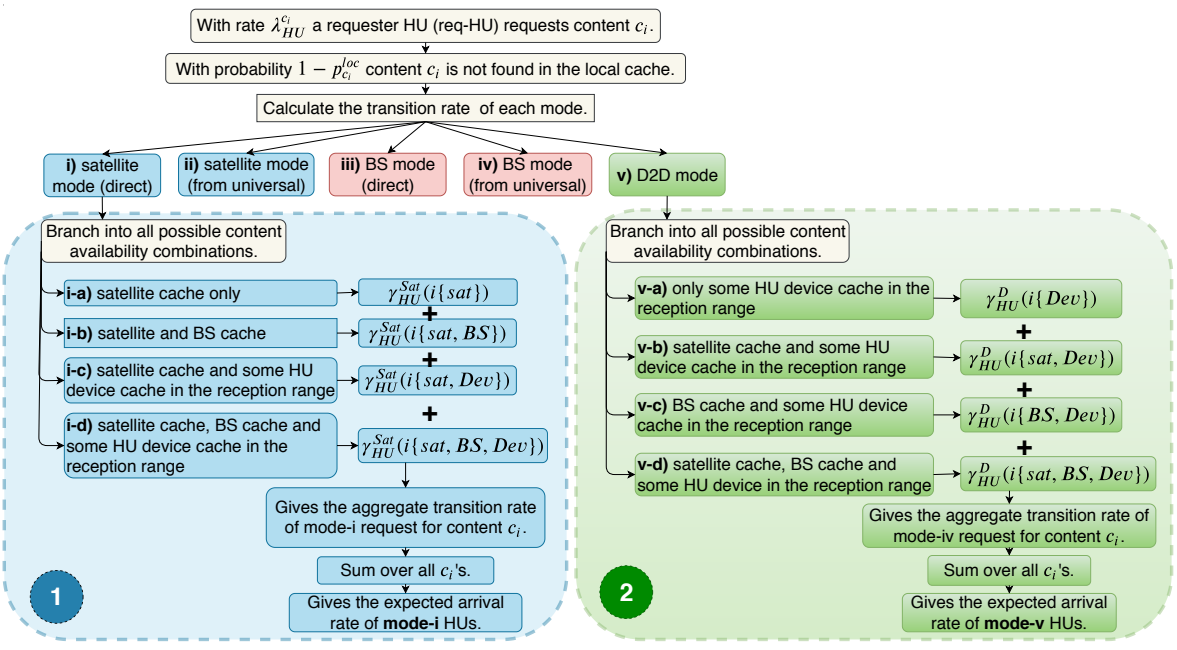

Figure 3.9. Algorithm for HU arrival state transition calculation(blue: satellite, red: BS, green: D2D) [1].

(shown in Figure 3.9) 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 Figure 3.9 (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 3.7. We sum over these rates to get the aggregate transition rate of the mode-i

Table 3.7. Mode-i HU arrival transitions originated at a generic state h_0.

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

service request for the retrieval of content c_i in Equation 3.16. By summing Equation 3.16 for all c_i's, we get the expected arrival rate of mode-i HUs into the network in Equation 3.17. This is the general scheme for the mode-i state transition calculation.

$$\Gamma^{sat}_{HU}(i) := \gamma^{sat}_{HU}(i,\{sat\}) + \gamma^{sat}_{HU}(i,\{sat,BS\}) + \gamma^{sat}_{HU}(i,\{sat,Dev\}) + \gamma^{sat}_{HU}(i,\{sat,BS,Dev\}) \tag{3.16}$$

$$\Gamma^{sat}_{HU} := \sum_{i=1}^{N}\Gamma^{sat}_{HU}(i) \tag{3.17}$$

Now, let us explain some of the mode-i branch calculations. For instance, when we look at the branch (i-a) (in Table 3.7), 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 also available in Table 3.7.

Table 3.8. Mode-*ii* HU arrival transitions originated at a generic state h_0.

	Content availability	**Dest. State**	**Prob. of content availability**	**Transition Rate**
ii-a	not available	$s_{(i_{HU}^{sat(u)}+1)}$	$\mathbb{P}_{(\emptyset)}^{c_i}=[1-p_{c_i}^{lo}][1-p_{c_i}^{sat}][1-p_{c_i}^{BS}][1-p_{c_i}^{D(f_1)}(h_0)]$	$\gamma_{HU}^{sat(u)}(i,\emptyset)=\lambda_{HU}^{c_i}\mathbb{P}_{(\emptyset)}^{c_i}$ $[\frac{R_{sat}(h_0)}{R_{sat}(h_0)+R_{BS}(h_0)}]1_{(R_{sat}(h_0)+R_{BS}(h_0)>0)}$
ii-b	BS cache only	$s_{(i_{HU}^{sat(u)}+1)}$	$\mathbb{P}_{(B)}^{c_i}=[1-p_{c_i}^{lo}][1-p_{c_i}^{sat}]p_{c_i}^{BS}[1-p_{c_i}^{D(f_1)}(h_0)]$	$\gamma_{HU}^{sat(u)}(i,\{BS\})=\lambda_{HU}^{c_i}\mathbb{P}_{(B)}^{c_i}$ $1_{([(idle_{t,\overline{f1}}(h_0)==0)\vee(r_{BS}==0)]\wedge(idle_s(h_0)>0)\wedge(r_{sat}>0))}$
ii-c	some HU device in the reception range of req-HU only	$s_{(i_{HU}^{sat(u)}+1)}$	$\mathbb{P}_{(D)}^{c_i}=[1-p_{c_i}^{lo}][1-p_{c_i}^{sat}][1-p_{c_i}^{BS}]p_{c_i}^{D(f_1)}(h_0)$	$\gamma_{HU}^{sat(u)}(i,\{Dev\})=\lambda_{HU}^{c_i}\mathbb{P}_{(D)}^{c_i}[\frac{R_{sat}(h_0)}{R_{sat}(h_0)+R_{BS}(h_0)}]$ $1_{(R_{sat}(h_0)+R_{BS}(h_0)>0)}$ $1_{((r_{dev}==0)\vee(i_{HU}^{D(f_1)}(h_0)==D_{max})\vee(i_{PU}^{ter(f_1)}(h_0)==1))}$
ii-d	BS cache and some HU device in the reception range of req-HU	$s_{(i_{HU}^{sat(u)}+1)}$	$\mathbb{P}_{(B,D)}^{c_i}=[1-p_{c_i}^{lo}][1-p_{c_i}^{sat}]p_{c_i}^{BS}p_{c_i}^{D(f_1)}(h_0)$	$\gamma_{HU}^{sat(u)}(i,\{BS,Dev\})=\lambda_{HU}^{c_i}\mathbb{P}_{(B,D)}^{c_i}$ $1_{((R_{BS}(h_0)+R_{D2D}(h_0)==0)\wedge(idle_s(h_0)>0)\wedge(r_{sat}>0))}$

For the transition rate of mode-*ii* (*satellite mode from universal*) calculation, we branch into content availability combinations ii-a, ii-b, ii-c and ii-d. These branches and corresponding transition rates are provided in Table 3.8. Similar to mode-*i* case, by the summation of these rates, we get the aggregate transition rate of mode-*ii* service request for the retrieval of content c_i in Equation 3.18. The summation of $\Gamma_{HU}^{sat(u)}(i)$ over all content c_i's outputs the expected arrival rate of mode-*ii* HUs into the system provided in Equation 3.19.

$$\Gamma_{HU}^{sat(u)}(i) := \gamma_{HU}^{sat(u)}(i,\{\emptyset\}) + \gamma_{HU}^{sat(u)}(i,\{BS\}) + \gamma_{HU}^{sat(u)}(i,\{Dev\}) + \gamma_{HU}^{sat(u)}(i,\{BS,Dev\}) \tag{3.18}$$

$$\Gamma_{HU}^{sat(u)} := \sum_{i=1}^{N}\Gamma_{HU}^{sat(u)}(i) \tag{3.19}$$

The mode-*iii* (*BS mode direct*) and *iv* (*BS mode from universal*) transition rate branches are elaborated in Table 3.9. The aggregate transition rate of mode-*iii* and -*iv* service request for retrieval of content c_i are given in Equation 3.20 and 3.21 respectively. The expected arrival rate of mode-*iii* and -*iv* HUs into the network are

Table 3.9. Mode-*iii* and -*iv* HU arrival transitions originated at a generic state h_0.

	Content availability	**Dest. State**	**Prob. of content availability**	**Transition Rate**
iii-a	BS cache only	$s_{(i_{HU}^{BS}+1)}$	$\mathbb{P}_{(B)}^{c_i}=[1-p_{c_i}^{lo}][1-p_{c_i}^{sat}]p_{c_i}^{BS}[1-p_{c_i}^{D(f_1)}(h_0)]$	$\gamma_{HU}^{BS}(i,\{BS\})=\lambda_{HU}^{c_i}\mathbb{P}_{(B)}^{c_i}$ $1_{((idle_{t,\overline{f_1}}(h_0)>0)\wedge(r_{BS}>0))}$
iii-b	satellite and BS cache	$s_{(i_{HU}^{BS}+1)}$	$\mathbb{P}_{(S,B)}^{c_i}=[1-p_{c_i}^{lo}]p_{c_i}^{sat}p_{c_i}^{BS}[1-p_{c_i}^{D(f_1)}(h_0)]$	$\gamma_{HU}^{BS}(i,\{sat,BS\})=\lambda_{HU}^{c_i}\mathbb{P}_{(S,B)}^{c_i}$ $[\frac{R_{BS}(h_0)}{R_{sat}(h_0)+R_{BS}(h_0)}]1_{(R_{sat}(h_0)+R_{BS}(h_0)>0)}$
iii-c	BS cache and some HU device in the reception range of req-HU	$s_{(i_{HU}^{BS}+1)}$	$\mathbb{P}_{(B,D)}^{c_i}=[1-p_{c_i}^{lo}][1-p_{c_i}^{sat}]p_{c_i}^{BS}p_{c_i}^{D(f_1)}(h_0)$	$\gamma_{HU}^{BS}(i,\{BS,Dev\})=\lambda_{HU}^{c_i}\mathbb{P}_{(B,D)}^{c_i}$ $[\frac{R_{BS}(h_0)}{R_{BS}(h_0)+R_{D2D}(h_0)}]1_{(R_{BS}(h_0)+R_{D2D}(h_0)>0)}$
iii-d	satellite cache, BS cache and some HU device cache in the reception range of req-HU	$s_{(i_{HU}^{BS}+1)}$	$\mathbb{P}_{(S,B,D)}^{c_i}=[1-p_{c_i}^{lo}]p_{c_i}^{sat}p_{c_i}^{BS}p_{c_i}^{D(f_1)}(h_0)$	$\gamma_{HU}^{BS}(i,\{sat,BS,Dev\})=\lambda_{HU}^{c_i}\mathbb{P}_{(S,B,D)}^{c_i}$ $[\frac{R_{BS}(h_0)}{R_{sat}(h_0)+R_{BS}(h_0)+R_{D2D}(h_0)}]$ $1_{(R_{sat}(h_0)+R_{BS}(h_0)+R_{D2D}(h_0)>0)}$
iv-a	not available	$s_{(i_{HU}^{BS(u)}+1)}$	$\mathbb{P}_{(\emptyset)}^{c_i}=[1-p_{c_i}^{lo}][1-p_{c_i}^{sat}][1-p_{c_i}^{BS}][1-p_{c_i}^{D(f_1)}(h_0)]$	$\gamma_{HU}^{BS(u)}(i,\emptyset)=\lambda_{HU}^{c_i}\mathbb{P}_{(\emptyset)}^{c_i}$ $[\frac{R_{BS}(h_0)}{R_{sat}(h_0)+R_{BS}(h_0)}]1_{(R_{sat}(h_0)+R_{BS}(h_0)>0)}$
iv-b	satellite cache only	$s_{(i_{HU}^{BS(u)}+1)}$	$\mathbb{P}_{(S)}^{c_i}=[1-p_{c_i}^{lo}]p_{c_i}^{sat}[1-p_{c_i}^{BS}][1-p_{c_i}^{D(f_1)}(h_0)]$	$\gamma_{HU}^{BS(u)}(i,\{sat\})=\lambda_{HU}^{c_i}\mathbb{P}_{(S)}^{c_i}$ $1_{([(idle_s(h_0)==0)\vee(r_{sat}==0)]\wedge(idle_{t,\overline{f_1}}(h_0)>0)\wedge(r_{BS}>0))}$
iv-c	some HU device in the reception range of req-HU only	$s_{(i_{HU}^{BS(u)}+1)}$	$\mathbb{P}_{(D)}^{c_i}=[1-p_{c_i}^{lo}][1-p_{c_i}^{sat}][1-p_{c_i}^{BS}]p_{c_i}^{D(f_1)}(h_0)$	$\gamma_{HU}^{BS(u)}(i,\{Dev\})=\lambda_{HU}^{c_i}\mathbb{P}_{(D)}^{c_i}[\frac{R_{BS}(h_0)}{R_{sat}(h_0)+R_{BS}(h_0)}]$ $1_{(R_{sat}(h_0)+R_{BS}(h_0)>0)}$ $1_{((r_{dev}==0)\vee(i_{HU}^{D(f_1)}(h_0)==D_{max})\vee(i_{PU}^{ter(f_1)}(h_0)==1))}$
iv-d	satellite cache and some HU device in the reception range of req-HU	$s_{(i_{HU}^{BS(u)}+1)}$	$\mathbb{P}_{(S,D)}^{c_i}=[1-p_{c_i}^{lo}]p_{c_i}^{sat}[1-p_{c_i}^{BS}]p_{c_i}^{D(f_1)}(h_0)$	$\gamma_{HU}^{BS(u)}(i,\{sat,Dev\})=\lambda_{HU}^{c_i}\mathbb{P}_{(S,D)}^{c_i}$ $1_{((R_{sat}(h_0)+R_{D2D}(h_0)==0)\wedge(idle_{t,\overline{f_1}}(h_0)>0)\wedge(r_{BS}>0))}$

calculated in Equation 3.22 and 3.23 respectively.

$$\Gamma_{HU}^{BS}(i) := \gamma_{HU}^{BS}(i,\{BS\}) + \gamma_{HU}^{BS}(i,\{sat,BS\}) + \gamma_{HU}^{BS}(i,\{BS,Dev\}) + \gamma_{HU}^{BS}(i,\{sat,BS,Dev\}) \tag{3.20}$$

$$\Gamma_{HU}^{BS(u)}(i) := \gamma_{HU}^{BS(u)}(i,\{\emptyset\}) + \gamma_{HU}^{BS(u)}(i,\{sat\}) + \gamma_{HU}^{BS(u)}(i,\{Dev\}) + \gamma_{HU}^{BS(u)}(i,\{sat,Dev\}) \tag{3.21}$$

$$\Gamma_{HU}^{BS} := \sum_{i=1}^{N} \Gamma_{HU}^{BS}(i) \tag{3.22}$$

$$\Gamma_{HU}^{BS(u)} := \sum_{i=1}^{N} \Gamma_{HU}^{BS(u)}(i) \tag{3.23}$$

Table 3.10. Mode-v HU arrival transitions originated at a generic state h_0.

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

For mode-v (*D2D mode*) transition, all the content availability branches and corresponding rates are provided in Table 3.10. The aggregate transition rate of mode-v service request for the retrieval of content c_i is given in Equation 3.24. When transitions over all contents are aggregated, we get the expected arrival rate of mode-v HUs into the network ($\Gamma_{HU}^{D(f_1)} := \sum_{i=1}^{N}\Gamma_{HU}^{D(f_1)}(i)$). The general scheme for this calculation (2) is available in Figure 3.9.

$$\Gamma_{HU}^{D(f_1)}(i) := \gamma_{HU}^{D(f_1)}(i,\{Dev\}) + \gamma_{HU}^{D(f_1)}(i,\{sat,Dev\}) + \gamma_{HU}^{D(f_1)}(i,\{BS,Dev\}) + \gamma_{HU}^{D(f_1)}(i,\{sat,BS,Dev\}) \tag{3.24}$$

To exemplify a branch, when we look at (v-a) (in Table 3.10), the probability of finding content c_i *only* at some HU device in the reception range is $\mathbb{P}_{(D)}^{c_i}$. In this case c_i is retrieved only if $r_{dev} > 0$. Besides, either the terrestrial frequency f_1 should be idle ($1_{((i_{HU}^{D(f_1)}(h_0)==0)\wedge(i_{PU}^{ter(f_1)}(h_0)==0))}$) or the number of maximum concurrent D2D transmission(s) has not been reached ($1_{(0<i_{HU}^{D(f_1)}(h_0)<D_{max})}$).

The HU service completions and hence departures form our generic state h_0 are provided in Table 3.11 with the link type, the destination state and the corresponding transition rate respectively. Apart from the transitions originated at a generic state h_0, the incoming transitions destined to h_0 are needed to solve the balance equations.

Table 3.11. HU departure transitions originated at h_0.

Link	**Destination State**	**Transition Rate**
Sat	$s_{(i_{HU}^{sat(u)}-1)}$	$i_{HU}^{sat(u)}(h_0)\mu_{HU}^{sat(u)}$
Ter	$s_{(i_{HU}^{BS(u)}-1)}$	$i_{HU}^{BS(u)}(h_0)\mu_{HU}^{BS(u)}$
Sat	$s_{(i_{HU}^{sat}-1)}$	$i_{HU}^{sat}(h_0)\mu_{HU}^{sat}$
Ter	$s_{(i_{HU}^{BS}-1)}$	$i_{HU}^{BS}(h_0)\mu_{HU}^{BS}$
Ter	$s_{(i_{HU}^{D(f_1)}-1)}$	$i_{HU}^{D(f_1)}(h_0)\mu_{HU}^{D}$

For the sake of readability, we omit incoming transitions. By solving the complete set of balance equations, we 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.

The channel state space complexity is $O((D_{max})^{\alpha}(N_{f_{ter}}-\alpha)^3 N_{f_{sat}}^2)$ where α is the number of terrestrial frequencies dedicated for D2D operation. We take α=1, so the space complexity is $O(D_{max}(N_{f_{ter}}-1)^3 N_{f_{sat}}^2)$. Solving the system of balance equations to calculate the steady state probailities π_x's is $O(n^3)$ with the total number of n channel states.

3.4. 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 mode-*i*, -*ii*, -*iii*, -*iv* and -*v* HU effective arrival rates are defined from Equation 3.25 to 3.29, respectively.

$$\lambda_{eff(HU)}^{sat} := \sum_{x \in S} \Gamma_{HU}^{sat}(x)\pi_x \tag{3.25}$$

$$\lambda_{eff(HU)}^{sat(u)} := \sum_{x \in S} (\sum_{i=1}^{N} \gamma_{HU}^{sat(u)}(i,x))\pi_x \tag{3.26}$$

$$\lambda_{eff(HU)}^{BS} := \sum_{x \in S} \Gamma_{HU}^{BS}(x)\pi_x \tag{3.27}$$

$$\lambda_{eff(HU)}^{BS(u)} := \sum_{x \in S} (\sum_{i=1}^{N} \gamma_{HU}^{BS(u)}(i,x))\pi_x \tag{3.28}$$

$$\lambda_{eff(HU)}^{D2D} := \sum_{x \in S} \Gamma_{HU}^{D(f_1)}(x)\pi_x \tag{3.29}$$

We have already defined Γ_{HU}^{sat} above in Equation 3.17 originating at state h_0. For mathematical completeness, we use $\Gamma_{HU}^{sat}(x)$ with x denoting the channel state in our calculations. $\Gamma_{HU}^{sat}(x)$ is the overall mode-i HU rate retrieving content directly from the satellite for a given channel state x. The weighted sum of $\Gamma_{HU}^{sat}(x)$ over all channel states gives the expected mode-i effective HU arrival rate into the network provided in Equation 3.25. As a remark, this is an effective arrival rate since the blocking of HUs are taken into account in the calculation of $\Gamma_{HU}^{sat}(x)$.

The mode-ii HU effective arrival rate retrieving contents from the universal source across the satellite to req-HUs is given in Equation 3.26. $\gamma_{HU}^{sat(u)}(i,x)$ is the overall mode-ii HU retrieval rate of content c_i for a given channel state x (in Table 3.8). By summing $\gamma_{HU}^{sat(u)}(i,x)$ over all contents ($\sum_{i=1}^{N} \gamma_{HU}^{sat(u)}(i,x)$) and getting the weighted sum over all channel states ($\sum_{x \in S}(\sum_{i=1}^{N} \gamma_{HU}^{sat(u)}(i,x))\pi_x$), we get the expected mode-ii effective HU arrival rate into the network. The definitions of other expected effective HU arrival rates $\lambda_{eff(HU)}^{BS}$, $\lambda_{eff(HU)}^{BS(u)}$ and $\lambda_{eff(HU)}^{D2D}$ are given in Equation 3.27, 3.28 and 3.29.

Now, we look at the dropping probability of HUs in BS mode provided in Equation 3.31. In the channel state x, when there is at least one HU getting service in BS mode $\rightarrow 1_{(i_{PU}^{ter(\overline{f_1})}(x) \neq (N_{f_{ter}} - 1))}$and no idle frequency in the frequency set {f_2, f_3,...

$f_{N_{f_{ter}}}\}$ $(1_{(idle_{t,\overline{f_1}}(x)==0)})$, due to PU appearance HU cannot find an idle frequency to continue its operation in BS mode and that HU is forcibly dropped. The probabilities of being at channel states that cause such HU drops are summed to U in Equation 3.30. Multiplying U with the arrival rate of PUs to the terrestrial frequency set $\{f_2,\ f_3,...\ f_{N_{f_{ter}}}\}$ $(\frac{\lambda_{PU}^{ter}(N_{f_{ter}}-1)}{N_{f_{ter}}})$, we get the forcibly terminated rate of HUs in BS mode due to PU arrivals. The division of this forcibly terminated HU rate in BS mode over the aggregate effective HU arrival rate in BS mode $(\lambda_{eff(HU)}^{BS}+\lambda_{eff(HU)}^{BS(u)})$ gives p_{drop}^{BS} in Equation 3.31.

$$U:=\sum_{x\in S}(\pi_x\cdot 1_{((idle_{t,\overline{f_1}}(x)==0)\wedge(i_{PU}^{ter(\overline{f_1})}(x)\neq(N_{f_{ter}}-1)))}) \tag{3.30}$$

$$p_{drop}^{BS}:=\frac{\frac{\lambda_{PU}^{ter}(N_{f_{ter}}-1)}{N_{f_{ter}}}U}{\lambda_{eff(HU)}^{BS}+\lambda_{eff(HU)}^{BS(u)}} \tag{3.31}$$

The dropping probability of HUs in D2D mode is defined in Equation 3.32. PU arrives at terrestrial frequency f_1 if no PU exists there $(i_{PU}^{ter(f_1)}(x)==0)$. Due to overlaying in D2D mode, upon a PU arrival there all active D2D operations will be dropped. The probability of being at channel states that cause such incidents is summed as $(\sum_{x\in S}(i_{HU}^{D(f_1)}(x)\cdot\pi_x\cdot 1_{(i_{PU}^{ter(f_1)}(x)==0)}))$. Multiplying this probability with the PU arrival rate to the terrestrial frequency f_1 $(\frac{\lambda_{PU}^{ter}}{N_{f_{ter}}})$, we get the forcibly terminated HU rate in D2D mode due to PU arrival. The division of this forcibly terminated HU rate in D2D mode over the effective HU arrival rate in D2D mode $(\lambda_{eff(HU)}^{D2D})$ gives p_{drop}^{D2D} in Equation 3.32.

$$p_{drop}^{D2D}:=\frac{\frac{\lambda_{PU}^{ter}}{N_{f_{ter}}}\cdot(\sum_{x\in S}(i_{HU}^{D(f_1)}(x)\cdot\pi_x\cdot 1_{(i_{PU}^{ter(f_1)}(x)==0)}))}{\lambda_{eff(HU)}^{D2D}} \tag{3.32}$$

For any content c_i, $p_{c_i}^{lo}$ is the local cache content availability probability as given in Section 3.2. By multiplying $p_{c_i}^{lo}$ with the content request rate $\lambda_{HU}^{c_i}$ and summing over all contents $(\sum_{i=1}^{N}(\lambda_{HU}^{c_i}p_{c_i}^{lo}))$, we get the locally served HU rate. Dividing this over the HU request rate $(\sum_{i=1}^{N}\lambda_{HU}^{c_i})$ gives the probability of an HU getting service from

its local cache p_{local} in Equation 3.33.

$$p_{local} := \frac{\sum_{i=1}^{N}(\lambda_{HU}^{c_i} p_{c_i}^{lo})}{\sum_{i=1}^{N} \lambda_{HU}^{c_i}} \tag{3.33}$$

3.4.1. 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 $\lambda_{eff(HU)}^{sat}$ in Equation 3.25 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_{HU}^{sat} (contents fetched directly from the satellite). The HUs throughput in mode-ii is calculated similarly as $Th_{HU}^{sat(u)} := \lambda_{eff(HU)}^{sat(u)} \cdot s(\hat{v_b})$.

The mode-iii HU throughput Th_{HU}^{BS} is $\lambda_{eff(HU)}^{BS} \cdot (1 - p_{drop}^{BS}) \cdot s(\hat{v_b})$. For the effective service rate, dropped contents are excluded by 1-p_{drop}^{BS} since they do not contribute to successful transmissions. Multiplying mode-iii HUs arrival rate $\lambda_{eff(HU)}^{BS}$ in Equation 3.27 with 1-p_{drop}^{BS} gives the effective service rate of mode-iii HUs. The mode-iv HUs throughput $Th_{HU}^{BS(u)} := \lambda_{eff(HU)}^{BS(u)} \cdot (1 - p_{drop}^{BS}) \cdot s(\hat{v_b})$ and mode-v HUs throughput (D2D mode) $Th_{HU}^{D} := \lambda_{eff(HU)}^{D2D} \cdot (1 - p_{drop}^{D2D}) \cdot s(\hat{v_b})$ are calculated similarly.

For local hits, we look at the G_{HU}^{local} 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 $s(\hat{v_b})$.

$$G_{HU}^{local} := \lambda_{HU} \cdot p_{local} \cdot s(\hat{v_b}) \tag{3.34}$$

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

$$G_{HU} := G_{HU}^{local} + Th_{HU}^{sat} + Th_{HU}^{sat(u)} + Th_{HU}^{BS} + Th_{HU}^{BS(u)} + Th_{HU}^{D} \tag{3.35}$$

3.4.2. 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 Equation 3.37. It is calculated by the division of the overall consumed power P_{all} in Equation 3.36 over the overall system goodput of HUs in Equation 3.35.

$$P_{all} := P_{BS}^{tx} + P_{BS(u)}^{rec+tx} + P_{D2D}^{tx} + P_{loc} \tag{3.36}$$

$$EPB_{HU} := \frac{P_{all}}{G_{HU}} \tag{3.37}$$

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}^{tx}, *b)* $P_{BS(u)}^{rec+tx}$, *c)* P_{D2D}^{tx} and *d)* P_{loc}.

P_{BS}^{tx} in Equation 3.38 is the BS transmission power consumption for mode-*iii* HU services either *completed* or *dropped*. $\lambda_{eff(HU)}^{BS} \cdot (1 - p_{drop}^{BS})$ is the effective service rate of completed mode-*iii* HUs while the BS consumes $P_{BS}^{ch}/\mu_{HU}^{BS}$ 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.

$$P^{tx}_{BS} := (\lambda^{BS}_{eff(HU)} \cdot (1 - p^{BS}_{drop}) \cdot \frac{P^{ch}_{BS}}{\mu^{BS}_{HU}}) + (\lambda^{BS}_{eff(HU)} \cdot p^{BS}_{drop} \cdot \frac{P^{ch}_{BS}}{2 \cdot \mu^{BS}_{HU}}) \tag{3.38}$$

$P^{rec+tx}_{BS(u)}$ in Equation 3.39 is the BS power consumption for mode-*iv* HU services consisting of two service types: *a)* completed *b)* dropped. While calculating $P^{rec+tx}_{BS(u)}$, we consider additional cost imposed by the universal source integration, namely the BS reception energy. In type-*a* services, the multiplication of $\lambda^{BS(u)}_{eff(HU)} \cdot (1 - p^{BS}_{drop})$ and $\frac{P^{ch}_{BS}}{\theta_{BS}} \cdot \frac{s(\hat{v_b})}{C^{BS(u)}_{HU}}$ gives the BS reception power consumption for contents fetched from the universal source to the BS cache first and $\{\lambda^{BS(u)}_{eff(HU)} \cdot (1 - p^{BS}_{drop}) \cdot (\frac{P^{ch}_{BS}}{\mu^{BS}_{HU}})\}$ is the BS transmission power consumption for contents transmitted from the BS cache to HU requester devices. In type-*b* services, $\lambda^{BS(u)}_{eff(HU)} \cdot p^{BS}_{drop} \cdot (\frac{P^{ch}_{BS}}{\theta_{BS}} \cdot \frac{s(\hat{v_b})}{C^{BS(u)}_{HU}})$ is the BS reception power consumption for contents fetched from the universal source to the BS cache first. $\Delta^{BS(u)}_{HU}$ is the mean total service duration from the universal source across the BS to the requester HU as defined in (3.6). Assuming no bias, the duration of an incomplete service is in average half of a complete service, so $\frac{\Delta^{BS(u)}_{HU}}{2}$ is expected duration for an incomplete HU service from the universal source across the BS to the requester but dropped. $\frac{\Delta^{BS(u)}_{HU}}{2} - \frac{s(\hat{v_b})}{C^{BS(u)}_{HU}}$ is the duration of an incomplete HU service with content started to be transmitted from the BS cache but dropped. The multiplication of $P^{ch}_{BS} \cdot (\frac{\Delta^{BS(u)}_{HU}}{2} - \frac{s(\hat{v_b})}{C^{BS(u)}_{HU}})$ with $\lambda^{BS(u)}_{eff(HU)} \cdot p^{BS}_{drop}$ outputs the BS transmission power consumption starting from the transmissions of contents from the BS cache until they are dropped.

$$\begin{aligned} P^{rec+tx}_{BS(u)} := {} & \{\lambda^{BS(u)}_{eff(HU)} \cdot (1 - p^{BS}_{drop}) \cdot ([\frac{P^{ch}_{BS}}{\mu^{BS}_{HU}}] + [\frac{P^{ch}_{BS}/\theta_{BS}}{C^{BS(u)}_{HU}/s(\hat{v_b})}])\} \\ & + \{\lambda^{BS(u)}_{eff(HU)} \cdot p^{BS}_{drop} \cdot ([(P^{ch}_{BS} \cdot (\frac{\Delta^{BS(u)}_{HU}}{2} - \frac{s(\hat{v_b})}{C^{BS(u)}_{HU}})] + [\frac{P^{ch}_{BS}/\theta_{BS}}{C^{BS(u)}_{HU}/s(\hat{v_b})}])\} \end{aligned} \tag{3.39}$$

P^{tx}_{D2D} in Equation 3.40 is the transmission power consumption of HU devices operating in mode-v and is constructed with a similar logic of P^{tx}_{BS}.

$$P^{tx}_{D2D} := (\lambda^{D2D}_{eff(HU)} \cdot (1 - p^{D2D}_{drop}) \cdot \frac{P^{tx}_{dev}}{\mu^{D}_{HU}}) + (\lambda^{D2D}_{eff(HU)} \cdot p^{D2D}_{drop} \cdot \frac{P^{tx}_{dev}}{2 \cdot \mu^{D}_{HU}}) \tag{3.40}$$

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

3.5. Connectivity Mode Assignment

In our connectivity mode managing resource allocation strategy, we utilize system unit caches (the cache of requester device itself, satellite, BS and the caches of other HU devices in predetermined vicinity) and channel (satellite or terrestrial) states. Apart from that, the assignment of mode weight vector r:=[r_{sat}, r_{BS}, r_{dev}] is used to configure the connectivity mode selection rates. This vector r consists of r_{sat}, r_{BS} and r_{dev} elements designating the weight factor for each idle satellite mode operating frequencies (r_{sat}), for each idle BS mode operating frequencies (r_{BS}) and for D2D mode operating frequency as long as its maximum overlaying limit D_{max} is not reached (r_{dev}). A brief summary of our system model with mode details is provided in Figure 3.10.

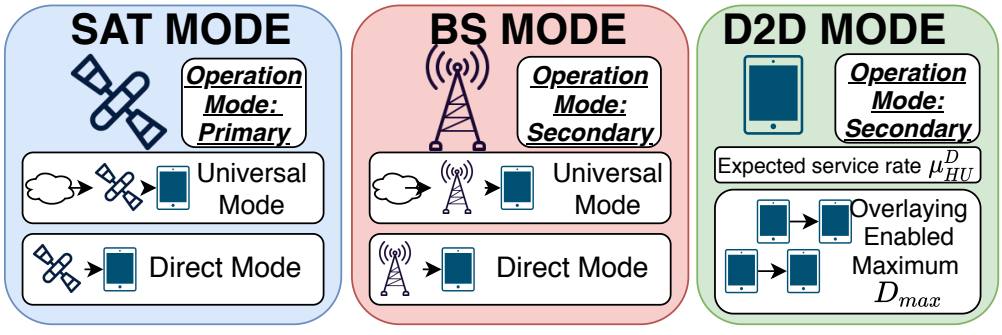

Figure 3.10. System model and different connectivity modes © 2021 IEEE [3].

The assigned values of r are used to configure the selection rate of different network modes which in turn determines the channel steady state probabilities, effective

arrival rates, dropping probabilities and local hit probability. Accordingly, $EPB_{HU}(r)$ and $G_{HU}(r)$ are determined as explained in Section 3.4. We find the sub-optimal assignment of r that decreases the energy consumed per successfully transmitted bit EPB_{HU} (the decrease in EPB_{HU} contributes to EE) while overall goodput G_{HU} is kept above a threshold. Our optimization problem is as follows:

$$
\begin{aligned}
&\min_{r} && EPB_{HU}(r) \\
&s.t. && G_{HU}^{threshold} \leq G_{HU}(r) \\
& && r_{sat} + r_{BS} + r_{dev} = 1 \\
& && 0 \leq r_x \leq 1, \qquad r_x \in \mathbb{R}, \quad x \in \{sat, BS, dev\}
\end{aligned}
$$

The space complexity for this problem is calculated by counting the number of all channel states provided in Figure 3.5. The upper bound of active satellite mode HU services are $N_{f_{sat}}^2$. β is the number of dedicated terrestrial frequencies for D2D mode with the terrestrial frequency set $\{f_1, ..., f_\beta\}$ operating in D2D mode only and hence $(D_{max})^\beta$ gives the maximum number of concurrent D2D services. The other terrestrial frequencies in the set $\{f_{\beta+1}, ... \ f_{N_{f_{ter}}}\}$ operate in BS mode only. At these terrestrial frequencies, the upper bound of active operations for both HU (BS mode) and PU services is $(N_{f_{ter}} - \beta)^3$. Hence, $\mathrm{O}((D_{max})^\beta (N_{f_{ter}} - \beta)^3 N_{f_{sat}}^2)$ is the space complexity of channel states. As we take $\beta = 1$, this complexity reduces to $\mathrm{O}(D_{max}(N_{f_{ter}} - 1)^3 N_{f_{sat}}^2)$. Solving the system of balance equations is $\mathrm{O}(n^3)$ with total n channel states.

In the determination of the optimization type for our problem definition, we inspect the linearity and convexity of our objective $EPB_{HU}(r)$. A function is said to be linear if it satisfies *i) superposition principle* and *ii) homogeneity*. We first recall the definitions of these two properties for a function $f : \mathbb{R}^N \to \mathbb{R}$ with $N \geq 1$. For $\forall x, y \in \mathbb{R}^N$ if $f(x + y) = f(x) + f(y)$ then *superposition principle* holds. In the homogeneity property $f(c \cdot x) = c \cdot f(x)$ $\forall x \in \mathbb{R}^N$ and $\forall c \in \mathbb{R}$. When we look at our objective, $EPB_{HU}(0 \cdot r) = 1.96$ nJbp $\neq 0 \cdot EPB_{HU}(r)$ for any r and thus property-*ii* is violated and our objective is non-linear. Next, we delve into the convexity examination.

A function $f : \mathbb{R}^N \rightarrow \mathbb{R}$ is said to be convex if for $\forall x, y \in \mathbb{R}^N$ and $\forall c \in [0, 1]$, $f(c \cdot x + (1-c) \cdot y) \leq c \cdot f(x) + (1-c) \cdot f(y)$ is satisfied. We investigate $r_1 = [0.3,\ 0.2,\ 0.5]$, $r_2 = [0.3,\ 0.3,\ 0.4]$ and $c = 0.2$ scenarios. In our system, we consider weight vector scenarios with the largest D2D mode weight r_{dev} (0.5 and 0.4) among all weights because our studies that are also output of this thesis [1, 2] have revealed that D2D operations are consuming low energy with considerably good channel conditions and hence service capacity. We consider scenarios with the satellite mode weight r_{sat} (0.3) as large as or larger than the BS mode weight r_{BS} (0.2 and 0.3) since the satellite is solar-powered. In these realistic scenarios with $r_1 = [0.3,\ 0.2,\ 0.5]$, $r_2 = [0.3,\ 0.3,\ 0.4]$ and $\lambda = 0.2$, our objective function EPB_{HU} does not satisfy the convexity condition with $EPB_{HU}(r_1) = 181.1$ nJpb and $EPB_{HU}(r_2) = 186.6$ nJpb. $EPB_{HU}(0.2 \cdot r_1 + (1-0.2) \cdot r_2)$ $= 185.57$ nJpb $\not\leq$ $0.2 \cdot EPB_{HU}(r_1) + (1-0.2) \cdot EPB_{HU}(r_2) = 185.50$ nJpb. By showing at least one instance not holding convexity property, our objective function is proven to be non-convex. Hence, our optimization problem is of type non-convex non-linear programming (NLP). It is NP-hard and hence to solve it in feasible time, we employ a heuristic to find a sub-optimal solution namely Pattern Search algorithm (*PSA*).

Table 3.12. Parameters for resource allocation schemes.

Parameter	**Explanation**	**Value**
$G_{HU}^{threshold}$	The minimum overall system goodput required from the network	48 Mbps
T_{const}	Tolerance on constraint G_{HU}	$1e^{-8}$
T_{obj}	Tolerance on objective function EPB_{HU}	$1e^{-9}$
T_{mesh}	Tolerance on mesh size	$1e^{-7}$
s_{in}	Initial mesh size	0.02
α_c	Contraction factor	0.5
α_e	Expansion factor	2

PSA is a derivative-free (black-box) search method [102]. We analyze the constraint function *overall goodput* G_{HU}. For $r = [0.3,\ 0.2,\ 0.5]$ and $c = 0.1$, we notice the violation of homogeneity property of linearity for the constraint since $G_{HU}(c \cdot r) = G_{HU}(0.1 \cdot [0.3,\ 0.2,\ 0.5]) = 47.89$ Mbps $\neq 0.1 \cdot G_{HU}([0.3,\ 0.2,\ 0.5]) = 0.1 \cdot 47.89$ Mbps. Due to the non-linear constraint, we employ Augmented Lagrangian PSA for our non-linearly constraint problem [103]. PSA uses polling mechanism as given in Figure 3.11 to create an assignment mode weight vector sequence $\{r_x\}$ for converging to a lower

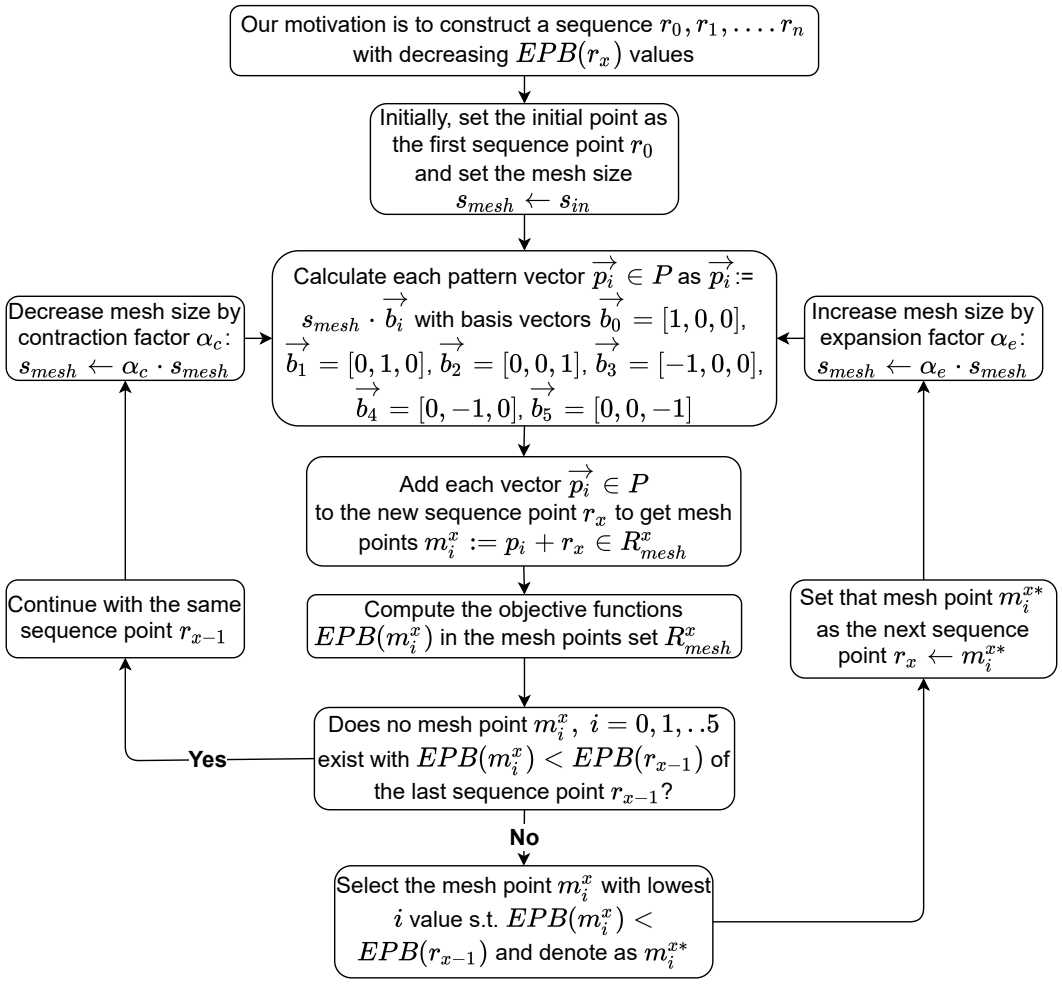

Figure 3.11. PSA polling mechanism.

value of objective EPB_{HU}. In polling, the PSA computes $EPB_{HU}(r_x)$ values at each iteration. In a successful poll, any mesh vector r_x improving the objective $EPB_{HU}(r_x)$ is inserted to the sequence $\{r_0, r_1, ..., r_{x-1}\}$ and for the next iteration the mesh size is increased by the multiplicative expansion factor α_e. In an unsuccessful poll, no improvement is observed so the sequence remains the same and for the next iteration the mesh size is decreased by the multiplicative contraction factor α_c. PSA algorithm stops when changes in objective, constraint or mesh size less than specified tolerance values is observed. The tolerance values and other PSA parameters are provided in Table 3.12.

PSA does not guarantee the global minimum. However, starting runs from different initial points, its solution can be improved. In this context, we use all possible on/off settings in different modes for the initial points as illustrated in Figure 3.12.

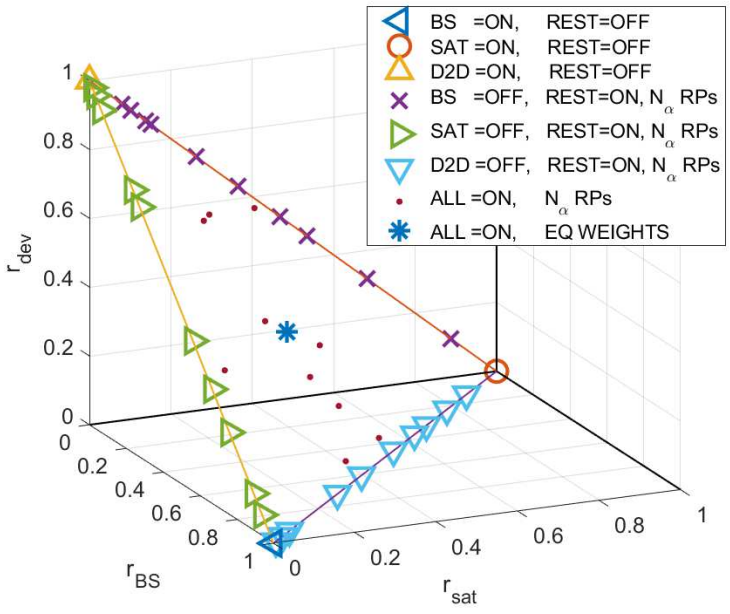

Figure 3.12. Set of initial points for PSA $[r_{sat}, r_{BS}, r_{dev}]$ (RP: random point, EQ: equal) © 2021 IEEE [3].

The corners of the triangle in Figure 3.12 (the first three cases in the legend) show the initial point settings with only one mode ON. When we look at the lines (4^{th} to 6^{th} cases in the legend) drawn between these points , we have the set of initial points with one mode OFF only. For instance, the yellow line drawn between "only D2D ON" and "only BS ON" settings depicts the scenario of "only satellite OFF" (5^{th} case in the legend). N_α many random points on this line are selected to start the PSA with satellite inactive but other modes active settings. Finally, we draw N_α many random initial points depicted as dots in Figure 3.12 (7^{th} case in the legend) from the surface bounded by the three lines. These points surveil all modes ON settings. As a special initial point setting, we utilize for each mode equal weight point (8^{th} case in the legend).

3.6. 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 that is illustrated in Figure 3.13. 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 3.3.1. 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 Equations 3.13-3.15 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 3.13. 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 3.6.4, the results for varying D2D mode weight are presented.

3.6.1. Caching Dynamics and Popularity-Driven Caching (PDC)

We compare the popularity-driven caching (PDC) to the baseline random caching. We assume that universal source is on, overlaying mechanism for D2D operation mode is enabled and all mode weights r_{sat}, r_{BS}, r_{dev} are assigned to 1/3. With increasing

Table 3.13. Simulation parameters and values for the multi-mode HetNet.

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
$N_{f_{sat}}$	The total number of satellite frequencies	2
$N_{f_{ter}}$	The total number of terrestrial frequencies	3
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
D_{max}	The maximum concurrent D2D operations allowed by the network	5
N	Total number of contents	20
α	Zipf parameter	1.2
λ_{HU}	The mean content request rate of hybrid users	2.4 $\frac{user}{sec}$
λ_{PU}^{ter}	The mean arrival rate of primary users at terrestrial link	0.03 $\frac{user}{sec}$
C_{Sat}^{cache}	The satellite cache capacity	125 Mbs
C_{BS}^{cache}	The base station cache capacity	100 Mbs
C_{Dev}^{cache}	The HU device cache capacity	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

arrival rate of HUs for content request λ_{HU}, EPB_{HU} is decreased (leading to improvement in EE) for both caching techniques as provided in Figure 3.14. With increasing λ_{HU}, the network gets more crowded and therefore the probability of satellite channel or BS mode operable terrestrial frequencies being idle declines (since these services durate longer than D2D mode operations). Consequently, D2D mode utilization boosts and the reduction of EPB_{HU} is observed. Besides, we also monitor no obvious difference in EPB_{HU} for two different caching methods in any λ_{HU} rate. When we analyze the G_{HU}, as expected it increases with increasing λ_{HU} in both caching mechanisms as shown in Figure 3.14. G_{HU} results for PDC are greater than that of random caching especially for larger λ_{HU}'s since the PDC technique has preference to cache more popular contents with a greater probability and thus leading to greater local service rate. The simulation results follow the same trend with the analytical EPB_{HU} and G_{HU} values in both caching mechanisms.

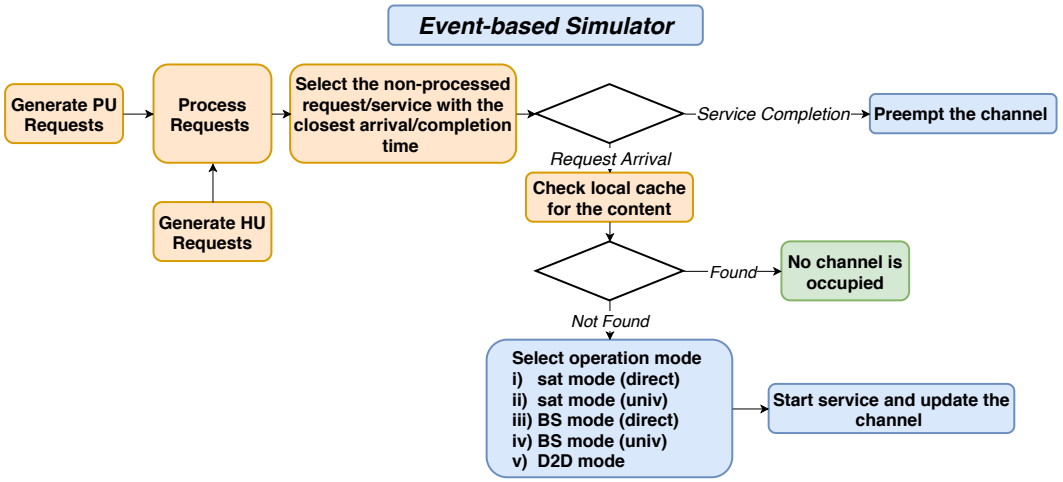

Figure 3.13. Simulation mechanism for the multi-mode HetNet.

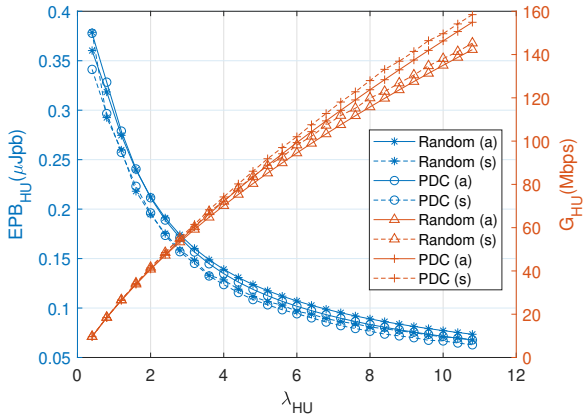

Figure 3.14. EE and goodput results (a: analytical, s: simulation) [1, 2].

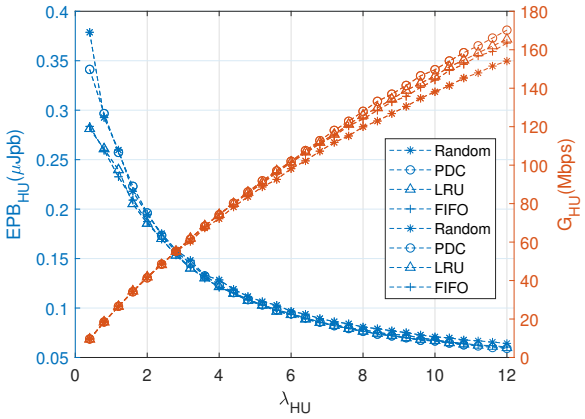

Figure 3.15. Simulation EE and goodput results [2].

For large λ_{HU}'s in Figure 3.14, the simulation G_{HU} results are larger than corresponding analytical results in both caching techniques. The reason is that the impact of D2D mode services is greater for large λ_{HU} and $p_{c_i}^{D(f_1)}(x)$ is a lower bound for the D2D content availability probability of any content c_i as explained in Subsection 3.3.2. This is utilized in the analytical calculation but not in the simulations. Hence, we are not restricted by this lower bound in the simulations and we obtain more precise and larger G_{HU} values.

For a more thorough analysis, we compare simulation results of PDC mechanism with simulation results of caching methodologies *Least Recently Used (LRU)*, *First In*

First Out (FIFO) in addition to random caching in Figure 3.15. Again, we assume that universal source is on, overlaying for D2D mode operation is enabled and all mode weights r_{sat}, r_{BS} and r_{dev} are equal to 1/3. Random and PDC are time-independent caching strategies that do not depend on the caching time of contents. On the contrary, LRU and FIFO need the caching time during the decision phase of eviction from the cache for the sake of recently received content. With increasing λ_{HU}, EPB_{HU} decreases and G_{HU} increases in all techniques. For λ_{HU} values less than or equal to 2 $^{user}/_{sec}$, the time-dependent caching strategies (LRU and FIFO) have lower EPB_{HU} results compared to time-independent ones. For larger λ_{HU}, the gap of EPB_{HU} for time-dependent and -independent caching strategies disappears. In Figure 3.15, the G_{HU} results for PDC, LRU and FIFO caching are greater than random caching especially for larger λ_{HU}'s. All these three caching strategies (LRU, FIFO, PDC) make use of content information (LRU: time of a request for some content, FIFO: arrival time of content to some system unit cache, PDC: popularity distribution of contents) in order to cache contents efficiently. Therefore, they outperform random caching in terms of G_{HU} especially under high request rate λ_{HU}. PDC algorithm attains slightly better G_{HU} results than that of LRU and FIFO for larger λ_{HU} (For $\lambda_{HU} = 12^{user}/_{sec}$ G_{HU} of FIFO is 163.5 Mbps, G_{HU} of LRU is 165.5 Mbps while G_{HU} of PDC is 170.2 Mbps). As a consequence, the utilization of global content popularity distribution in PDC outperforms local decisions only considering content request or arrival history in LRU or FIFO techniques respectively.

3.6.2. 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

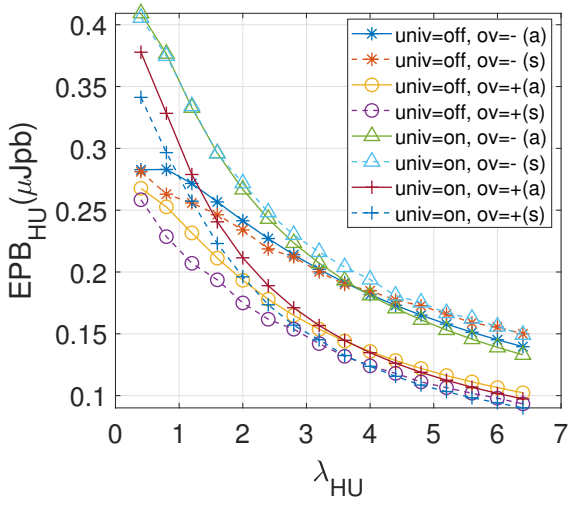

Figure 3.16. EE results (a: analytical, s: simulation, -: disabled, +: enabled, ov: overlay) [1, 2].

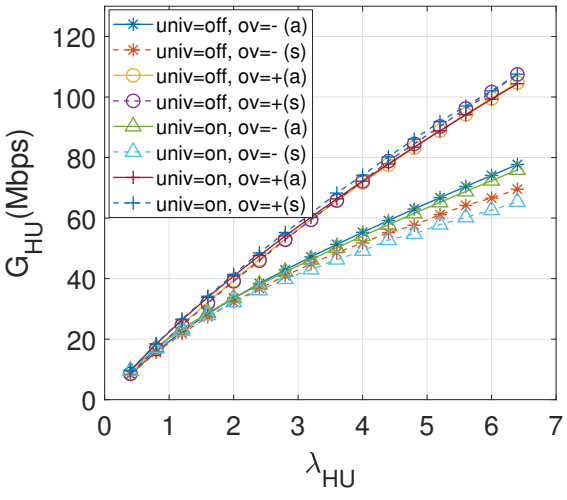

Figure 3.17. Goodput results(a: analytical, s: simulation, -: disabled, +: enabled, ov: overlay) [1, 2].

out the effect of different system unit weighting (i.e., no favored transmission mode) to specifically focus on universal source and overlaying mechanisms. Apparently, for all settings (universal source on/off, overlaying enabled/disabled) EPB_{HU} decreases (Figure 3.16) and G_{HU} increases (Figure 3.17) 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 Figure 3.16. 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 Figure 3.16, 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 Figure 3.17. 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 frequen-

cies 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 balance 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 Figure 3.17.

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. Therefore, 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.

3.6.3. 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 Figure 3.18. 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] $^{\text{user}}/_{\text{sec}}$

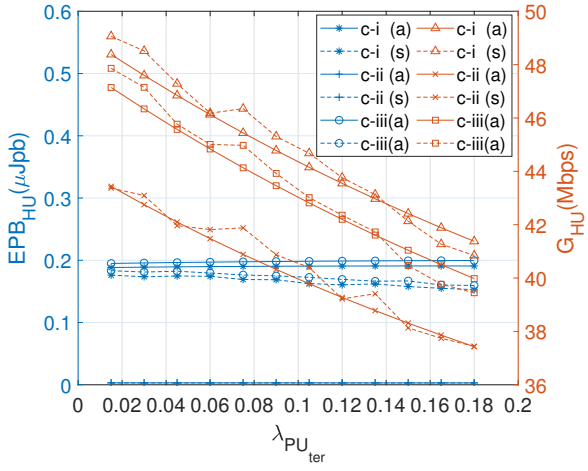

Figure 3.18. EE and goodput results for varying PU arrivals in terrestrial link (a: analytical, s: simulation, c: constellation) [1, 2].

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*) 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 Figure 3.18, 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 Figure 3.18, *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 [97]. Therefore, 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 [97], 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.

3.6.4. 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} (Figure 3.19) and G_{HU} (Figure 3.20) 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 Figure 3.19. 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 analyti-

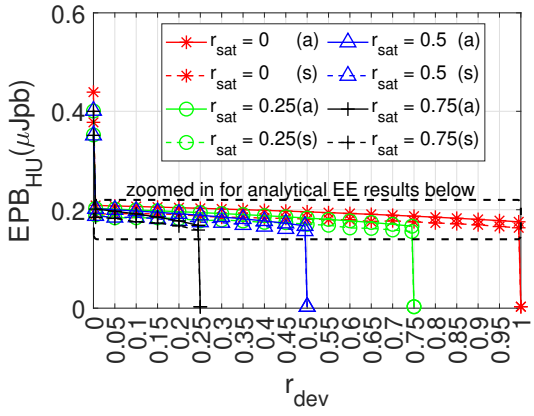

Figure 3.19. EE results for varying r_{dev} where r_{sat} is fixed ($r_{dev} = 1 - r_{sat} - r_{BS}$, a: analytical, s: simulation) [1, 2].

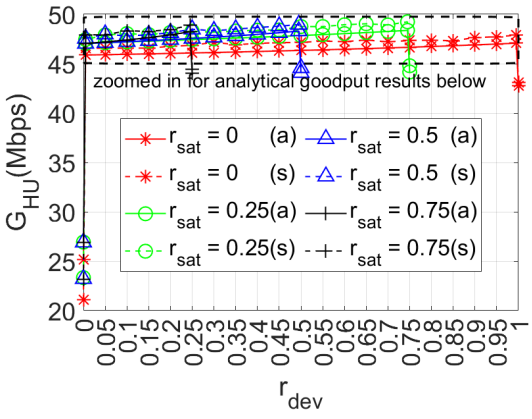

Figure 3.20. Goodput results for varying r_{dev} where r_{sat} is fixed [1, 2].

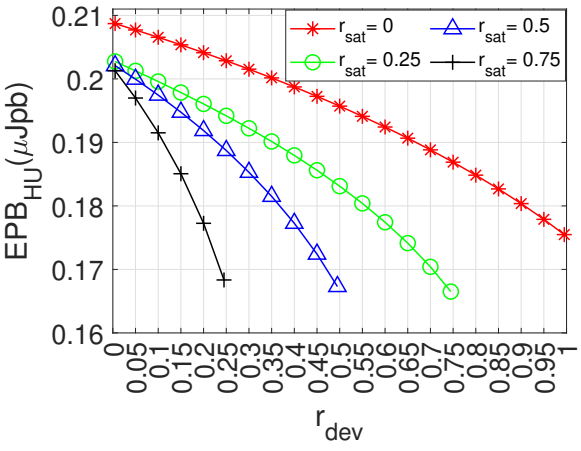

Figure 3.21. Analytical EE results [1, 2].

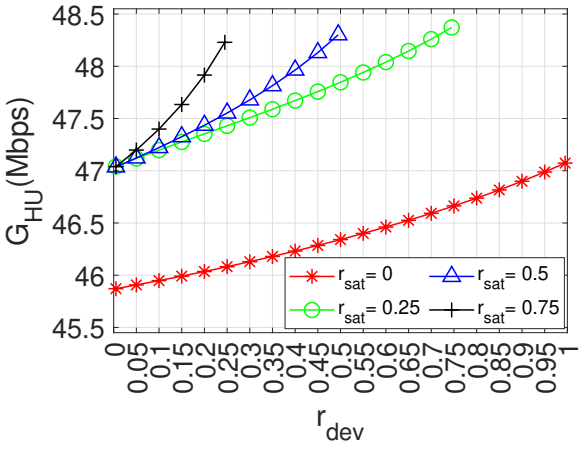

Figure 3.22. Analytical goodput results [1, 2].

cally.) as shown in Figure 3.20. 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 Figure 3.20.

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 Figure 3.21, 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 Figure 3.21 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 ($P_{dev}^{tx} < P_{BS}^{ch}$) 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 Figure 3.22. 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 Figure 3.22. 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.

3.6.5. PSA

We look at the sub-optimal assignment of $\mathbf{r}$ by PSA. We assume the universal source is on and the D2D overlaying is enabled, $N_{f_{sat}} = 2$ and $N_{f_{ter}} = 3$. The simulation parameter list is provided in Table 3.14. The sub-optimal assignment of $\mathbf{r}$ by PSA is $r_{PSA} = [0.09 + \epsilon_1, \epsilon_2 - \epsilon_1, 0.91 - \epsilon_2]$ where $0 < \epsilon_i$ for $i = 1, 2$, $\epsilon_i \to 0$ such that $\epsilon_2 > \epsilon_1$. The objective of PSA is $EPB_{HU} = 0.166$ μJpb and the constraint $G_{HU} = 48.43 Mbps$. The convergence of a mode weight to zero does not necessarily result in no selection of the corresponding mode. For r_{PSA}, even the BS mode weight converges to zero the corresponding mode selection rate is 15.7%. All mode selection rates of several assignment configurations among non-blocked HU services are provided in Table 3.15.

Table 3.14. Simulation parameters for PSA.

Par.	**Explanation**	**Value**
$N_{f_{sat}}$	The total number of satellite frequencies	2
$N_{f_{ter}}$	The total number of terrestrial frequencies	3
λ_{HU}	The mean content request rate of hybrid users	2.4 $\frac{user}{sec}$
P_{sat}^{ch}	Per channel transmission power of the satellite	48 W
P_{BS}^{ch}	Per channel transmission power of the BS	6 W
P_{dev}^{tx}	The transmission power of a hybrid user device	80 mW
d_{sat}	The distance from LEO satellite to earth	300 km
d_{BS}	Mean distance of an HU to the BS	150 m
d_{D2D}	Mean distance between receiver and sender HU devices	30 m
D_{max}	The maximum number of concurrent D2D operations that is allowed by the network	5
N_{α}	The number of random initial points for PSA in only one mode OFF scenarios and for all modes ON scenarios	10

Table 3.15. Energy efficiency results and mode selection rates.

Configuration	**EPB$_{\mathbf{HU}}$**	**Sat mode**	**BS mode**	**D2D mode**
i) *PSA*	*0.166μJpb*	5.4%	*15.7%*	78.9%
ii) EQ_{ALL}	0.189μJpb	6.6%	18.4%	75.1%
iii) ACT_{ALL}	0.192μJpb	6.6%	18.7%	74.7%
iv) *CAP* [104]	0.352μJpb	24.6%	75.4%	-
v) CAP_{EX}	0.177μJpb	5.5%	16.9%	77.6%
vi) $LQ(T_d$=0.1) [105]	0.176μJpb	-	17.0%	83.0%
vii)$LQ_{EX}(T_d$=0.1)	0.171μJpb	4.23%	16.23%	79.54%

We explore three mode availability options for the PSA assignment:

(i) *BS mode is an option and other option(s) (satellite, D2D mode) are also available*: The BS mode weight r_{BS} is smaller than others and hence this mode is selected with a low probability. The satellite is solar-powered and D2D mode operations consume less energy than BS mode (due to lower power level and shorter service duration in D2D mode compared to the BS mode) and thereof decreasing the BS selection improves EE with decreased EPB_{HU}.

(ii) *BS mode is the only option*: For serving requester HUs (req-HUs), BS mode (directly or indirectly) is used. We utilize the BS cache and corresponding terrestrial frequencies, which is a contributing factor for keeping G_{HU} above the designated threshold.

(iii) *BS mode is not an option*: Either due to the requested content unavailability in the BS cache and/or the terrestrial frequencies that can be used for BS operation are not idle. Then this scenario forks into three availability options: *(a)Only satellite mode. (b)Only D2D mode. (c)Both satellite and D2D mode*: In option *(c)* r_{sat} and r_{dev} are content traffic determinants. According to PSA result, r_{sat} attains lower value than r_{dev}. Although HU energy consumption in D2D mode is greater than in the satellite mode (satellite is solar powered but HU devices consume $\frac{P_{dev}^{tx}}{\mu_{HU}^{D}}$ energy per successful and on average $\frac{P_{dev}^{tx}}{(2\cdot\mu_{HU}^{D})}$ per unsuccessful service),

the HU service durations over the satellite link are long and thereof, the satellite link saturates. So PSA returns larger r_{dev} (0.91-ϵ_2) compared to r_{sat} (0.09+ϵ_1).

We investigate EE of the PSA technique by a comprehensive comparison to different assignment configurations of types: *i)* baseline, *ii)*existing works in the literature. The investigation results are listed in Table 3.15. The baseline assignments are listed as follows:

(i) EQ_{ALL}: has equal weights $[r_{sat}, r_{BS}, r_{dev}] = [1/3, 1/3, 1/3]$.

(ii) ACT_{ALL}: the average of N_α random assignments with all modes on $[r_{sat}, r_{BS}, r_{dev}]$ $= [\gamma^i_{sat}, \gamma^i_{BS}, 1 - \gamma^i_{sat} - \gamma^i_{BS}]$ where $1 > \gamma^i_{sat} > 0$, $1 > \gamma^i_{BS} > 0$, $i \in \{1, 2, ..., N_\alpha\}$.

The assignments from the existing works in the literature are listed below:

(i) CAP [104]: It chooses the flow through the satellite/cellular links by the ratio of the corresponding link capacity over the total capacity.

(ii) CAP_{EX}: It is an extended version of the CAP with D2D mechanism and mode selection is updated accordingly.

(iii) LQ [105]: The quality of the service link affects the service capacity in wireless networks. To this end, better quality link selection is the main objective of this assignment and accordingly D2D mode is selected if $T_d \cdot d^{-n_{D2D}}_{t:r(D2D)} > d^{-n_{BS}}_{t:r(BS)}$ and BS mode is selected otherwise. Note than $d_{t:r(D2D)}$ and $d_{t:r(BS)}$ denote the distance between the transmitter and receiver in the D2D and BS modes, respectively. n_{D2D} and n_{BS} are the corresponding path loss exponents. Here T_d is the BS to D2D mode offloading factor to tune the selection mechanism.

(iv) LQ_{EX}: It is an extended version of the LQ with the satellite mode. It utilizes the satellite only when other modes are not available (due to cache/channel conditions). The reason is that the unstable nature of the satellite link renders poor link quality compared to the terrestrial counterparts.

In terms of EE, the sub-optimal assignment PSA achieves EE improvement with 13.8% lower EPB_{HU} (0.166 μJpb) over the average EPB_{HU} value (0.192 μJpb) of

ACT_{ALL}. Similarly, PSA assignment has EPB_{HU} better than EQ_{ALL} counterpart in terms of EE: 0.166 μJpb vs. 0.189 μJpb, a 12.4% lower figure. CAP assignment does not operate in D2D mode and therefore it is very poor in terms of EE with more than double EPB_{HU} value of the PSA (0.352μJpb vs. 0.166 μJpb). When we monitor its extended version CAP_{EX}, the PSA achieves 6.5% improvement over the CAP_{EX} ($EPB_{HU} = 0.177$ μJpb). PSA also has 6.0% lower EPB_{HU} value (0.166 μJpb) than LQ with D2D offload factor $T_d = 0.1$ (0.176 μJpb). Thus, the sub-optimal assignment PSA is more energy-efficient than LQ. With the satellite extension of the LQ assignment (LQ_{EX}), the difference in terms of EE reduces between PSA and LQ_{EX}. However, the PSA has still slightly lower EPB_{HU} figure. In a nutshell, our sub-optimal assignment scheme PSA outperforms all of these aformentioned assignments in terms of EE.

According to the PSA scheme, the highest mode weight is assigned to the D2D communication. Even though the satellite is solar-powered, a key bottleneck is the rapid saturation problem of the link and accordingly D2D mode selection is shown to be more beneficial in terms of EE. Furthermore, cellular services require high energy consumption because of the large BS transmission power. Consequently, BS mode selection occurs when it is the only option available. Otherwise, BS mode is mostly ignored for the sake of improving system-wise EE. D2D communication improves EE due to its low power consumption and high speed transmission enabling greater channel rates. Hence, sub-optimal mode management scheme PSA verifies the importance of D2D systems with the highest weight assigned to the D2D communication mode.

Satellite and cellular communications are also utilized if D2D mode is not used due to content or channel unavailability. They are more successful in keeping contents for services with their larger cache capacities compared to end-user devices and with their support, overall goodput is kept still at an acceptable level. Based on this phenomenon, the trade-off between EE and goodput is illustrated.

3.6.6. Discussion

In this chapter, we have modeled a satellite integrated cognitive and D2D multi-mode HetNet and analyzed it from the multimedia service aspect rigorously. By the integration of the universal source concept, under highly loaded HetNet traffic, EE of the system is preserved and moreover the content availability is improved by allowing access to the exterior content servers. According to the mode selection analysis, EE is boosted by increasing D2D mode weight. Furthermore, both regarding the EE and system capacity, a remarkable improvement is observed by enabling the D2D overlaying mechanism with controlled interference. Finally, we developed a PSA method in such multi-mode natured HetNets for improving network-wise EE constrained by the capacity and observed that PSA achieves the best EE with the greatest D2D mode selection rate rather than other modes. Besides, PSA outperforms the EQ_{ALL}, ACT_{ALL}, CAP, CAP_{EX}, LQ and LQ_{EX} approaches in terms of EE. In a nutshell, D2D mechanism is an instrumental approach for EE improvement in HetNet systems. In the future with the advent of 6G technologies, such HetNets will include mega satellite constellations such as Starlink, Hongyan etc. as well. In that regard, the state space complexity of the analytical analysis will be a limitation for a compound analysis of 6G systems with D2D services and mega satellite constellations. To overcome this issue, the satellites can be grouped by their distances to the earth, traffic load, congestion rate etc. and each group can be exemplified by one instance within the HetNet state diagram for tractability.

4. CONTENT MODELING AND CACHING

In this chapter, we develop a popularity, chunking and layering based video content model for content-centric D2D edge networks. To the best of our knowledge, our work is the first proposal that models video contents according to all these dimensions — especially from the perspective of caching in D2D networks. Additionally, based on our novel content model, we propose caching algorithms via prioritization on content attributes in such systems. We also investigate the impact of caching on the energy consumption, goodput and energy efficiency. Some part of our studies in this chapter are presented and available in the conference proceedings [4].

4.1. System Model

In this section, we consider the wireless nodes in the network edge exchanging content via D2D communications in an infrastructure-independent manner [106]. This architectural layout refers to emerging mobile edge computing scenarios such as augmented reality (AR) and edge-accelerated content streaming. Devices in this network setting need to be protected against excessive energy consumption due to video traffic while enjoying very high bitrates. In that regard, we focus on video content modeling and caching in these ad hoc D2D networks.

In our system, users are dispersed in the spatial domain without access to a base station for content delivery. For modeling the user locations, Poisson Point Process (PPP) is a commonly utilized spatial distribution [107]. In our network, users are distributed according to PPP with mean density λ_{users}. They have devices with storage that is capable of storing contents. These devices can exchange video content with each other via D2D communications. When a content is requested, first the requester will check its local cache. If the content is not found, it will try to use D2D transmissions. It will fetch the requested content from the closest accessible device that stores that content. All users have equal priority while accessing the wireless medium. For the

D2D wireless channel, the employed pathloss model for a given distance d is:

$$P_r(d) = P_{D2D}^{tx} - 20 \cdot log_{10}(\frac{4\pi f d_0}{c}) - 10 \cdot log_{10}(\frac{d}{d_0})^{n_{D2D}} \tag{4.1}$$

where P_{D2D} is the transmit power of a device and $P_r(d)$ received power, n_{D2D} is the path loss exponent of D2D transmission and d_0 is a reference distance of the device antenna. The D2D channel capacity for service is calculated by $C_{D2D} = B \cdot log_2(1 + \frac{P_r(d)}{B \cdot N_0})$ where B is the bandwidth and N_0 is the noise power density.

4.1.1. Video Content Model

To explain the rationale behind our three dimensional (popularity, chunking and layering) video content model, we first branch into these dimensions and describe them:

- *Popularity:* Popularity is a key content attribute that is used to optimize caching according to content request characteristics. The emergence of content-centric networking requires popularity profiling of contents. In the literature, the Zipf distribution $Zipf(\alpha, N)$ is widely used for generic modeling of content requests [97, 108]. Here, N stands for the total number of contents in the system while α determines the skewness of the distribution.
- *Chunking:* The partitioning of contents into chunks infuses link bandwidth gain [57]. The chunking also leverages the caching gain [109]. Besides, it is a practical strategy for designing simpler caching schemes and enabling differentiation among different parts of a content. In that regard, it is beneficial to be utilized in the content model.
- *Layering:* In scalable coding, the base layer is the standard quality (SQ) video segment while enhancement layers improve the video quality [110]. The upper layers require low quality layer portions for successful decoding. Scalable video coding provides adaptability for different network conditions [111] such as congestion or packet loss. Thus, it is integrated into our content model.

Table 4.1. Video sequences utilized for determining characteristics [6].

Video	**Genre**
Citizen Kane	Drama
Silence of the Lambs	Drama
Jurassic Park I	Action
Die Hard I	Action
The Terminator I	Action
Total Recall	Action
Star Wars IV	Sci-fi
Star Wars V	Sci-fi
Aladdin	Cartoon
Cinderella	Cartoon
The Firm	Drama
Tonight Show	Late Night Show
Baseball	Game 7 of the 2001 World Series
Snowboarding	Snowboarding Competition

For empirically determining the video characteristics in our content model, we utilize 60 minutes long quarter common intermediate format (QCIF) formatted temporal scalable encoded videos [6] listed in Table 4.1. $IBBPBBPBBPBBPBB...$ is the group of picture (GoP) structure of these videos with frame rate 30 fps. In [6], layering dimension is used where the trace statistics of temporal scalable encoded videos are provided. I and P frames constitute the base layer while B frames form the enhancement layer. The calculation of mean video frame sizes of base and enhancement layer $\overline{X^{\gamma}}$ is given in Equation 4.2 [112]. X_n is the size of the n^{th} frame for $n = 0, 1, \cdots, N_{frame} - 1$ while X_n^b=X_n for I and P (base layer) frames and X_n^b is zero for B (enhancement layer) frames. On the contrary, X_n^e is zero for base layer frames and X_n^e=X_n for enhancement layer frames.

$$\overline{X^{\gamma}} = \frac{1}{N_{frame}} \sum_{n=0}^{N_{frame}-1} X_n^{\gamma} \quad , \qquad \gamma \in \{b, e\}. \tag{4.2}$$

For video coding, our sample videos consist of frames partitioned into 8×8 sample blocks of luminance, hue and intensity and all of them are mapped to 8×8 transform

coefficient blocks via discrete cosine transform (DCT). These blocks are quantized based on a quantization scale where low scale means higher quality and high scale entails lower quality in [6]. In our work, we consider 10, 14 and 16 as quantization scales for I, P and B frames, respectively. For the given quantization scale, the mean base frame size $\overline{X^b}$ is 0.3727 kB while the mean enhancement frame size $\overline{X^e}$ is 0.176 kB [6]. The average size of 60 minutes long stardard quality (SQ) videos of frame rate 30 Hz $\overline{s_{SQ}}$ is then calculated as:

$$\overline{s_{SQ}} := 3600\ s \cdot 30\ \text{Hz} \cdot \overline{X^b}\ kB \cdot 8 \cdot 10^3\ \frac{bits}{kB} \tag{4.3}$$

Accordingly the average size of high quality (HQ) videos (60 minutes long, frame rate 30 Hz) $\overline{s_{HQ}}$ is calculated using the value $\overline{s_{SQ}}$ and the additional enhancement layer contribution as shown below:

$$\overline{s_{HQ}} := \overline{s_{SQ}} + (3600\ s \cdot 30\ \text{Hz} \cdot \overline{X^e}\ kB \cdot 8 \cdot 10^3\ \frac{bits}{kB}) \tag{4.4}$$

Then the average sizes according to employed video sequences are $\overline{s_{SQ}} = 322$ Mb and $\overline{s_{HQ}} = 474$ Mb.

Popularity dimension reveals the essence of content request characteristics. The Zipf distribution $Zipf(\alpha, N)$ is applicable for generic modelling of content request characteristics. We will investigate how Zipf exponent characterizing parameter α affects the caching mechanisms in the Section 4.4.

The ratio of high quality consumers is denoted as p_{HQ} and $p_{HQ} \in [0, 1]$. For the inspection of layering dimension, we will look for the impact on the caching mechanisms of different p_{HQ} values in the Section 4.4.

[67, 113] benefit fixed size content chunking for caching. In particular in [109] it is stated that the simple homogeneous (equal sized) content partitioning is sufficient to gain benefits in caching. Thus, we employ equipartioning, i.e. partition the contents

into equal sizes as the chunking technique [67, 109] and take the base chunk size as 16 Mbits. The request characteristic of chunks is also required to be studied. [109] utilizes the $Weibull(\lambda, k)$ distribution to model the request characteristics of content chunks with λ and k the scale and shape parameters, respectively. We will portrait the effect of Weibull distribution on the caching algorithms in the forthcoming Section 4.4.

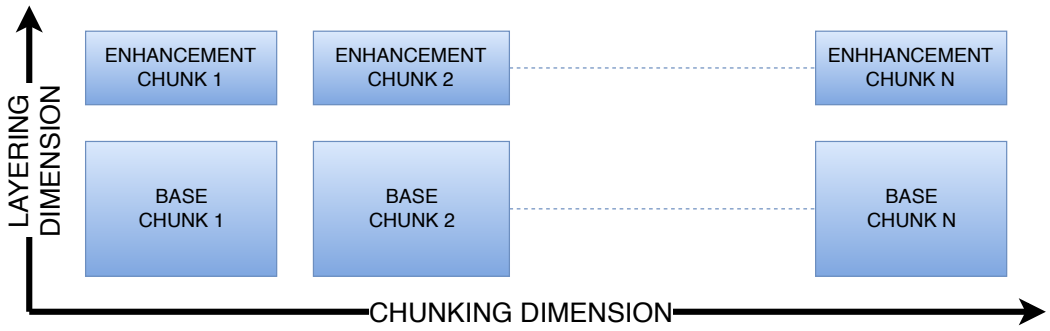

Figure 4.1. Layer and chunk dimensions of the video content model © 2019 IEEE [4].

The two layers in our video content model with chunking are shown in Figure 4.1. We assume two video layers with equal sized chunks in these layers.

4.2. Multidimensional Caching Schemes for D2D Edge Networks

Now, we propose content based caching algorithms. Before focusing on the caching proposals, we look at the basis of our caching that is the content model. Given in Figure 4.1, contents are partitioned into chunks each consisting of two distinct layers: *i)* base *ii)* enhancement. In our scheme, each content *unit* is uniquely defined by the popularity, chunk and layer order.

Some content parts are more useful in terms of amplifying the caching gains. In that regard, the utilization of content popularity differentiation in caching is applicable. For instance, Suksomboon et al. propose *PopCache* in content-centric networks which stores popular contents close to the requesters [67]. However, apart from the inter-content patterns such as relative popularity, intra-content features are promising as some content portions are more beneficial in terms of caching gains. For instance, from the layering aspect, the base is more essential than enhancement layer(s) since videos cannot be rendered and displayed when it is unavailable [111, 114]. When we delve into

the chunking analysis, the initial chunks are more worthy owing to the intra-content decreasing request patterns with the initial chunks demanded more frequently (e.g., the beginning segments of a video compared to the end ones) [115, 116]. Based on these content related observations, we develop our prioritized content caching techniques.

Our proposals are constructed with the aim of preserving "important" content segments (layer/chunk) in caches for improving the system performance. Our caching algorithms are designated as follows: *i)* Layer Prioritized Popularity Based Caching ($LPPC$), *ii)* Chunk Prioritized Popularity Based Caching ($CPPC$). The proposed algorithms $CPPC$ and $LPPC$ are provided in Figure 4.2 and 4.4, respectively. We provide an example scenario of the $LPPC$ algorithm in Figure 4.3 as well.

In both caching algorithms $CPPC$ and $LPPC$ when the newly requested content unit c_{new} cannot fit into the cache of capacity C due to cache capacity limitations, some content unit(s) need to be evicted for storing the newly requested unit c_{new}. The set of content units to be dismissed are decided as follows: First the content units in the cache S_c are sorted based on the content dimensions. Starting from the lowest order the content units are discarded from the cache, until the free cache capacity is sufficient to store the new unit.

Both in $LPPC$ and $CPPC$ techniques, contents are first sorted on the content popularity dimension in descending order. This way, both algorithms give the highest priority to content popularity attribute. Next, the caching mechanisms focus on either on the chunk or layer order. $LPPC$ first sorts on layer dimension of a given content. For breaking ties among the same layer of the same content, it sorts on the chunk order. Thus, layering dimension dominates chunking dimension in $LPPC$ algorithm. On the contrary, $CPPC$ first sorts on the chunking dimension from the initial chunk to final one of any given content and then sorts on the layering dimension for breaking ties. In contrast to $LPPC$, $CPPC$ technique assigns greater importance to the chunking dimension compared to layering.

INPUTS

S_c: The set of content units in the cache

c_{new}: The newly requested content unit

C: The cache capacity

LPPC(S_c, c_{new}, C){

$Cur_{Cap} = Capacity(S_c)$;

if *(*$Cur_{Cap} + size(c_{new}) \leq C$*)* **then**

  return $S_c \cup \{c_{new}\}$;

else

  $S_{sorted} \leftarrow$ sort(S_c, POP);

  $S_{sorted} \leftarrow$ sort(S_{sorted}, LAYER);

  $S_{sorted} \leftarrow$ sort(S_{sorted}, CHUNK);

  //$S_{sorted} = \{s_1,\ s_2,\ \ldots,\ s_k\}$ ordered from s_1 to s_k

  j = k;

  while *(*$j \geq 1$*)* **do**

    $S_{sorted} \leftarrow S_{sorted} \setminus \{s_j, s_{j+1}, \ldots s_k\}$;

    if *(*$Cur_{Cap} + size(c_{new}) - \sum_{\theta=j}^{k} size(c_\theta) \leq C$*)* **then**

      return $S_{sorted} \cup \{c_{new}\}$;

    end

    $j \leftarrow j - 1$;

  end

end

}

Figure 4.2. Layer prioritized popularity based caching (LPPC) © 2019 IEEE [4].

By the contrasting domination factor of content model dimensions in caching techniques, we look at the performance analysis of our network and analyze how the different prioritization of dimensions affect the results. We also compare our proposals with standard caching technique Least Recently Used (LRU).

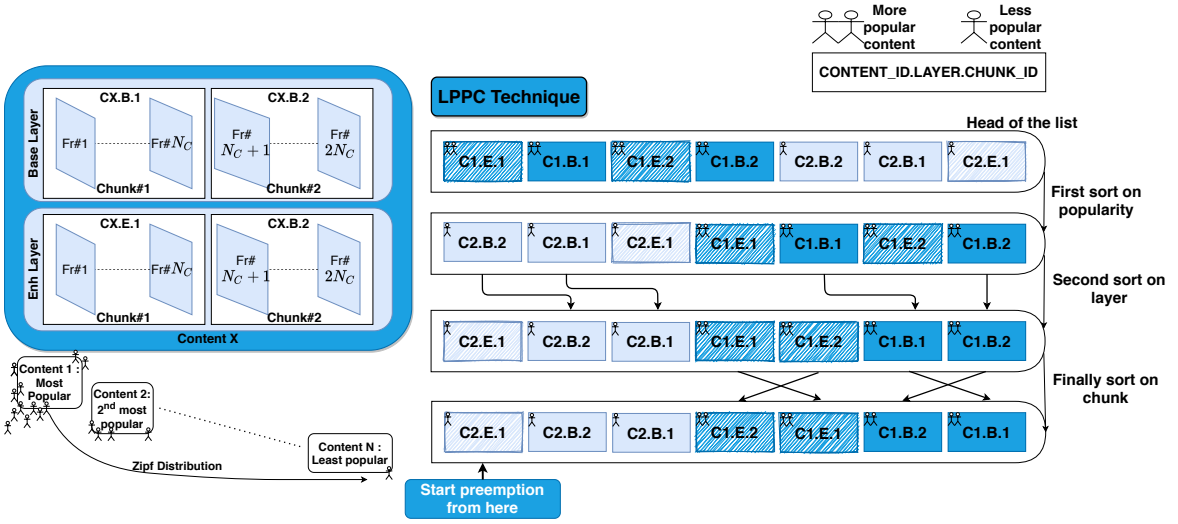

Figure 4.3. An example scenario of the $LPPC$ algorithm.

4.2.1. Time Complexity

The complexity of any caching algorithm is important for practical purposes. Therefore, we investigate the complexity of our proposed algorithms. Let N_c be the number of contents, N_l be the number of layers and N_{ch}^{max} be the maximum number of chunks of a content. Both in $LPPC$ and $CPPC$ contents are first sorted on content popularity with time complexity $O(N_c\ log\ N_c)$. In $LPPC$ after the sorting on content popularity, it sorts initially on the layer dimension on each content with time complexity $O(N_c(N_l\ log\ N_l))$. Finally, it sorts on the chunk dimension for each layer of all contents in the cache. This phase has time complexity $O(N_c\ N_l(N_{ch}^{max}\ log\ N_{ch}^{max}))$. Overall, the time complexity of $LPPC$ algorithm is $O(N_c\ log\ N_c + N_c(N_l\ log\ N_l) + N_c\ N_l(N_{ch}^{max}\ log\ N_{ch}^{max}))$.

In $CPPC$ algorithm, the sorting order of layering and chunking dimensions are the direct opposite of $LPPC$. It has time complexity $O(N_c\ log\ N_c + N_c(N_{ch}^{max}\ log\ N_{ch}^{max}) + N_c\ N_{ch}^{max}(N_l\ log\ N_l))$. Consequently, both of our proposed algorithms $LPPC$ and $CPPC$ operate in polynomial time.

CPPC(S_c, c_{new}, C){

$Cur_{Cap} = Capacity(S_c)$;

if *(*$Cur_{Cap} + size(c_{new}) \leq C$*)* **then**

  return $S_c \cup \{c_{new}\}$;

else

  $S_{sorted} \leftarrow$ sort(S_c, POP);

  $S_{sorted} \leftarrow$ sort(S_{sorted}, CHUNK);

  $S_{sorted} \leftarrow$ sort(S_{sorted}, LAYER);

  //$S_{sorted} = \{s_1,\ s_2,\ ...\ ,\ s_k\}$ ordered from s_1 to s_k

  j = k;

  while *(*$j \geq 1$*)* **do**

    $S_{sorted} \leftarrow S_{sorted} \setminus \{s_j, s_{j+1}, ...s_k\}$;

    if *(*$Cur_{Cap} + size(c_{new}) - \sum_{\theta=j}^{k} size(c_\theta) \leq C$*)* **then**

      return $S_{sorted} \cup \{c_{new}\}$;

    end

    $j \leftarrow j - 1$;

  end

end

}

Figure 4.4. Chunk prioritized popularity based caching (CPPC) © 2019 IEEE [4].

4.3. Performance Metrics

We investigate our proposed caching algorithms in terms of the performance metrics *(i) energy*, *(ii) goodput* and *(iii) energy efficiency*. Before we delve into the definitions of these metrics, we define some basic notations that we utilize in Table 4.2.

4.3.1. Energy

One of the main energy consumption components is the local cache hits of requested content units E_{loc}. $P^u_{loc} \cdot \frac{|s_u|}{C_{loc}}$ is the energy consumption of each local service for some content unit u ($req_u \in S_{(n,n)}$). The summation of local services for all content units and devices in the analyzed network region gives E_{loc} in Equation 4.5.

Table 4.2. Notations for performance metrics.

Parameter	**Explanation**
T_{sim}	The simulation duration
P^{u}_{loc}	The power consumption of a local content unit retrieval
P^{tx}_{D2D}	The transmission power consumption of a device transmitting some content unit
C_{loc}	The local service capacity for a content unit
$C_{D2D}(n,m)$	The D2D channel capacity between the n^{th} and m^{th} devices
E_{block}	The activation energy of devices from the sleeping to the idling state
N_D	The total number of devices located in the cell
S_U	The set of content units identifiable by content, chunk and layer id uniquely
req_u	The request for the content unit u
s_u	The size of the content unit u
$S_{(n,m)}$	The set of services from the n^{th} device to m^{th} one
$Comp$	The set of requests for a content where all the base chunks are transmitted successfully (service completed successfully)
$Fail$	The set of requests for content units that have failed

The aggregate transmission energy of devices operating in the D2D technique E_{D2D} in Equation 4.6 is also an important element for the system energy usage. $P^{tx}_{D2D} \cdot \frac{|s_u|}{C_{D2D}(n,m)}$ is the energy consumption level of the D2D service for some content unit u transmitted between the device pair n and m. E_{D2D} gives the addition of all D2D utilizing services for all content units from all device pairs.

Note that some content unit requests are blocked due to the limitation of the network and cache capacities. For these requests, the devices are transferred from the sleeping to the idling state. In that regard, the activation energy of devices for content units $E_{block}(s_u)$ is summed to get the total blocking energy consumption in Equation 4.7.

$$E_{loc} := \sum_{u \in S_U} \sum_{n \in N_D} \sum_{req_u \in S_{(n,n)}} P^u_{loc} \cdot \frac{|s_u|}{C_{loc}} \tag{4.5}$$

$$E_{D2D} := \sum_{u \in S_U} \sum_{\substack{n,m \in N_D \\ n \neq m}} \sum_{req_u \in S_{(n,m)}} P^{tx}_{D2D} \cdot \frac{|s_u|}{C_{D2D}(n,m)} \tag{4.6}$$

$$E_{block} := \sum_{u \in S_U} \sum_{n \in N_D} E_{block}(s_u) \tag{4.7}$$

$$E_{all} := E_{loc} + E_{D2D} + E_{block} \tag{4.8}$$

To sum up, we obtain the overall energy consumption of our system E_{all} in Equation 4.8 with the summation of all the expenditures among Equations 4.5-4.7.

4.3.2. Goodput

Initially, the aggregate number of received bits at local hits of requested content units over the course of the simulation is defined as G_{loc} in Equation 4.9. Note that for successful reception, any content unit request req_u should be in the set of $Comp$. This is because none of the corresponding content units of a given content request has contribution to goodput, if that given content request has incomplete base chunk(s).

The overall goodput provided by the network through D2D technique is defined as G_{D2D} in Equation 4.10. For leading to a contribution in the D2D goodput, any content unit request should be in the $Comp$ set due to the same reasoning explained above. Any unit request should also be a member of $\overline{Fail}$ set. The $Comp$ and $\overline{Fail}$ set are not necessarily the same. A content request can have all of its base chunks successfully transmitted, thus $req_u \in Comp$. However, some enhancement chunk unit u might have failed for that content and $req_u \in Fail$ and thus not contributing to the goodput via D2D technique.

By the addition of all of these contributions, we get the overall network goodput G_{all} in Equation 4.11.

$$G_{loc} := \frac{\sum_{u \in S_U} \sum_{n \in N_D} \sum_{\substack{req_u \in Comp \\ req_u \in S_{(n,n)}}} |s_u|}{T_{sim}} \tag{4.9}$$

$$G_{D2D} := \frac{\sum_{u \in S_U} \sum_{\substack{n,m \in N_D \\ n \neq m}} \sum_{\substack{req_u \in Comp \\ req_u \in S_{(n,m)} \\ req_u \in \overline{Fail}}} |s_u|}{T_{sim}} \tag{4.10}$$

$$G_{all} := G_{loc} + G_{D2D} \tag{4.11}$$

4.3.3. Energy Efficiency

The division of the overall energy consumption of our system over the total number of transmitted bits gives the energy efficiency (EE) as shown in Equation 4.12.

$$EE := \frac{E_{all}}{G_{all} \cdot T_{sim}} \tag{4.12}$$

4.4. Performance Evaluation

We assess the system performance in terms of *(i) energy, (ii) goodput* and *(iii) energy efficiency* with varying parameters in each given subsection. The default simulation parameters are provided in Table 4.3. We compare our strategies to the baseline *Least Recently Used (LRU)* scheme that replaces the chunk(s) least recently accessed from the cache when the cache capacity is not sufficient to store a newly requested chunk. It is a common algorithm extensively utilized in cache-based systems.

4.4.1. Impact of Zipf parameters

We start our inspection with the popularity dimension and first inspect how our system performs with varying Zipf distribution parameter α for the aforementioned metrics. According to the comparison of the system energy consumption, it is observed

Table 4.3. Default parameters for dimension prioritized caching algorithms.

Parameter	**Explanation**	**Value**
T_{sim}	The simulation duration	1200 sec
α	The Zipf distribution exponent	0.8
λ	The Weibull distribution scale parameter	1
k	The Weibull distribution shape parameter	0.6
p_{HQ}	The ratio of high quality consumers	1
λ_{users}	The mean density of users located in a cell according to Poisson Point Process	$0.0015 \frac{user}{m^2}$
R_z	The radius of the investigation zone	330 m
$R_{D2D}^{\nparallel}$	The radius of interference free D2D transmission zone	120 m
C	The cache capacity of devices	47.1 Mbits
P_{D2D}^{tx}	The transmission power consumption of a device transmitting some content unit	80 mW
d_0	The reference distance of device antenna	1 m
n_{D2D}	The path loss exponent of D2D transmission	3
B	The bandwidth of the terrestrial channel	2 MHz
N_0	The noise power density	-95 dBm

that with increasing α, the total system energy decreases in all caching mechanisms LRU, $LPPC$ and $CPPC$ as shown in Figure 4.5. For larger α, the popularity gap between contents increases. The most popular contents are stored in devices more often and the overall local hit rates are improved. Increasing α benefits energy due to two factors. First, the local hit power consumption level P_{loc}^{u} is less than device transmission power P_{D2D}^{tx}. Second, local hits attain larger local service capacity C_{loc} compared to the D2D service capacity. As shown in Figure 4.5, the $CPPC$ and $LPPC$ policies deplete less energy than the LRU algorithm for any fixed α values. Our proposed techniques cache content units based on the priorities at the popularity, chunking and layering dimensions with the aim of preserving important content portions in local caches for prospective requests and thus reducing network traffic and consuming less energy. The energy consumption of $CPPC$ ($LPPC$) has an improved performance gap with the classical LRU technique ranging from 11.49% (9.05%) to 19.63% (12.88%). With increasing α, the gap between content popularities increases. The request rate for highly popular contents rises and as our heuristics preserve such contents in local

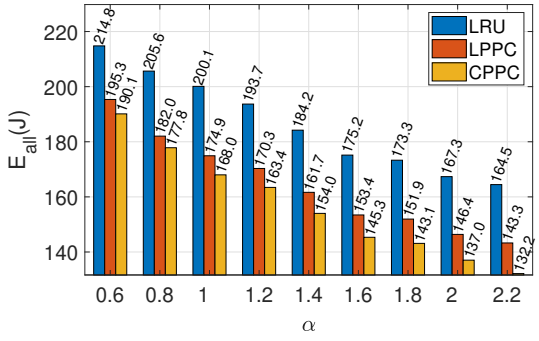

Figure 4.5. Energy performance of caching mechanisms for different α values © 2019 IEEE [4].

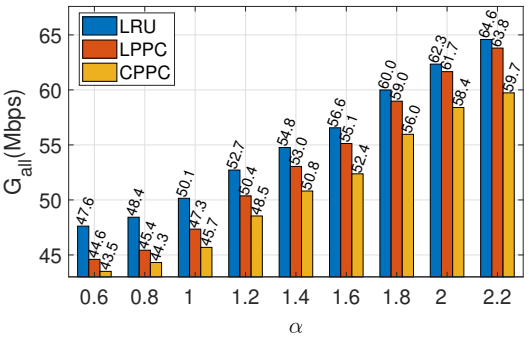

Figure 4.6. Goodput performance of caching mechanisms for different α values © 2019 IEEE [4].

caches with greater priority, we observe larger improvement in total network energy. As shown in Figure 4.5, the maximal improvement of $CPPC$ ($LPPC$) over the LRU is attained at the largest $\alpha = 2.2$ in the Zipf parameter observation space with 19.63% improvement from 164.5 Joule to 132.2 Joule (12.88% improvement from 164.5 Joule to 143.3 Joule). $CPPC$ is more beneficial for energy consumption in contrast to $LPPC$ especially for larger α's. This means for the minimization of the total network energy, chunking has more dominant impact compared to the layering dimension of the video model for large α's.

According to the performance evaluation studies, the system goodput improves with increasing α in all caching strategies given in Figure 4.6. The local availabilities are greater due to the rise in the popularity differentiation among content units at large α regime and therefore we observe the improvement in the system goodput. As shown in Figure 4.6, the $CPPC$ and $LPPC$ policies achieve poorer system goodput in contrast to the LRU algorithm. Compared to the LRU technique, the worst degradation in the goodput is observed at the $CPPC$ algorithm with 8.9% reduction from 50.1 Mbps to 45.7 Mbps at $\alpha = 1$. For the entire inspected Zipf parameter α regime, the $LPPC$ technique is more beneficial that the $CPPC$ in terms of the system goodput. We deduce that layering has greater impact than the chunking on the service quality, unlike the energy consumption case.

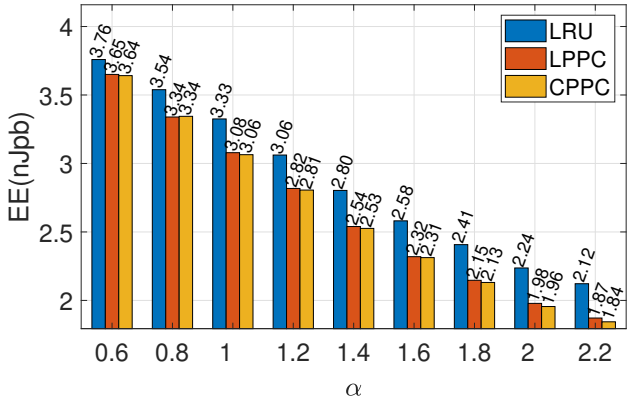

Figure 4.7. EE performance of caching mechanisms for different α values © 2019 IEEE [4].

Despite the system goodput degradation in our caching strategies, they introduce a gain in the energy consumption with a greater dominance. Therefore, our schemes are profoundly energy efficient. To analyze the system energy efficiency, we look at the Figure 4.7. Evidently, our heuristics perform better compared to the LRU with increasing α. For the largest $\alpha = 2.2$ in our investigation domain, the improvement of our strategy $CPPC$ ($LPPC$) over the LRU in terms of EE is 13.10% from 2.12 nJpb to 1.84 nJpb (11.81% from 2.12 nJpb to 1.87 nJpb). No apparent EE difference between our policies is observed. Hence, the different prioritization ordering of our proposed algorithms do not significantly alter the EE characteristic of the network.

4.4.2. Impact of Weibull parameters

The investigation of chunking dimension focuses on the Weibull scale and shape parameters λ and k respectively to portrait the intra-content (among chunks) popularity. For the investigation of varying λ's, we take $k = 0.6$ and during the study of varying k's, λ is taken as 1. We observe that the energy decreases in all caching techniques LRU, $CPPC$ and $LPPC$ with decreasing k (Figure 4.8) and increasing λ (Figure 4.10). The initial parts of video contents are highly requested and thereof cached more often in devices and this increases the local hits and hence we observe improvement in the energy consumption with decreasing k or rising λ. The increase of local hits benefits energy due to the same reasons explained above in Subsection 4.4.1.

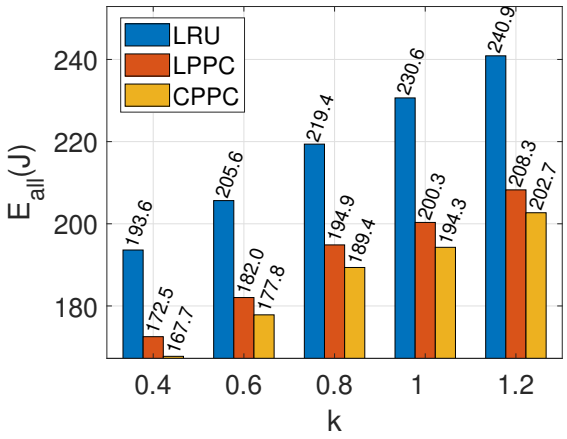

Figure 4.8. Energy performance of caching mechanisms for varying Weibull shape parameter k.

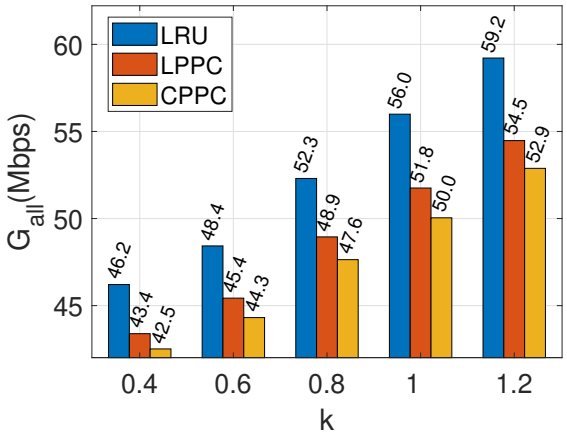

Figure 4.9. Goodput performance of caching mechanisms for varying Weibull shape parameter k.

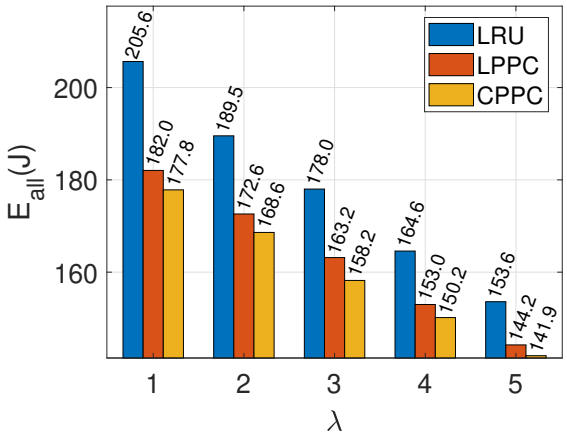

Figure 4.10. Energy performance of caching mechanisms for varying Weibull scale parameter λ.

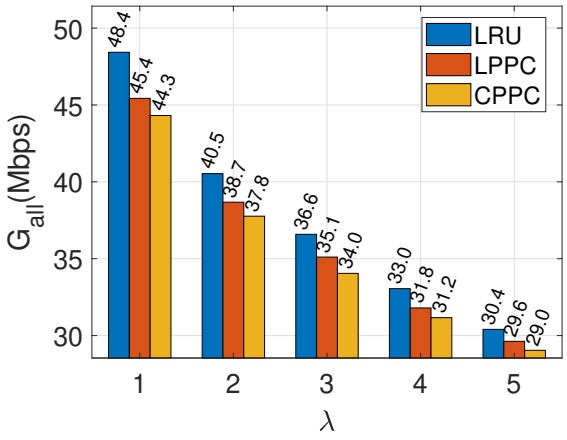

Figure 4.11. Goodput performance of caching mechanisms for varying Weibull scale parameter λ.

Our proposed caching techniques consume lower energy than the baseline profile LRU $\forall k \in \{0.4,\ 0.6,\ 0.8,\ 1,\ 1.2\}$ and $\forall \lambda \in \{1,\ 2,\ 3,\ 4,\ 5\}$ depicted in Figure 4.8 and 4.10 respectively. By preserving important units in caches via priority based caching strategies, we reduce the system energy consumption. The energy consumption of $CPPC$ ($LPPC$) has a improvement over the baseline LRU ranging from 13.37% to 15.86% (from 10.9% to 13.55%) $\forall k \in \{0.4,\ 0.6,\ 0.8,\ 1,\ 1.2\}$. For the parameter set $\lambda = \{1,\ 2,\ 3,\ 4,\ 5\}$, the performance gain of the $CPPC$ ($LPPC$) over the LRU program ranges from 7.62% to 13.52% (from 6.09% to 11.479%). $LPPC$ technique assigns the lowest importance to the chunking dimension. Recall that for breaking ties of a content at the same layer, the chunks are ordered from the initial to the last one. $CPPC$ also maintains intra-chunk prioritization with a greater priority on the chunking dimension compared to $LPPC$. With increasing λ and/or decreasing k, the gap widens between the popularity among chunks of a content. The initial chunks are requested with greater probability and our algorithms $CPPC$ and $LPPC$ cache the initial chunks locally with greater priority and hence local availability increases, leading to a reduction in the total network energy. In the scale parameter domain, the greatest energy improvement of $CPPC$ ($LPPC$) compared to the LRU is achieved at the lowest scale parameter $\lambda = 1$ with 13.52% from 205.6 Joule to 177.8 Joule (11.47% from 205.6 Joule to 182.0 Joule). For the investigation of the shape parameter $k \in \{0.4,\ 0.6,\ 0.8,\ 1,\ 1.2\}$, the largest improvement of $CPPC$ ($LPPC$) over the LRU technique occurs at the largest $k = 1.2$ from 240.9 Joule to 202.7 Joule with 15.86% reduction (from 240.9 Joule to 208.3 Joule with 13.55% reduction).

In all caching techniques, a system goodput improvement is observed with increasing k illustrated in Figure 4.9. Similarly, goodput improves with decreasing λ as given in Figure 4.11. In Figure 4.9 and 4.11 it is seen that our proposals attain lower system goodput compared to the LRU. First we focus on the Weibull shape parameter k. According to the observations in Figure 4.9, the largest decline in the system goodput is observed for the $CPPC$ algorithm at $k = 1.2$ with 10.7% decrease from 59.2 Mbps to 52.9 Mbps. For all $k \in \{0.4,\ 0.6,\ 0.8,\ 1,\ 1.2\}$, the $LPPC$ caching strategy is slightly more beneficial in terms of the system goodput compared to the $CPPC$.

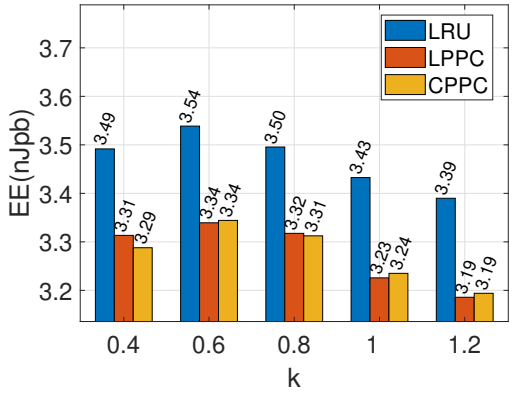

Figure 4.12. EE performance of caching mechanisms for varying Weibull parameter k.

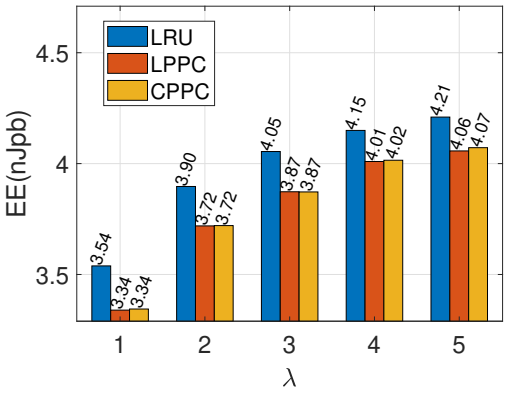

Figure 4.13. EE performance of caching mechanisms for varying Weibull parameter λ in EE results.

When we look at the Weibull scale parameter λ, the worst loss in the system goodput is monitored for the $CPPC$ policy at $\lambda = 1$ with 8.5% reduction from 48.4 Mbps to 44.3 Mbps. In all scale parameter λ domain values, $LPPC$ caching is slightly better than $CPPC$ from the system goodput aspect.

Even if the system goodput is worse for $LPPC$ and $CPPC$ caching algorithms compared to the LRU, as shown in Figure 4.13 and 4.12 they are more energy efficient than LRU. The reason is that the gain in the energy consumption dominates the fall back in the system goodput. Evidently EE levels of our policies are close to each other for all observed Weibull ranges. In that regard, the different prioritization schemes for chunking and layering do not impact the EE of the system.

4.4.3. Impact of p_{HQ} values

In this subsection of the performance evaluation, we study the network operation with respect to layering dimension. The caching mechanisms are investigated for varying HQ consumer ratio p_{HQ}. In Figure 4.14, it is seen that with increasing p_{HQ} the energy expenditure of the system increases for all caching algorithms. The request for enhancement layers increases with increasing p_{HQ} and this induces larger traffic volume and hence the system energy consumption rises.

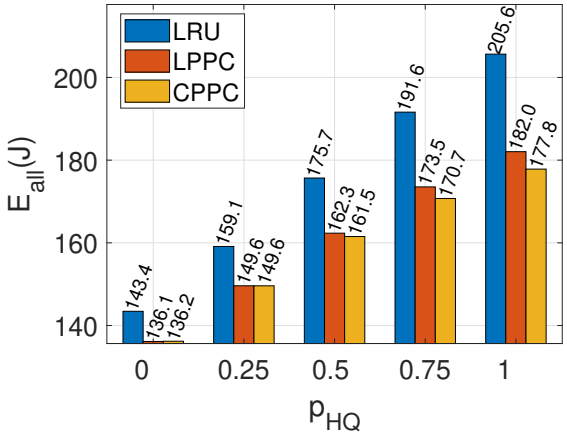

Figure 4.14. The comparison of caching mechanisms for varying p_{HQ} values in terms of energy.

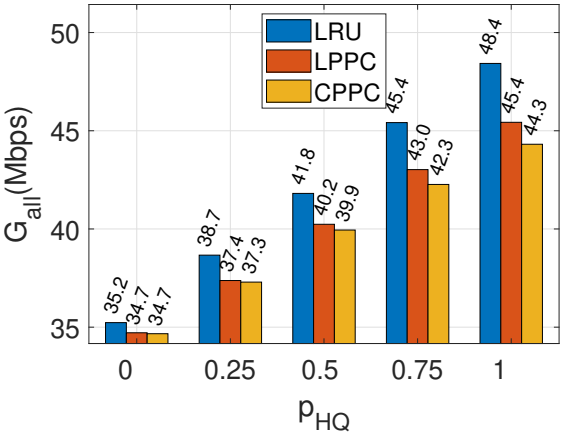

Figure 4.15. The comparison of caching mechanisms for varying p_{HQ} values in terms of goodput.

In Figure 4.14, our policies perform better than LRU caching in terms of the energy consumption for any p_{HQ}. The prioritization of the caching schemes in our proposals leads to the reduction in the system energy. The energy consumption of the $CPPC$ ($LPPC$) caching is improved from 5.05% (5.09%) to 13.52% (11.47%) compared to the LRU across the observation domain of $p_{HQ} \in \{0,\ 0.25,\ 0.5,\ 0.75,\ 1\}$. With increasing p_{HQ}, we observe greater energy improvement. Increasing p_{HQ} results in larger request for the enhancement units. The requested units are locally found with greater probability due to content characteristic aware caching mechanism. Hence, we observe an improvement in the energy. The greatest improvement of $CPPC$ ($LPPC$) over LRU technique is achieved at the highest $p_{HQ} = 1$ with 13.52% (11.47%) from 205.6 Joule to 177.8 Joule (11.47% from 205.6 Joule to 182.0 Joule).

For the system goodput evaluation given in Figure 4.15, we notice that goodput improves with increasing p_{HQ} for all caching mechanisms. The increase of the HQ consumer ratio p_{HQ} induces larger request to the system. With increased request rate, the system needs to serve larger amount of data either locally or via D2D technique and therefore the system goodput rises with increasing p_{HQ}. The $CPPC$ and $LPPC$ techniques reach lower goodput levels compared to the LRU as shown in Figure 4.15. The biggest system goodput decline is monitored for $CPPC$ algorithm at $p_{HQ} = 1$ with 8.5% decrease from 48.4 Mbps to 44.3 Mbps. $LPPC$ first prioritizes the layering

and then the chunking dimension as explained in Section 4.2. It is more beneficial than $CPPC$ in terms of system goodput especially for larger p_{HQ}'s. Hence, we realize that layering dimension has greater impact than the chunking for the system goodput.

Our proposed algorithms $CPPC$ and $LPPC$ fall back of the LRU in terms of system goodput. On the contrary, they are gainful in terms of energy expenditure compared to the LRU and this energy gain dominates the loss in terms of system goodput. Hence, our algorithms $CPPC$ and $LPPC$ are more energy efficient than LRU as depicted in Figure 4.16. We do not observe any significant difference between the energy efficiency levels of $CPPC$ and $LPPC$ in the observation domain of $p_{HQ} \in \{0,\ 0.25,\ 0.5,\ 0.75,\ 1\}$. Therefore, we conclude that the different prioritization order in our algorithms do not impact the energy efficiency nature of our system.

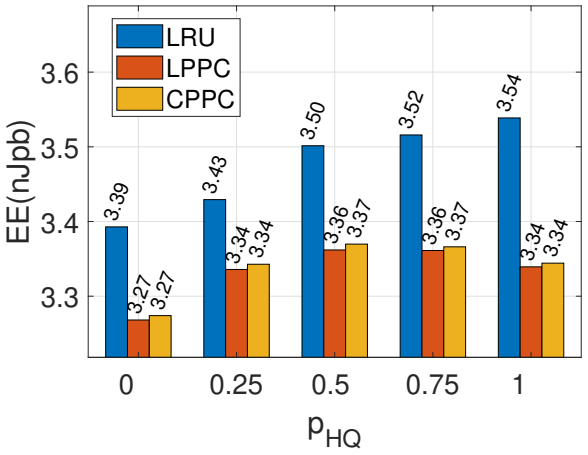

Figure 4.16. The comparison of caching mechanisms for varying p_{HQ} values in terms of energy efficiency.

One of the problems in D2D networks is the battery depletion of devices. Energy harvesting is one of the techniques to counteract this problem [117]. Power allocation in energy harvesting based D2D networks is studied to properly manage scheduling in an energy efficient way [118]. Clustering method is another technique to resolve the energy consumption problem of devices in D2D networks [119]. However, the cluster heads may encounter heavy energy depletion. To overcome this problem cluster head rotation can be used [42]. Another battery depletion resolving technique is to allow

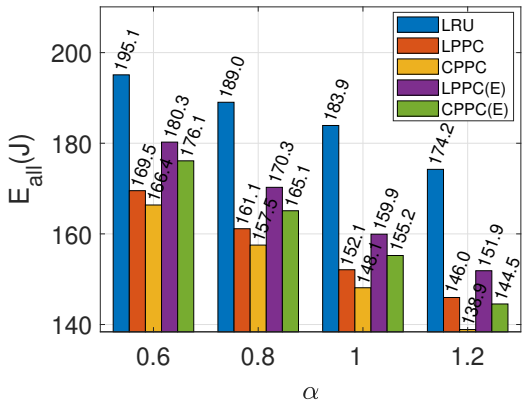

Figure 4.17. Energy performance of caching mechanisms for different α's.

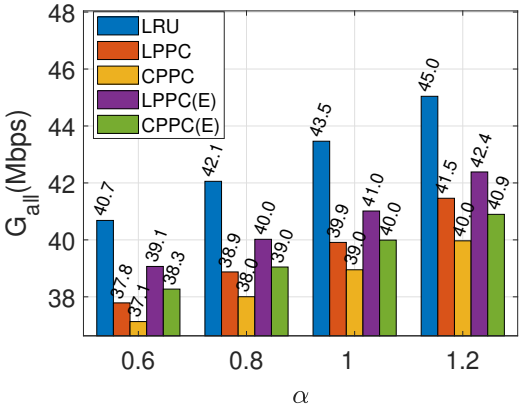

Figure 4.18. Goodput performance of caching mechanisms for different α's.

original transmitter to cooperate with other devices that have enough battery and use those devices as a relay to complete the service with less burden on the original transmitter [120]. Similarly, we will also utilize satisfactory battery level in our proposal. However, we differentiate from [120] by not using the relay concept. In our D2D network we introduce the threshold concept for remaining battery level of devices. First, the devices that have battery level above the threshold are considered to be candidate transmitters. If no such device exists, the devices with limited battery become candidate transmitters. Hence, instead of being transmitter oriented as in [120], our study is requester oriented and focuses on choosing a transmitter with proper battery level. We integrated this threshold based energy management scheme to both $CPPC$ and $LPPC$ and call them $CPPC(E)$ and $LPPC(E)$ respectively. We evaluate how this energy scheme impacts *energy*, *goodput* and *energy efficiency* results with respect to Zipf distribution parameter α.

In Figure 4.17, for all caching techniques E_{all} decreases with increasing α due to the same reasoning explained previously. Our proposals both energy management scheme integrated and not integrated perform better than the classical LRU cache replacement technique. The improvement of energy management scheme integrated $CPPC(E)$ ($LPPC(E)$) over the LRU ranges from 9.73% to 17.05% (from 7.61% to 13.04%) in general ascending with increasing α. However, $CPPC(E)$ and $LPPC(E)$ fall back off the corresponding energy non-integrated schemes $CPPC$ and $LPPC$ re-

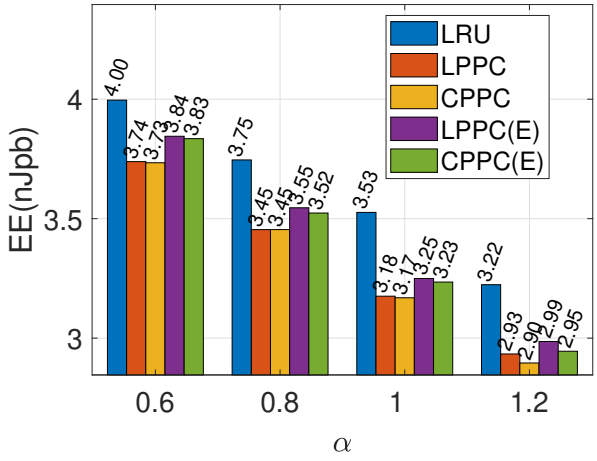

Figure 4.19. EE performance of caching mechanisms for different α values.

spectively. In threshold based energy management utilizing cachings, not necessarily the closest transmitter but the closest device that has sufficient battery is selected. Therefore, in some instances not the shortest D2D service is used and this leads to a slight rise in the total energy consumption. The performance degradation in energy management scheme integrated caching algorithms rises with decreasing α in contrast to non integrated ones. For $\alpha = 1.2$ the decrease of $CPPC(E)$ over $CPPC$ ($LPPC(E)$ over $LPPC$) is 4.07% (4.05%) while it rises up to 5.86% (6.32%) with decreasing α down to 0.6.

We observe that system goodput improves for all caching techniques with increasing α as shown in Figure 4.18. As already explained above, high popularity differentiation among content units for greater α's leads to improved caching benefits. Therefore, we observe an improvement in the system goodput with increasing α. For any α, we monitor a decline in system goodput for $CPPC$ and $LPPC$ compared to LRU. The reason is that our content model utilizing proposals evict non-popular units with a great probability and lead to less content diversity. $CPPC(E)$ and $LPPC(E)$ also have lower system goodput than that of LRU for any α. However, they show improvement over their corresponding content feature utilizing techniques $CPPC$ and $LPPC$. $CPPC(E)$ and $LPPC(E)$ first use devices with remaining battery above a threshold for D2D transmission with the aim of not allowing device batteries exhaust rapidly and therefore the shortest possible D2D service rate is not necessarily used. On the other hand, previously fast depleting devices are live and serve longer thereby dominating

the sacrificed shortest service rates. This improvement of $CPPC(E)$ ($LPPC(E)$ over the corresponding technique $CPPC$ ($LPPC$) is ranging from 2.33% to 3.07% (from 2.23% to 3.39%) with decreasing α from 1.2 to 0.6.

In the energy efficiency investigation, we observe improvement with increasing α in all caching techniques as shown in Figure 4.19. When we compare threshold based energy management integrated caching mechanisms $CPPC(E)$ and $LPPC(E)$ to the classical LRU, we observe energy efficiency improvement for all $\alpha \in \{0.6, 0.8, 1, 1.2\}$. The largest improvement of $CPPC(E)$ ($LPPC(E)$) over LRU is 8.64% at $\alpha = 1.2$ (7.85% at $\alpha = 1$). But they are less energy efficient than corresponding non-integrated versions. In conclusion, we do not observe an obvious EE difference between the results of $CPPC(E)$ and $LPPC(E)$. We deduce that the different prioritization order for the cache replacements does not impact the EE for threshold based energy management integrated mechanisms.

4.5. Cooperative Caching in the Edge Network

Cooperative caching is a promising technique to alleviate the multimedia traffic burden on wireless networks and therefore we also study multimedia caching in D2D networks from the cooperation aspect in this work. We propose cooperative cache replacement algorithms as a contribution. Moreover, the investigation of content scene change dynamics on layer based content model is considered as another technical contribution. We highlight the impact of cooperative caching by investigating through extensive simulations. The cooperative caching study in this section is available in [5].

4.5.1. Scene Change Dynamics

The dynamicity of multimedia has a significant impact on characteristics of video streams and chunks. One such factor is the scene change dynamics. Some contents are more stable with small inter-frame changes while others have rapid and evident inter-frame changes like action movies.

To investigate the scene change dynamics, we utilize 30 fps temporal scalable video traces where I and P frames constitute the base layer while B frames form the enhancement in the temporal scalable encoding [6]. We use the *terse traces* around 30 minutes of *Tokyo Olympics, The Silence of the Lambs, the Star Wars IV* and *NBC News* videos of *GoP* size 16 and the number of consecutive B frames 3 with quantizer 16. For the content classification *spatial* and/or *temporal* features are broadly utilized and the frame difference variance is considered to project temporal multimedia dynamicity [121]. In that regard, for a given content c with the total number of frames F_c and the size of a given frame f as $s_{(f)}$, the average frame size difference in Equation 4.13, we calculate the variance of the frame size differences of aformentioned videos according to Equation 4.14.

$$Avg(c) := \frac{1}{F_c - 1} \sum_{f=1}^{F_c-1} s_{(f+1)} - s_{(f)} \tag{4.13}$$

$$Var(c) := \frac{1}{F_c - 1} \sum_{f=1}^{F_c-1} ([s_{(f+1)} - s_{(f)}] - Avg(c))^2 \tag{4.14}$$

Next, we look at the ratio of the base layer size over the enhancement counterpart $R_c(B,E) := \frac{\sum_{b\in\mathbb{B}_c}(s_b)}{\sum_{i\in\mathbb{I}_c}(s_i)+\sum_{p\in\mathbb{P}_c}(s_p)}$ with $\mathbb{B}_c$, $\mathbb{I}_c$ and $\mathbb{P}_c$ as the B-, I- and P-frame sets of some content c, respectively. We get the statistical information about the ratio of the base layer size over the enhancement by calculating $R_c(B,E)$'s for the aforementioned videos with the provided video statistics in [122]. By monitoring the correlation between the variance of frame size differences and ratio of layer sizes, we observe the Pearson correlation coefficient ρ=0.992 showing a high linear dependency. Thus, we can utilize the layer size ratios to interrogate temporal scene change dynamics as well. In that regard, in our construct we keep the base layer size fixed to have the same SQ content compositions and for the scene change adaptation we change the enhancement size s_e and multiply it with α_{enh}. In the temporal characteristics of contents, the high variance of frame size differences are classified as high temporal activity [121]. As the layer size ratio and the mentioned variance definition are highly linearly correlated, the enhancement layer having larger size for a fixed base layer size implies that the content

has higher temporal changing dynamic. On the contrary, having smaller enhancement size reveals that the content is more stable with little temporal variation.

4.5.2. Cooperative Caching Algorithms

Cooperation is a broadly utilized mechanism for caching techniques [72, 73, 77]. It has a great potential to improve caching gains. In cooperative caching techniques users utilize each others' information to decide on the caching and act accordingly. Here we propose two cooperative caching algorithms and study how they impact the network *energy* and *goodput* results.

We make use of neighbour caches to determine the content unit(s) to be evicted in our cooperative caching mechanisms. Both of our proposals rely on the cooperation of devices within a vicinity of the requester. That cooperation is done for making eviction decisions according to the caching status of neighboring devices in the reception range. In both of our proposed techniques, the requester cache capacity C, the set of content units S_U^D in the requester device cache and the newly retrieved unit u' are taken as system parameters. They utilize all the neighbour device caches in the reception range.

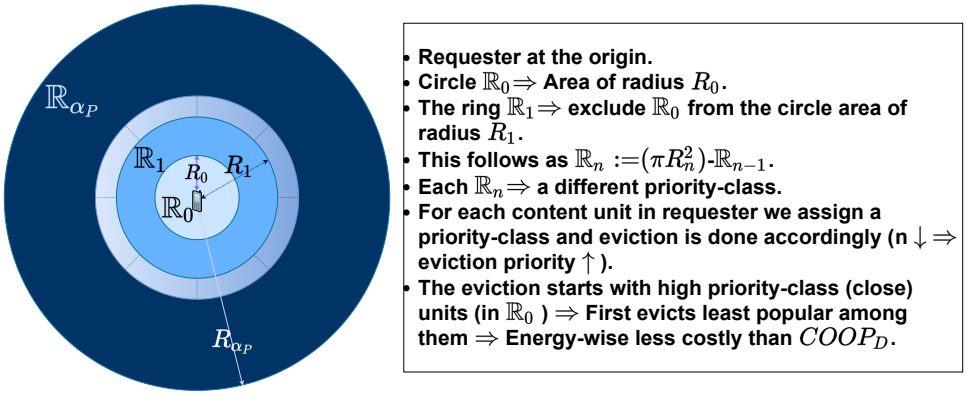

Figure 4.20. D2D reception partitions $\mathbb{R}_0, \mathbb{R}_1, ..., \mathbb{R}_{\alpha_P}$ © 2020 IEEE [5].

INPUTS:

$\mathbf{S_U^D}$: The list of content units available in the cache of device me

$\boldsymbol{c_{new}}$: Newly requested content unit by device me, $\boldsymbol{C}$: The cache capacity of device me

```
%select the set of units to be evicted
if (TotalSize(S_U^D) + size(c_new) > C) then
    sortedU_me ← sort("CPPC", S_U^D); %first sort on the CPPC for tie break
    %initially mark all units as out of reception range R_D2D
    sortedU_me ← [S_U^D, 0]; %priority-(α_P+1) class
    for (i = 1:1:|S_U^D|) do
        for (r=0:1:α_P) do
            if (u_i ∈ R_r) then
                %if u_i available in R_r pin as priority-r class member
                sortedU(i) ← [u_i, α_P + 1 − r];
                break;
            end
        end
    end
end
%sort units based on the priority class from priority-α_P (lowest) to priority-0 (highest) class
sortedU_me ← sortOnPriority(sortedU_me);
***
%pin the highest eviction priority unit(last one) in sortedU_me
w ← highestPriority(sortedU_me);
%for all neighbours in R_D2D, select the devices that also have same highest prospective eviction priority unit w (for
these devices in the very first new request w will be evicted next)
wHolders ← selectNgh(R_D2D, w);
%If this device is locked down for w eviction, decrease w's priority to the lowest and release lock down of w in me
[lockDown, sortedU] ← LDUpdates(lockDown, w, me, sortedU);
%start the eviction from the last unit (from higher priority) in sortedU_me and iterate until there is enough free space
for c_new
sortedU_me ← cacheUpdate(sortedU_me, c_new);
%if neighbours exist with w as the highest prospective eviction priority unit lock down these devices unless already
locked down
lockDown ← nghLDDecide(wHolders, lockDown, w);
%multicast lockdown decisions to related neighbour devices in the reception range R_D2D for prospective requests
multicast(wHolders, lockDown);
%broadcast cache state updates to all devices in the reception range R_D2D for prospective requests
broadcast(me, sortedU_me);
```

Figure 4.21. $COOP_A$ algorithm © 2020 IEEE [5].

In the first scheme $COOP_A$ in Figure 4.21, we construct priority-classes based on closeness to requesters and content unit availabilities in neighbours. We look at each unit in the local cache and assign them to a priority-class. The requester device located at the origin with radius R_{D2D} has reception range $\mathbb{R}_{D2D}$ that is partitioned into mutual exclusive area portions $\{\mathbb{R}_0, \mathbb{R}_1, \ldots ,\mathbb{R}_{\alpha_P}\}$ as depicted in Figure 4.20 with $\alpha_P + 1$ many distinct mutual exclusive areas. We define the radius of each D2D reception partition by $R_n{:=}R_{n-1}{+}R_0$ and $R_0{:=}\frac{R_{\alpha_P}}{\alpha_P+1}$ in our model. $\mathbb{R}_0$ is defined to be the inner circle with radius R_0 with the requester located at the origin. The ring area

$\mathbb{R}_1$ is defined to be exclusion of $\mathbb{R}_0$ from the circle area of radius R_1 with the requester located at the origin. This recursively follows as $\mathbb{R}_n := (\pi R_n^2)\text{-}\mathbb{R}_{n-1}$ with $R_{\alpha_P}{:=}R_{D2D}$. For each of these areas $\mathbb{R}_n$ we define a priority-class in terms of eviction. Content units in the requester cache that are also available in areas closer to the requester have greater eviction priority. Content units already available in at least one neighbour in area $\mathbb{R}_0$ are *priority-0* class members with the *highest priority* in terms of eviction. Next, the units already available in at least one neighbour in $\mathbb{R}_1$ are *priority-1* class members that have lower priority than priority-0 class. In general, the units already available in at least one neighbour in $\mathbb{R}_n$ are members of *priority-n* class having lower priority than priority-(n-1) class. The priority-(α_P+1) class is for the units out of D2D reception range $\mathbb{R}_{D2D}$ with the lowest priority. After the labeling of priorities, units are priority-sorted form the lowest to the highest in $sortedU_{me}$.

For requests to devices in the reception range of each other with the same unit w with the highest eviction priority, they will all evict w and erase it from the neighbourhood, which may lead to a caching performance degradation due to this *bandwagon effect*. To overcome this problem, the first content requester device signals others to "lock down" w and then that first requester evicts w. Thus, the eviction priority for w at neighbours is decreased until their prospective requests. Thereby, the cooperation postpones the aggressive eviction and via this "lock-down" concept, we open up an opportunity to receive cache update signals from the neighbours for proper caching decisions and make w to survive in the network for a longer amount of time. This action is achieved by $selectNgh$ and $LDUpdates$ functions in Figure 4.21.

In function *cacheUpdate*, we start the eviction with high priority labeled units and in our range partitioning scheme we label close vicinity area with a great priority in terms of the eviction. Hence, we first evict already in-close neighbour available units and thereof, prospective requests for such units at the same requester can be handled via D2D mechanism in a short range. Besides, our scheme has another dimension for the service capacity improvement. It assigns lowest priority for local units that are not found in reception range and this leads to a greater probability to preserve such units in the cache and in the future serve them locally. When a prospective request for

these units from some other device occurs, they are transmitted within the reception range via D2D technique with a great probability as well. Apart from these aspects, by removing in-neighbour available units from the local, we open up more space for larger number of different units and hence this impacts the overall system performance. For tie breaking purpose of the units in the same priority class, we utilize our technique $CPPC$ that is proposed in Section 4.2. This technique is called availability based cooperation ($COOP_A$) algorithm and given in Figure 4.21.

After the *cacheUpdate* operation, we lock down in-range neighbor devices that store unit w as the next highest priority eviction candidate by applying the function *nghLDDecide* to block them from starting eviction with w and allow w to have a longer lifespan. Next, these *lockDown* decisions are conveyed to those related neighbors for cooperation by the *multicast* function. Finally, we broadcast cache state updates to neighbor devices in the reception range as well.

INPUTS: $\mathbf{S_U^D}$: The list of content units available in the cache of device me,
$\boldsymbol{c_{new}}$: Newly requested content unit by device me, $\boldsymbol{C}$: The cache capacity of device me

if *(***TotalSize**($\mathbf{S_U^D}$) + **size**($\mathbf{c_{new}}$) > C*)* **then**
  %first sort on $CPPC$ for tie break
  sortedU$_{me}$ ← **sort**("**CPPC**", $\mathbf{S_U^D}$);
  sortedU$_{me}$ ← [$\mathbf{S_U^D}$, **Inf**];%indicate all units as in-neighbour unavailable
  for *(i = 1:1:$|S_U^D|$)* **do**
    %get the set of devices that have u_i available in the reception range $\mathbb{R}_{D2D}$ of this c_{new} requester device me
    curtxDevs ← **getCandidateTxs**($\mathbf{u_i}$);
    if *(***size**(**curtxDevs**) > 0*)* **then**
      %among these devices having u_i choose the distance of the closest one to me
      dist ← **closestDevDistance**(**curtxDevs**);
      sortedU$_{me}$(**i**) ← [$\mathbf{u_i}$, **dist**];
    end
  end
  %sort units based on the distance of the closest available neighbour starting from the distant (lowest priority) to closest (highest priority) one
  sortedU$_{me}$ ← **sortOnDistance**(**sortedU$_{me}$**);
  ***
end

Figure 4.22. $COOP_D$ algorithm © 2020 IEEE [5].

In the other proposed algorithm $COOP_D$ shown in Figure 4.22, for each local content unit u_i that is also available in some neighbour device, that u_i is attributed by the distance from the local device to the closest neighbour having u_i. All of these local units are sorted by these attributed distances from distant to closest one in *sortOnDistance*.

For the *cacheUpdate*, the closest available content unit has the highest priority for the eviction and the priority descends with increasing distance. The non-available ones are assigned with distance infinity. By prioritizing the closest in-neighbour available unit(s) for eviction, for prospective requests short-distance D2D communication will be utilized. The units with the same distance have the same priority. To break ties, we utilize $CPPC$ again. This technique is called distance based cooperation ($COOP_D$). In this algorithm, the lock-down cooperation mechanism and utilized functions are the same with the $COOP_A$ (marked as ***). An example of the $COOP_D$ technique is illustrated in Figure 4.23. In this example, the requester device at the origin requests unit u_2. For storing it eviction is needed. At the requester device cache, unit u_3 is available in a device r_{u_3} away while u_1 and u_5 are available in a closer device with distance r_{u_1}. For tie breaking, $CPPC$ is applied and accordingly unit u_5 is evicted first.

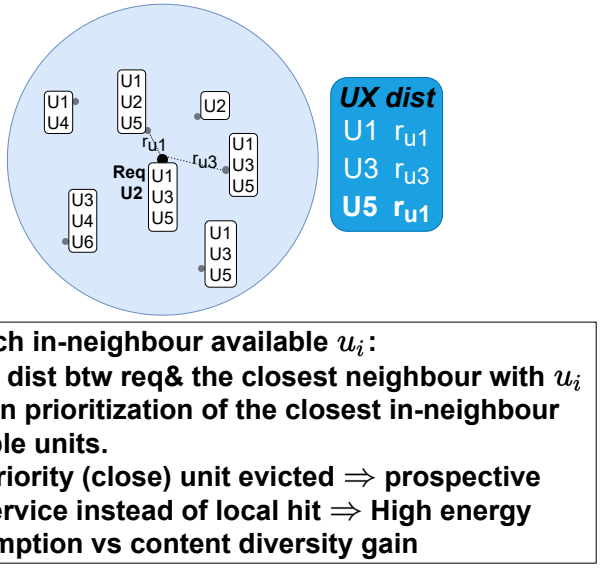

Figure 4.23. An example for $COOP_D$ algorithm.

4.5.3. Performance Evaluation of Cooperative Caching Mechanisms

In this subsection, our main motivation is to analyze our cooperative caching mechanism and the impact of scene change dynamics in terms of the energy consumption and goodput in our D2D network. We compare our cooperative $COOP_D$ and $COOP_A$ algorithms to the baseline LRU, our technique $CPPC$ that is proposed in Section 4.2, minimum-access (MIN-ACC) strategy [123] and Size*Order (SXO)

Table 4.4. System and simulation parameters for cooperative caching algorithms.

Par.	**Value**	**Explanation**
T_{sim}	1800 s	The total simulation duration
N_l	2	The number of content layers
N_p	23.56 Mbits	The average partition size
α_P	1	The number of mutual exclusive partitions of the D2D reception range is α_P+1
λ_D	$1.46e^{-3}\frac{dev}{m^2}$	The mean density of device distribution in PPP
α	1	The skewness parameter of Zipf distribution
λ	1	The Weibull distribution scale parameter
k	0.6	The Weibull distribution shape parameter
p_{HQ}	1	The ratio of high quality consumers
C	47.1 Mbits	The cache capacity of devices
R_{D2D}	120 m	The reception range radius of a requester
s_b	322.0 Mbits	The average base layer size
s_e	152.1 Mbits	The average enhancement layer size
P^u_{loc}	40 mW	The power consumption of local content unit retrieval
P^{tx}_{D2D}	80 mW	The transmission power consumption of a device
P^{rec}_{D2D}	16 mW	The reception power consumption of a device
N_c	250	The number of contents
$N_{f_{ter}}$	4	The total number of system frequencies
B	2 MHz	The operation bandwidth of each frequency
α_{enh}	1	The interplay multiplier for changing average enhancement layer size

cache replacement policy [76]. The system and simulation parameters are available in Table 4.4. We implemented our event-based simulator in MATLAB R2020 environment and ran the simulations on a laptop device with Intel i7-8550U CPU @1.8GHz and 16 GB RAM. For each experiment case with 1800 sec, we run the simulations 10 times and take their mean result. The simulator processes incoming content unit requests (arrivals) and service completions. Each D2D service has B bandwidth for transmission and in total there are $N_{f_{ter}}$ frequencies. The service completions are handled by frequency preemption. The content request rate is 8 $\frac{user}{sec}$ and in the request management scheme, the request manager first checks the local cache for content unit requests. In case of a local hit, no transmission occurs. Otherwise, the closest device in the reception range serves via D2D technique. The reception range $\mathbb{R}_{D2D}$ is the circle with radius R_{D2D} centered at the requester device. We assume the inter-user

distances in a reception range are known by users via a signaling mechanism to discover neighbour devices [75] and cooperative calculation in this D2D network [124]. Therefore, devices in the reception range of each other can exchange their locations for the determination of inter-device distances. For the D2D interference management, no two D2D transmissions are allowed to co-operate at a frequency in a close proximity by realizing the following arrival criteria: *i)* the new requester is at least R_{D2D} away from all active transmitters *ii)* the new transmitter is out of reception range of all active receivers. In our simulations we take D2D interference into account. At a frequency with simultaneous D2D transmissions, for each receiver device the received power strength of other transmitters directed for other receivers are summed as the corresponding interference. In that regard, the channel capacity is updated. For any new D2D arrival, the interference of currently active transmitters to the new receiver at that frequency and its corresponding capacity are calculated dynamically. The interference of new transmitter to each active receiver at that frequency and their capacities are also dynamically adopted. For a D2D service completion, the ceased interference of terminated transmitter to active receivers at that frequency are erased. The device cache replacement management is our *main use case* and therefore we specifically focus on device services and energy levels with respect to different cache replacement schemes that are utilized when there is not enough free cache space for the requested and retrieved content and an eviction decision is required to take action.

In Figure 4.24, it is shown that with increasing α_{enh} values the total system energy consumption increases for all caching techniques. Intuitively, with increasing α_{enh} the enhancement size and subsequently the total content size rises. Hence, the system requires greater energy consumption even for a local hit. Besides, the probability of local availability decreases and hence the content has to be transmitted in D2D mode and this requires larger energy than that of a local hit. The reason for D2D technique leading to heavier energy consumption is due to two aspects. The first aspect is the larger power level of D2D compared to the local hit. The second one is larger service duration of a D2D mode transmission.

When we compare the cache management technique $CPPC$ with other algorithms, we observe that $CPPC$ outperforms others in terms of energy consumption E_{all} as seen in Figure 4.24. The largest observed improvement of $CPPC$ over the LRU is from 341.6 J to 275.1 J at $\alpha_{enh} = 2.5$. As $CPPC$ makes us of content characteristics, its improvement over LRU (MIN-ACC) algorithm becomes more evident with larger enhancement content sizes. $CPPC$ also outperforms SXO algorithm in terms of E_{all} and the improvement rate reaches up to 19.3% (from 341.1 J to 275.1 J) at α_{enh}=2.5. In SXO algorithm, large popular units are prioritized for the eviction and in our content model large units are base units. The eviction prioritization of base units in SXO leads to more transmissions in D2D communication mode for such units rather than local hits for them. Thereby, the energy consumption of SXO is increased compared to the $CPPC$ algorithm. When we compare $COOP_D$ to the LRU (MIN-ACC), its consumption falls back for all α_{enh}'s in the inspection domain because of the prioritization nature of the replacement unit selection in this cooperative technique $COOP_D$. It evicts in-neighbour available units based on closeness prioritization. The more popular a content unit is, the higher probability it has to be found in a close device and hence to be selected for eviction with a greater priority. The prospective requests for these evicted popular units can be served in D2D mode instead of local hits. Therefore, it has higher energy consumption compared to the LRU (MIN-ACC) ranging from 24.7% to 28.3% (17.4% to 25.5%). $COOP_D$ also falls back of the SXO for all α_{enh}'s in the inspection domain. As already explained above in $COOP_D$ popular units that are already available in the neighbourhood are given a greater eviction priority and requests for such units can be served by D2D transmission instead of local hits. SXO considers local cache and selects highly popular large units that are of type base for eviction and then will cover up for the requests of these units via D2D mechanism. However, $COOP_D$ will also require D2D transmission for popular enhancement units available in a close vicinity and hence it will consume more energy than SXO ranging from 24.9% to 34.0%. $COOP_A$ has also higher results compared to the LRU (SXO) except for the largest α_{enh}=2.5. $COOP_A$ outperforms MIN-ACC in terms of energy for $\alpha_{enh} \geq 1.5$ and the maximum improvement rate observed is 7.3% from 363.1 J to 336.7 J at $\alpha_{enh} = 2.5$.

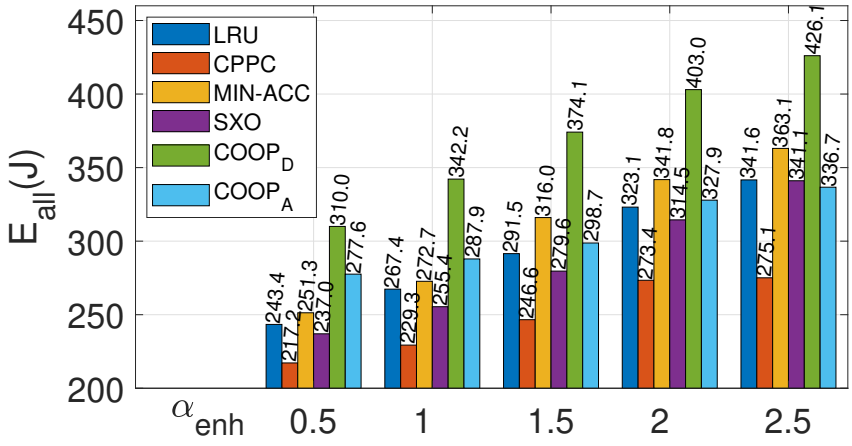

Figure 4.24. Cooperative caching energy results for varying α_{enh} values © 2020 IEEE [5].

Until now, we have compared cooperative techniques to the *LRU*, *MIN-ACC* and *SXO* algorithms. Next, we compare cooperative techniques $COOP_A$ and $COOP_D$ to the *CPPC* and we observe a rise in E_{all} the total system energy consumption. In both cooperative techniques (either based on the distance or not), in-neighbour available units are to be evicted first and such units have higher popularity compared to ones not available in neighbours. When prospective requests for such more popular units occur again, this time they are not found locally but served by D2D transmission. However, in *CPPC* such units are not necessarily evicted and thereof their prospective requests are handled locally. Thus, cooperative techniques result in an increase in the energy consumption in contrast to *CPPC*.

When we compare cooperative techniques with each other, we observe that $COOP_A$ algorithm has better energy performance than the distance based one $COOP_D$ for all α_{enh}'s as shown in Figure 4.24. This improvement reaches up to 21.0% from 426.1 to 336.7 J at α_{enh}=2.5. In distance based cooperative cache replacement technique, local content units are sorted based on the distance from the closest unit-storing neighbour. The more popular a unit is, the higher probability it is found in a close device. Thus, more popular units are assigned higher priority for the eviction. Prospective requests for such units can be handled via D2D transmissions instead of low-cost local hits. However, in $COOP_A$ with $\alpha_P = 1$ content unit availability in a neighbor is sufficient

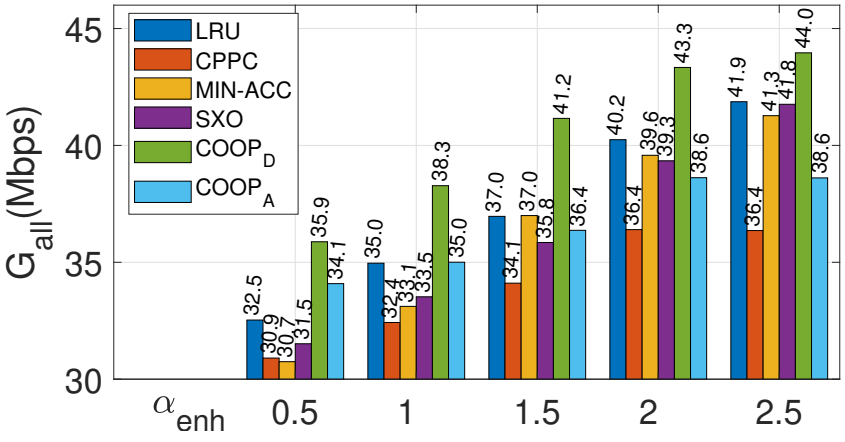

Figure 4.25. Cooperative caching goodput results for varying α_{enh} values © 2020 IEEE [5].

to be selected for the eviction. Among these in-neighbour available units, $CPPC$ selects the least popular one as the prioritized candidate for the eviction. Therefore, $COOP_A$ preserves popular units among in-neighbour units and thus prospective requests for them can be low energy cost local hits. This explains the effective improved performance of $COOP_A$ over distance based $COOP_D$ algorithm.

When we focus on the goodput performance of our system, we observe the results in Figure 4.25. With increasing α_{enh}, the requested content size rises and the system serves with larger capacity in all caching mechanisms. $COOP_D$ has better goodput values than LRU (MIN-ACC) for all α_{enh}'s in the investigation range, with improvement reaching to 11.3% from 37.0 Mbps to 41.2 Mbps at α_{enh}=1.5 (16.7% from 30.7 Mbps to 35.9 Mbps at α_{enh}=0.5). This is due to the erasure of high popular content units that are available in the neighbourhood and opening up caching opportunity to new units and increasing the diversity of units. Hence, more requests can be served and the rise in overall goodput G_{all} is observed. $COOP_D$ has also greater goodput G_{all} over the SXO for all α_{enh} values in the investigation range. This improvement in terms of G_{all} reaches to 14.8% from 35.8 Mbps to 41.2 Mbps at $\alpha_{enh} = 1.5$. $COOP_D$ algorithm gives high eviction priority to popular units in all types. However, SXO gives high eviction priority to popular base units only. Enhancement units are less likely to be selected as eviction candidates with their smaller size. The low eviction rate of highly popular

enhancement units in SXO results in less opening up opportunity to new units and lower unit diversity in contrast to the $COOP_D$ algorithm. This leads to an improved $COOP_D$ goodput over the SXO.

When we look at the $COOP_A$ algorithm, we observe lower goodput values than the classical LRU approach for large α_{enh}'s with $\alpha_{enh} > 1.0$. At $\alpha_{enh} = 0.5$, $COOP_A$ technique outperforms the LRU and a slight improvement with 4.8% is observed from 32.5 Mbps to 34.1 Mbps in Figure 4.25. $COOP_A$ has lower goodput values than the $MIN\text{-}ACC$ (SXO) algorithm for α_{enh} >1.0 (α_{enh} >1.5). With decreasing α_{enh}, our proposed cooperative $COOP_A$ algorithm performs better than the $MIN\text{-}ACC$ (SXO). At α_{enh}=0.5, its improvement over the $MIN\text{-}ACC$ (SXO) reaches to 10.9% (8.2%) from 30.7 to 34.1 (31.5 to 34.1) Mbps as shown in Figure 4.25.

In the investigation of cooperative techniques compared to $CPPC$, they have improved goodput values for all α_{enh}'s in the inspection domain as shown in Figure 4.25. The improvement rate of $COOP_D$ ($COOP_A$) over the $CPPC$ reaches to 20.9% from 36.4 to 44.0 Mbps at $\alpha_{enh} = 2.5$ (10.3% from 30.9 to 34.1 Mbps at $\alpha_{enh} = 0.5$). Cooperative techniques open up caching opportunity to new units during replacement and hence we observe a total service capacity G_{all} rise compared to the $CPPC$ algorithm. When we compare cooperative techniques $COOP_D$ and $COOP_A$ with each other, we notice that the distance based algorithm $COOP_D$ has greater goodput results than the $COOP_A$ for all α_{enh} values. At $\alpha_{enh} = 0.5$, the goodput improvement of $COOP_D$ over $COOP_A$ is 5.3% from 34.1 Mbps to 35.9 Mbps. The improvement rises up to 13.9% at $\alpha_{enh} = 2.5$ from 38.6 to 44.0 Mbps. Distance based cooperative algorithm outperforms the other one in terms of system goodput. The reason is that the distance based cooperative algorithm opens up more space to new units as it starts eviction from units common in the neighbourhood while non-distance based algorithm $COOP_A$ starts the eviction process from units that are also available in vicinity but not so popular due to the sorting mechanism in $CPPC$. Thus, an in-neighbour unit that is not so popular is a replacement candidate at first and if the neighbour also evicts that unit due to its low importance until a prospective request, then content unit diversity is not preserved as in $COOP_D$ and such requests are not served. This results in a lower system capacity

in the $COOP_A$ than the $COOP_D$ algorithm.

4.5.4. Discussion

In this chapter, we have elaborated on the D2D edge part of our HetNet architecture. We have developed an enriched multimedia content model consisting of *popularity*, *chunking* and *layering* dimensions and we have proposed dimension-prioritized content cache replacement algorithms $LPPC$ and $CPPC$. In terms of energy reduction the chunking dimension is more beneficial while the layering is a system goodput improving factor. Regarding the EE, both $LPPC$ and $CPPC$ outperform the LRU. We have proposed availability-based and distance-based cooperative caching algorithms $COOP_A$ and $COOP_D$, respectively as well. For all scene change dynamics, $COOP_D$ outperforms the LRU, $MIN\text{-}ACC$, SXO and $CPPC$ in terms of goodput. However, there is a trade-off factor increased energy consumption. To alleviate this, $COOP_A$ can be utilized that has much the same energy consumption as LRU (less energy consumption than $MIN\text{-}ACC$) under rapidly changing scene regime. We have compared our proposals with respect to scene change dynamic aspect of the multimedia. The main component of our system in terms of modeling is the multimedia and the motivator for improved performance is its features. In that regard, further investigation of the multimedia service will be required. Our D2D edge communication for multimedia services is the initial ground for the ultra high definition machine type communications in the 6G era. With the integration of mega satellite constellations, ultra-low latency for multimedia contents will be a fundamental requirement for such network systems. In that regard, the derivation of closed form latency in particular regarding the satellite for both the propagation and content delivery aspects is essential. Besides, monitoring the service quality by rate-distortion curves and calculating QoE values can also be applied.

5. ENERGY MINIMIZING OPTIMAL AND HEURISTIC CACHING TECHNIQUES FOR CELLULAR D2D NETWORKS

In this chapter, we investigate the edge cache management techniques in D2D cellular networks owing to the fact that in-network caching is regarded as an efficient facilitator for improving system-wise energy consumption. We have a comprehensive service-mode based energy model and define content energy expenditures based on the prospective consumption across the network. We formulate the optimization for the cache replacement problem in D2D networks and solve it with dynamic programming. Besides, we propose a heuristic algorithm *Energy Prioritized D2D Caching (EPDC)* to improve the time complexity of the content services. Finally, a rigorous performance analysis is performed while focusing on the service rate and energy consumption in different operation modes separately.

5.1. System Model

In this section we present the network architecture, content model and content request management. The network architecture is shown in Figure 5.1. In our network architecture, there is one cell with one base station (BS) and D2D enabled user devices. As broadly utilized in the literature [105, 125], user devices are spread across the cell according to Point Poisson Process (PPP) with mean density λ_{users}. There exists the universal source to serve the contents that are not available in any system unit (e.g. BS, device) at the edge network. In our network architecture, there are four main service operation modes: (i) local hit, (ii) D2D mode, (iii) BS mode (direct), (iv) BS mode (from the universal source) as marked in Figure 5.1.

In our content model, request probabilities are determined by popularity that is based on the Zipf distribution. The probability of content requests are calculated via $Zipf(\alpha, N)$ where α determines the distribution skewness and N is the total number

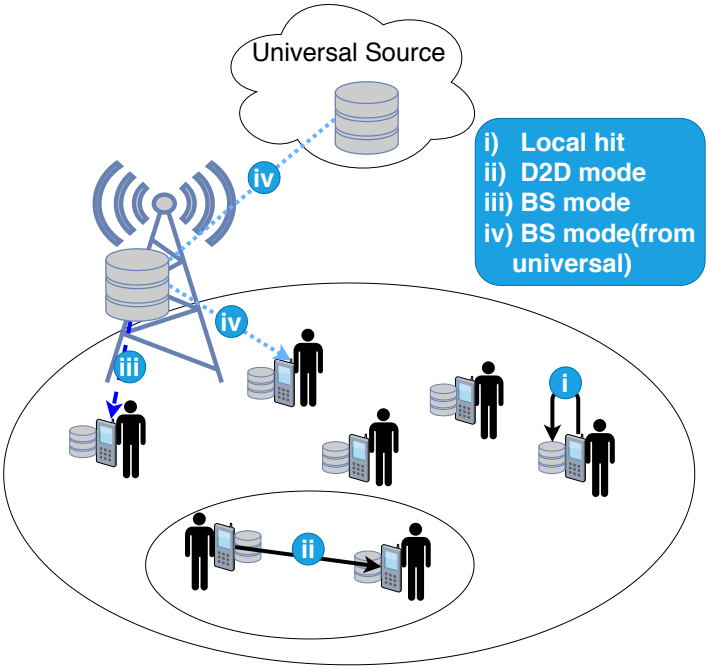

Figure 5.1. Cellular D2D edge network architecture.

of contents in the system. Content chunking is also used in communication systems with the aim of performance improvement [109, 126]. Hence, contents are partitioned into chunks in our model. Inter-chunk popularity differentiation is another approach used for improving transmission and caching performance in content-driven networks. In that regard, the request rate of content chunks are calculated by the commonly used $Weibull(\lambda, k)$ distribution with the scale and shape parameter λ and k respectively.

Our contents are scalable coded (SVC) videos. Each content is of high quality (HQ) consisting of one *base* and one *enhancement* layer. For standard quality content consumption, the base layer is sufficient whereas both layers are required for HQ visualization. The request rate for HQ contents in our network scenario is $p_k \in [0, 1]$. We define each layer of a given content's particular chunk uniquely as a *content unit tuple* by its content, chunk and layer ids respectively as $\{i, j, k\}$. Each such tuple is mapped to a unique *content unit identifier* u for tractability. The average content sizes are 322 and 474 Mbits for SQ and HQ SVC videos, respectively as provided in Section 4.1. Some utilized video samples are *Citizen Kane*, *Jurassic Park I*, *Star Wars IV*, *Aladdin* and *Tonight Show*. For the full video list, please refer to Table 4.1. In Section 5.2, we preserve the content characteristics of our content model and only change

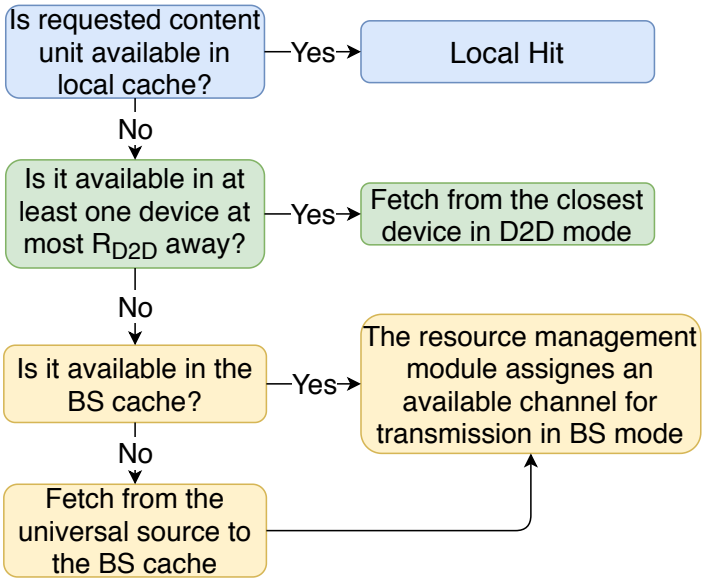

Figure 5.2. Content request management in the cellular D2D edge network.

the total number of contents to push the system to its limits for the feasibility inspection of the time complexity analysis. In the performance evaluation section, we also keep the total number of contents fixed and investigate the impact of some other system parameters as described in Section 5.4.

We extend the wireless D2D network in the previous chapter with cellular support. In this context, the BS transmission power P_{BS}^{tx} and path loss exponent of BS transmission n_{BS} are introduced for the calculation of the BS power strength in the BS channel model. The general content request management scheme is shown in Figure 5.2. For a requested content unit, initially the local cache is explored. If that unit is not locally available, then either D2D mode or BS mode is selected. As long as that requested unit is found in at least one device in the reception range of (at most R_{D2D} away from) the requester, D2D mode is selected. Otherwise, the BS serves to the requester. The BS mode operation branches as follows: *i)* that unit is available in the BS cache, *ii)* that unit is not available in the BS cache. In the branch-*i*, that unit is directly served by the BS cache. In the branch-*ii*, that unit is first fetched from the universal source to the BS and then served to the requester device to allow access to the exterior of the edge network. All requests and network using operation modes (ii–iv) acquire the same network access priority. If a content unit cannot be served

due to all channels being busy, then that unit and prospective units of that content are all dropped because of the bufferless operation scheme.

The BS and devices have storages for caching multimedia contents. The capacities of these units are provided in Table 5.4. In this paper, the cache replacement is our main focus. We elaborate on the cache replacement techniques running both on user devices and the BS. During the cache replacement, the system unit (device / BS) evicts its cache until there is sufficient free space for the requested and retrieved video content unit. The decision for the evictions are explained in Section 5.2.

5.2. Optimal and Energy Prioritized D2D Caching (EPDC)

5G systems is considered to be an enabler for multimedia services. However, the tremendous demand for the multimedia increases the energy cost of the wireless networks and this leads to an environmental issue that needs to be resolved. To this end, 5G communications need to improve energy efficiency performance [20]. As a 5G technology, the D2D mechanism in the edge network provides service capabilities closer to end users and thus reduces the burden on the network energy. Besides, in-network caching can be utilized to alleviate the transmissions and improve the energy efficiency of the network. Accordingly, we formulate the cache replacement problem in cellular D2D edge network for reducing energy consumption. We employ optimal and heuristic approaches to solve it for our system.

Initially, we solve the scalable video caching in D2D networks with the backhaul support optimally. For the ease of tractability, the model and system parameter notations are listed in Table 5.1. The content units residing in the local cache of some user are denoted by S_c. Some user requests a content unit c_{new} and it is retrieved from some other system component via wireless transmission. If the size of content units in the set S_c and new arrived unit c_{new} together exceed the cache capacity C_{Dev}^{cache}, then a subset of content units S_c residing in the local cache need to be evicted. The eviction decision is essentially a 0/1 Knapsack problem [127]. For each content unit you either decide to sustain that unit in the local cache or remove it. The content units that are

Table 5.1. Notations for energy-based caching algorithms.

Parameter	**Explanation**
N	The total number of contents in the system
α	The Zipf distribution skewness parameter
λ	The Weibull distribution scale parameter
k	The Weibull distribution shape parameter
p_i	The probability of content i being requested
p_j	The probability of content chunk j being requested
p_k	The probability of content layer k being requested
$\{i, j, k\}$	The unique *content unit tuple* (The i^{th} content's chunk j of layer k)
u	The unique *content unit identifier* such that each $\{i, j, k\} \mapsto u$
s_u	The size of the content unit u
R_{D2D}	The radius of the total reception range in D2D mode
N_{ngh}	The number of neighbouring devices at most R_{D2D} away from a requester
C_{Dev}^{cache}	The device cache capacity
C_{BS}^{cache}	The base station cache capacity
C_{loc}	The expected service capacity of content unit local hit
C_{D2D}	The expected service capacity of content unit retrievals via D2D technique
C_{BS}	The expected service capacity of content unit retrievals from the BS cache
$C_{BS(U)}$	The expected service capacity of content unit retrievals from the universal source to the BS cache
P_{loc}^{u}	The power consumption of local content unit retrieval
P_{D2D}^{tx}	The transmission power consumption of a device
P_{BS}^{tx}	The transmission power consumption of the BS
P_{BS}^{rec}	The reception power consumption of the BS
θ_{loc}	The local hit service power parameter
θ_{BS}	The base station reception power parameter
δ_{Dev}	The incremental units for the device cache
δ_{BS}	The incremental units for the BS cache

decided to remain in the local cache are designated by $\overline{S_c}$. In our network architecture at each caching decision phase, we want to minimize prospective energy consumption. Therefore, we sustain content unit set $\overline{S_c}$ that would induce the largest total energy expenditure. This way, content units with lower energy consumption will be evicted first and in their prospective requests, they will be fetched from other system components

with less energy cost. The optimization problem is as follows:

$$max \sum_{u \in S_c} E_{all}^{(u)} \; 1_{[u \in \overline{S_c}]} \tag{5.1}$$

$$s.t. \;\; s_{u'} + \sum_{u \in S_c} s_u \; 1_{[u \in \overline{S_c}]} \leq C_{Dev}^{cache} \tag{5.2}$$

where the indicator function $1_{[x]}$ is set to one if x is true and zero otherwise. Our objective function $\sum_{u \in S_c} E_{all}^{(u)} \; 1_{[u \in \overline{S_c}]}$ in (5.1) gives the total energy consumption of units across the network that are not evicted from the cache set S_c and residing in $\overline{S_c}$. With constraint (5.2), it is realized that the total size of new unit u' and the units not evicted from the cache does not exceed the cache capacity. We formulate the prospective energy consumption of any content unit u as $E_{all}^{(u)}$ in Equation 5.8 that is branched into the four scenarios that are shown in Figure 3.1: (i) Local hit, (ii) Transmission via D2D technique, (iii) Transmission from the BS cache, (iv) Transmission from the universal source across the BS

For the calculation of the prospective expected energy consumption $E_{all}^{(u)}$ of any *content unit* u that is mapped to some *content unit tuple* $\{i, j, k\}$, we utilize content unit availability probabilities of different system components. The availability probability in the local cache of a requester of some arbitrary content unit u (with the one-to-one and onto correspondent *content unit tuple* $\{i, j, k\}$) is defined in Equation 5.3 as $p_{loc}^{(u)}$. The multiplication of the i^{th} content request probability p_i, the j^{th} chunk request probability p_j and the k^{th} layer request probability p_k gives request probability of *content unit tuple* $\{i, j, k\}$ (*content unit* u). Aside from the request characteristic, the requester device storage capability is eminently prominent for the calculation of the content unit availability probability at a requester device. For the integration of this aspect, the device capacity C_{Dev}^{cache} is normalized over the total size of all content space

($\sum_{all} s_u$) by $\mathbb{F}_{loc}^{fit}$ function in Equation 5.4.

$$p_{loc}^{(u)} := p_i \cdot p_j \cdot p_k \cdot \mathbb{F}_{loc}^{fit} \tag{5.3}$$

$$\mathbb{F}_{loc}^{fit} := \frac{C_{Dev}^{cache}}{\sum_{all} s_u} \tag{5.4}$$

The availability probability of some arbitrary content unit u in the BS cache $p_{BS}^{(u)}$ in Equation 5.5 is calculated similarly. The corresponding normalization function $\mathbb{F}_{BS}^{fit}$ that outputs the BS storage capability is given in Equation 5.6.

$$p_{BS}^{(u)} := p_i \cdot p_j \cdot p_k \cdot \mathbb{F}_{BS}^{fit} \tag{5.5}$$

$$\mathbb{F}_{BS}^{fit} := \frac{C_{BS}^{cache}}{\sum_{all} s_u} \tag{5.6}$$

N_{ngh} is defined as the total number of devices that are at most R_{D2D} away from a requester. The availability probability of some content unit u in at least one of the neighbour devices of a requester $p_{D2D}^{(u)}$ is defined in Equation 5.7. $(1 - p_{loc}^{(u)})^{N_{ngh}}$ gives the probability of a requested content unit u not being available in any neighbouring device of a requester. By the complement, we obtain the availability probability of some content unit u being in at least one of the neighbours of a requester device.

$$p_{D2D}^{(u)} := 1 - (1 - p_{loc}^{(u)})^{N_{ngh}} \tag{5.7}$$

We formulate the prospective energy consumption level of an arbitrary content unit u in Equation 5.8.

$$\begin{aligned} E_{all}^{(u)} :=& p_{loc}^{(u)} E_{loc}^{(u)} + (1 - p_{loc}^{(u)}) p_{D2D}^{(u)} E_{D2D}^{(u)} + (1 - p_{loc}^{(u)})(1 - p_{D2D}^{(u)}) p_{BS}^{(u)} E_{BS}^{(u)} + \\ & (1 - p_{loc}^{(u)})(1 - p_{D2D}^{(u)})(1 - p_{BS}^{(u)}) E_{BS(U)}^{(u)} \end{aligned} \tag{5.8}$$

This energy consumption is branched into the following enumerated scenarios that are shown in Figure 5.1.

(i) Local hit
(ii) Transmission via D2D technique
(iii) Transmission from the BS cache
(iv) Transmission from the universal source across the BS

With the probability $p_{loc}^{(u)}$ the first scenario occurs and unit will be found in the local cache and $E_{loc}^{(u)}$ will be the corresponding prospective energy expenditure for the local hit of content unit u. The second scenario consumes $E_{D2D}^{(u)}$ energy when the content unit will not be available in the local cache but retrievable from some other device with probability $(1-p_{loc}^{(u)})p_{D2D}^{(u)}$. If no device in the transmission range (centered at the requester with radius R_{D2D}) can serve due to the lack of unit, the BS takes action and will serve the content unit from the BS cache with the energy expenditure level $E_{BS}^{(u)}$. Unless the relevant unit is in the BS cache, the transmission path from the universal source to the BS cache and then from there to the requester device will be utilized with $(1-p_{loc}^{(u)})(1-p_{D2D}^{(u)})(1-p_{BS}^{(u)})$ probability and $E_{BS(U)}^{(u)}$ energy will be consumed. The overall expected energy expenditure of unit u based on listed four scenarios is given in Equation 5.8. The expected energy expenditure of the given four scenarios are listed as follows:

$$E_{loc}^{(u)} := P_{loc}^{u} \cdot \frac{s_u}{C_{loc}} \tag{5.9}$$

$$E_{D2D}^{(u)} := P_{D2D}^{tx} \cdot \frac{s_u}{C_{D2D}} \tag{5.10}$$

$$E_{BS}^{(u)} := P_{BS}^{tx} \cdot \frac{s_u}{C_{BS}} \tag{5.11}$$

$$E_{BS(U)}^{(u)} := [P_{BS}^{rec} \cdot \frac{s_u}{C_{BS(U)}}] + E_{BS}^{(u)} \tag{5.12}$$

The expected energy expenditures are calculated by the product of power in some service type and its expected service duration for a given content unit. During local hits, the device consumes the lowest amount of energy compared to other system com-

ponents. We define the power level P^u_{loc} as $\frac{P^{tx}_{D2D}}{\theta_{loc}}$ while the expected service duration is $\frac{s_u}{C_{loc}}$. The largest service rate among the system components is C_{loc}. The product of P^u_{loc} and $\frac{s_u}{C_{loc}}$ gives the expected local hit energy consumption $E^{(u)}_{loc}$ for some content unit u as shown in Equation 5.9. The product of the device transmission power P^{tx}_{D2D} and expected prospective D2D service duration $\frac{s_u}{C_{D2D}}$ gives the D2D expected transmission energy $E^{(u)}_{D2D}$ (in Equation 5.10) for unit u. C_{D2D} is calculated taking average distance $\frac{R_{D2D}}{2}$ for future D2D transmission. Note that devices do not know the cache status of other devices and hence for future predictions they tune to average values for the calculation of prospective energy expenditures. In Equation 5.11, the expected energy consumption for a transmission of some unit u from the BS cache is calculated similar to the $E^{(u)}_{D2D}$. When we consider the fourth scenario, the content unit retrieval basically branches into two parts *i)* content unit retrieval from the universal source to the BS cache, *ii)* from the BS cache to requester device. The retrieval to BS cache results in expected $P^{rec}_{BS} \cdot \frac{s_u}{C_{BS(U)}}$ energy expenditure while the rest has $E^{(u)}_{BS}$ as given in Equation 5.12. For the content unit reception to the BS cache, the BS reception power is $P^{rec}_{BS} := \frac{P^{tx}_{BS}}{\theta_{BS}}$ and the expected future reception period for some unit u is $\frac{s_u}{C_{BS(U)}}$.

Dynamic programming algorithm is one of the techniques to solve the 0-1 knapsack problem optimally [128]. In our optimal solution, we utilize dynamic programming given in Figure 5.3 and further illustrated in Figure 5.4. In this optimal (OPT) algorithm, if the cache capacity C^{cache}_{Dev} does not suffice for the residing content units set S_c and the newly requested one c_{new}, then a subset of S_c will be evicted for the sake of newly arrived unit c_{new}. The core idea is to keep content units in the local cache that would otherwise lead to large prospective energy burden. We consider the cache as the knapsack. Each content unit u is an item that is weighted by its size s_u and valued by its energy consumption $E^{(u)}_{all}$. In dynamic programming, we break the optimization problem into subproblems, recursively solve these subproblems and store their results in E_{cum} array as shown in Figure 5.3. By backtracking E_{cum}, we select the units with the greatest energy sum to continue residing in the cache. These units are designated by the set $\overline{S_c}$. However, the time complexity of this algorithm is $O(|S_c| \cdot W)$ where W is the residual cache capacity after the new unit is inserted and $|S_c|$ is the number of

INPUTS

$\boldsymbol{S_c}$: The content units set residing in the device local cache

$\boldsymbol{c_{new}}$: The newly requested content unit, $\boldsymbol{C_{Dev}^{cache}}$: The device cache capacity

OPT$(\boldsymbol{S_c}, \boldsymbol{c_{new}}, \boldsymbol{C_{Dev}^{cache}})\{$

$Capacity \leftarrow TotalSize(S_c)$;

if $(Capacity + s_{c_{new}} \leq C_{Dev}^{cache})$ **then**

  return $S_c \cup \{c_{new}\}$; %the set kept in the cache

else

  %traverse all content units in the local device cache

  for $(i = 1 : 1 : |S_c|)$ **do**

    %traverse weights with δ_{Dev} (= 0.01 Mbs) incremental units

    for $j = 1 : 1 : \left\lfloor \frac{Capacity - s_{c_{new}}}{\delta_{Dev}} \right\rfloor$ **do**

      if $((i == 1) \vee (j == 1))$ **then**

        $E_{cum}(i, j) \leftarrow 0$; %first row and column are set to zero

      else if $(s_i < j)$ **then**

        %take maximum cumulative prospective energy consumption of in-cache units by either excluding i^{th} unit or including it

        $E_{cum}(i, j) \leftarrow max\{E_{cum}(i-1, j), E_{all}^{(i)} + E_{cum}(i-1, j-s_i)\}$;

      end

      else

        $E_{cum}(i, j) \leftarrow E_{cum}(i-1, j)$;

      end

    end

  end

  %start from the last row and column of E_{cum} backtrack and mark units that will continue to reside in the local device cache as $\overline{S_c}$

  $\overline{S_c} \leftarrow$ backtrack(E_{cum});

  return $\overline{S_c} \cup \{c_{new}\}$;

end

}

Figure 5.3. Dynamic programming caching algorithm.

content units in a given device cache [127]. This complexity does not provide a practical calculation time. Therefore, we propose a new energy prioritized D2D caching heuristic given in Figure 5.5.

The main motivation of this algorithm is to reduce the energy burden on the network by utilizing our analytically calculated expected energy consumption of units $E_{all}^{(u)}$. If the cache capacity C_{Dev}^{cache} does not suffice for the residing content units set S_c and the newly requested one c_{new}, eviction takes place. At the eviction decision phase, all content units in the set S_c are sorted based on their expected prospective energy costs $E_{all}^{(u)}$ in descending order. As shown in Figure 5.5 starting from the least energy

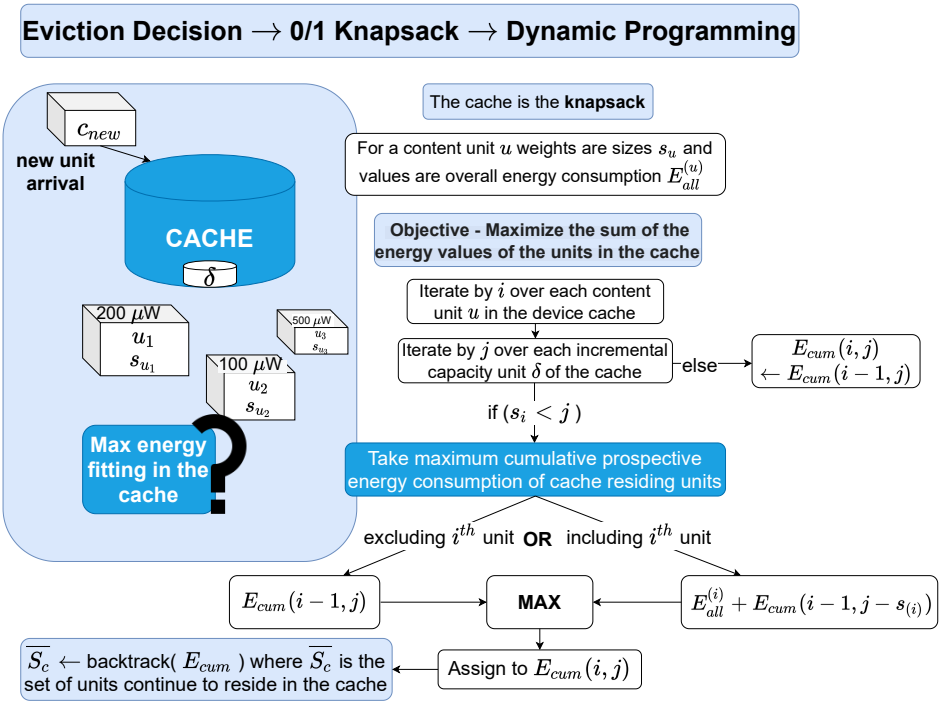

Figure 5.4. Dynamic programming for solving 0/1 caching knapsack problem.

consuming unit c_k, content units are eliminated from the cache until the free space is enough for storing new unit c_{new}. Unit(s) with low energy cost across the network are eliminated first and large energy consuming units over the network are sustained in local cache for less energy burden in future requests. In other words, the units with large prospective energy burden across the network are prioritized to be stored locally. Therefore, this scheme is named as energy prioritized D2D caching (EPDC). The time complexity of this procedure is bounded by the sorting procedure of elements in the set S_c based on their $E_{all}^{(u)}$ values. It is $O(|S_c| \cdot log|S_c|)$ with $|S_c|$ denoting the number of content units in a given device cache.

5.2.1. Time complexity analysis

Based on our complexity analysis, OPT algorithm has infeasible operation time. As shown in Figure 5.3, for all cached units OPT has an iteration range by the cache capacity with δ incremental units. For devices we take $\delta_{Dev} = 0.01$ Mbs and for the BS cache $\delta_{BS} = 0.1$ Mbs respectively. The average time consumed for devices and the BS cache by all techniques is provided in Table 5.2. Note that we test the time complexity

INPUTS

$\boldsymbol{S_c}$: The content units set residing in the device local cache

$\boldsymbol{c_{new}}$: The newly requested content unit

$\boldsymbol{C_{Dev}^{cache}}$: The device cache capacity

EPDC($\boldsymbol{S_c, c_{new}, C_{Dev}^{cache}}$){

$Capacity \leftarrow TotalSize(S_c)$;

if *($Capacity + s_{c_{new}} \leq C_{Dev}^{cache}$)* **then**
  return $S_c \cup \{c_{new}\}$; %the set kept in the cache

else
  %sort units in S_c based on their E_{all} values (in (5.8)) in descending order
  %$S_{sort} = \{c_1, c_2, ..., c_k\}$ with c_1 having largest E_{all}, c_k lowest
  $S_{sort} \leftarrow$ sort(S_c, E_{all}, DESCEND);
  j = k;
  while *($j \geq 1$)* **do**
    $S_{sort} \leftarrow S_{sort} \setminus \{c_j, c_{j+1}, ...c_k\}$;
    if *($Capacity + s_{c_{new}} - \sum_{\theta=j}^{k} s_{c_\theta} \leq C_{Dev}^{cache}$)* **then**
      return $S_{sort} \cup \{c_{new}\}$; %the set decided to be kept in the cache
    end
    $j \leftarrow j - 1$;
  end

end

}

Figure 5.5. Energy prioritized D2D caching algorithm (EPDC).

in an laptop device with Intel core i7-6500 CPU @2.5GHz and 8 GB RAM in MATLAB R2019 environment. We apply any chosen algorithm not only in devices but also in the BS cache for the sake of completeness and hence due to the large BS cache capacity even in a setup with moderate number of contents, we present the impracticality of applying OPT in real-time or near real-time scenarios. With $C_{Dev}^{cache} = 150$ Mbs and $C_{BS}^{cache} = 2.8$ Gbs even with 150 contents, OPT performs poorer than $EPDC$ in terms of service time especially for the BS caching. This is because $EPDC$ algorithm is only bounded by the sorting of content units but not cache capacity while OPT algorithm acts an iteration over the cache capacity for each in-cache content unit.

Table 5.2. Time complexity results for different caching techniques.

Technique	**Total Local Caching Time (s)**	**Total BS Caching Time (s)**
LRU	0.57	1.11
PDC	0.53	2.65
SXO	0.82	9.69
OPT	18.94	7319.11
$EPDC$	6.01	5.80

5.3. Performance Metrics

We analyze optimal scalable video caching algorithm (OPT) in D2D cellular networks with backhaul support and energy prioritized D2D caching ($EPDC$) algorithm in terms of *service rate* and *energy expenditure*. The system parameters are listed in Table 5.3. We compare OPT and $EPDC$ to the commonly utilized classical algorithm Least Recently Used (LRU) that does not consider content features. LRU is used as a baseline to compare the improvement of our algorithms over a non-content aware cache replacement algorithm. In the literature, popularity based caching techniques are widely employed as well [67, 93, 94]. Accordingly, we utilize the Popularity-driven caching (PDC) algorithm as another comparison technique [1], that is aware of content attribute popularity and manages the eviction decisions with the aim of keeping popular contents in cache with a greater *probability*. We compare our algorithms to the SXO technique as well [76]. In SXO, content features size and order (access rate) are taken into account for the eviction decision of the cache. Large contents with low access rate are evicted first. We compare our algorithms to content-aware cache replacement techniques PDC and SXO since multimedia content caching is our main focus.

We elaborate on the service rate and energy consumption in different operation modes rigorously. The energy consumption of our system consists of the following elements:

- E_{loc} for local hits
- E_{D2D} for D2D transmission

Table 5.3. System parameters for cellular D2D edge networks.

Parameter	**Explanation**
P_{BS}^{tx}	The transmission power consumption of the BS
P_{BS}^{rec}	The reception power consumption of the BS
$C_{BS}(n)$	The channel capacity between the BS and the n^{th} device
$C_{BS(U)}$	The capacity between the universal content repository and the BS
N_D	The total number of devices located in the cell
S_U	The set of content units uniquely identifiable by content, chunk, layer id
req_u	The request for the content unit u
s_u	The size of the content unit u
$S_{(n)}^{BS}$	The set of services from the BS to n^{th} device
$S_{(n)}^{BS(U)}$	The set of services from the universal content repository to the n^{th} device across the BS

- E_{BS} for direct services of the BS
- $E_{BS(U)}$ for indirect services of the BS
- E_{block} due to blocked services

E_{loc}, E_{D2D} and E_{block} are defined previously in Section 4.3. One of our new contributions in this chapter compared to the previous Chapter 4 is the BS integration to our D2D network architecture. Therefore, we need to define the BS energy expenditure branched into *i) direct services* and *ii) indirect services*. For these, E_{BS} and $E_{BS(U)}$ are defined in Equations 5.13 and 5.14, respectively. E_{BS} is the overall transmission energy consumption of the BS serving requests directly from the BS cache whereas $E_{BS(U)}$ is the transmission energy consumption for services from the universal source across the BS to requesters.

$$E_{BS} := \sum_{u \in S_U} \sum_{n \in N_D} \sum_{req_u \in S_{(n)}^{BS}} P_{BS}^{tx} \cdot \frac{|s_u|}{C_{BS}(n)} \tag{5.13}$$

$$E_{BS(U)} := \sum_{u \in S_U} \sum_{n \in N_D} \sum_{req_u \in S_{(n)}^{BS(U)}} \frac{P_{BS}^{tx} \cdot |s_u|}{C_{BS}(n)} + \frac{P_{BS}^{rec} \cdot |s_u|}{C_{BS(U)}} \tag{5.14}$$

By summing all of these expenditures as provided in Equation 5.15, we obtain E_{total} the total energy consumption of our system.

$$E_{total} := E_{loc} + E_{D2D} + E_{BS} + E_{BS(U)} + E_{block} \tag{5.15}$$

5.4. Performance Evaluation

We perform simulations for varying system parameters *cache capacity* and *D2D transmission radius*, and investigate how they impact the network in terms of *service capacity* and *energy consumption*. We compare $EPDC$ and OPT algorithms to the LRU, PDC [1] and SXO [76] techniques. The simulation parameters are provided in Table 5.4. In this table, the content unit popularity and request parameters (s, λ, k, p_k), the power consumption of system components (local, BS, etc.), local hit and BS reception power tune parameters (θ_{loc}, θ_{BS}), channel parameters (B, N_0, d_0, n_{D2D}, n_{BS}) are defined. Device and BS cache capacities (C_{Dev}^{cache}, C_{BS}^{cache}), device density $\lambda_{N_{HU}}$ and the radius for the reception range of a requester in D2D mode R_{D2D} are also remarkable system parameters. Based on the observations in practical systems, the BS transmission/reception powers are at the order of Watts quite a lot larger than the local and D2D power consumptions (in mW scale only). Mostly, the BS operations last longer than the D2D mode transmissions. Both the large power consumption and the long transmission time of the BS dictate it as the major network-wide energy consuming component. This will be further investigated in the forthcoming subsections. The channel parameters B, d_0 and n_{D2D} are adopted from our study [4] that is demonstrated in Chapter 4. We take the path loss exponent of the BS transmissions n_{BS} greater than that of the D2D mode n_{D2D} as broadly employed in the literature [129, 130].

We focus on the caching of multimedia contents. We simulate content requests based on the Zipf distributed popularity scheme. Contents are divided into chunks and inter-chunk popularity is demonstrated by the Weibull distribution. Our multimedia contents are layered into base and enhancement parts and the high quality (HQ) content consumption together with both layers is actualized by the HQ request probability

p_k. We have an event-based simulator that processes content unit request arrivals and service completions. For this simulator, network operations need power and channel parameters that are provided in Table 5.4 to actualize the arrivals. In each network scenario, we run simulations 10 times and take their average. In the upcoming subsections, we vary the device cache capacity C_{Dev}^{cache} and the radius for the reception range of a requester in D2D mode R_{D2D} to demonstrate how the D2D cache and service capability affect the network.

Table 5.4. Simulation parameters for cellular D2D edge network.

Symbol	**Explanation**	**Value**
T_{sim}	The simulation duration	400 s
α	The Zipf distribution skewness parameter	1
λ	The Weibull distribution scale parameter	5
k	The Weibull distribution shape parameter	0.8
p_{HQ}	The ratio of high quality consumers	1
P_{loc}^{u}	The power consumption of local content unit retrieval	40 mW
P_{D2D}^{tx}	The transmission power consumption of a device	80 mW
P_{BS}^{tx}	The transmission power consumption of the BS	6 W
P_{BS}^{rec}	The reception power consumption of the BS	1.2 W
θ_{loc}	The parameter for device local hit power consumption (The local hit power consumption of a device is $\frac{P^{tx}_{dev}}{\theta_{loc}}$)	2
θ_{BS}	The parameter per channel BS reception power consumption (Per channel reception power of the BS is $\frac{P^{ch}_{BS}}{\theta_{BS}}$)	5
C_{Dev}^{cache}	The device cache capacity	150 Mbits
C_{BS}^{cache}	The base station cache capacity	2.8 Gbits
B	The terrestrial channel operation bandwidth	2 MHz
N_0	The noise power density	-158 dBm
d_0	The reference distance of device antenna	1 m
n_{D2D}	The path loss exponent of D2D transmission	3
n_{BS}	The path loss exponent of BS transmission	4.2
λ_{users}	The mean density of user devices located in a cell according to Poisson Point Process	0.0015 $\frac{user}{m^2}$
R_{D2D}	The radius for the reception range of a requester in D2D mode	200 m
$\alpha_{density}$	The interplay multiplier for changing device density in a cell	1

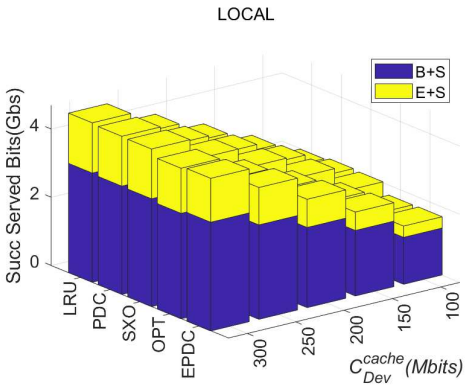

Figure 5.6. Locally served content units (B: of base layer, E: of enhancement layer, S: successful).

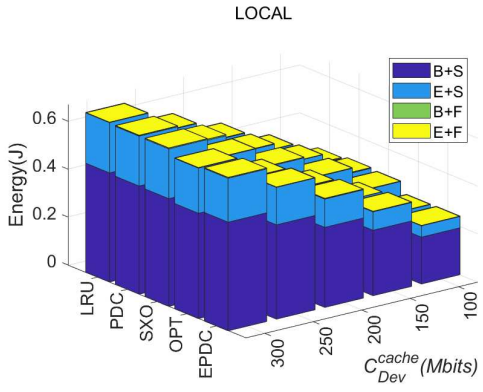

Figure 5.7. The energy consumed for local hits (B: of base layer, E: of enhancement layer, S: successful, F: fail).

5.4.1. Impact of device cache capacity C_{Dev}^{cache}

In this subsection, we investigate a key component in our system, the device cache capacity C_{Dev}^{cache}. We enlarge/shrink the cache capacity C_{Dev}^{cache} to observe how it impacts the caching techniques.

First, we compare our caching technique $EPDC$ with the OPT and the alternative strategies LRU, PDC and SXO in terms of the local hit service capacity and energy consumption. For any fixed C_{Dev}^{cache} value, the energy optimizing OPT attains the smallest local service capacity and in return the lowest local energy consumption given in Figure 5.6 and 5.7, respectively. Compared to the optimal case, our heuristic $EPDC$ is slightly worse in terms of energy consumption for local hits. However, the energy improvement over the LRU, PDC and SXO manifests the utilitarian nature of our proposal for local hits. This improvement of $EPDC$ over the alternatives becomes greater for decreased cache capacity. In the large cache capacity regime, the cache capacity has a greater impact on the energy expenditure as requested contents can be found locally with a greater probability. With decreasing cache capacity, the impact of cache replacement procedure can be observed more clearly. For $C_{Dev}^{cache} = 300$ Mbits the local hit energy consumption improvement rate of $EPDC$ over LRU (PDC / SXO)

is 4.0% from 0.67 to 0.64 J (3.5% from 0.66 to 0.64 J / 3.2% from 0.66 to 0.64 J). With decreasing C_{Dev}^{cache} to 100 Mbits, this improvement of $EPDC$ over LRU (PDC) rises up to 14.4% from 0.29 to 0.24 J (17.0% from 0.29 to 0.24 J). Similarly, the local hit energy improvement rate of $EPDC$ over SXO technique reaches 15.4%.

In all caching algorithms, with increasing device capacity C_{Dev}^{cache}, the direct BS service usage declines and accordingly less direct BS energy is consumed. The reason is that large device caches entail increased local and D2D service utilization. For the comparison of our heuristic $EPDC$ to other policies, in any C_{Dev}^{cache} value, our $EPDC$ technique achieves less direct BS mode energy consumption than LRU, PDC and SXO. The largest observed direct BS energy improvement of $EPDC$ over LRU (PDC) is 7.3% from 1.65 to 1.53 kJ (12.5% from 1.75 to 1.53 kJ) at $C_{Dev}^{cache} = 200$ Mbits. This improvement of $EPDC$ over SXO reaches up to 12.4% from 0.99 to 0.87 kJ at $C_{Dev}^{cache} = 300$ Mbits. $EPDC$ falls back from the OPT for the direct BS energy with at most 6.0% decrease at $C_{Dev}^{cache} = 100$ Mbits. Despite this slight degradation, the evident improvement of our proposal $EPDC$ over LRU, PDC and SXO demonstrates its benefit in terms of direct BS energy consumption.

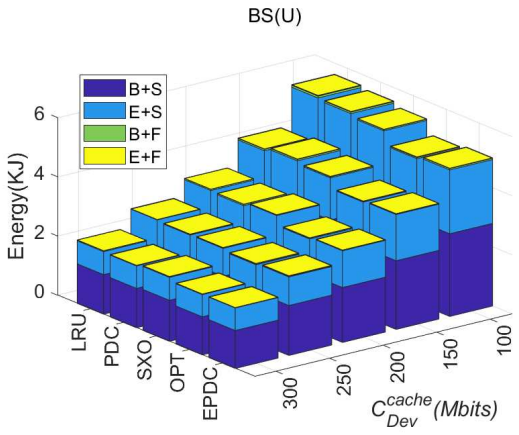

Figure 5.8. The energy consumed for retrievals in BS(U) mode (B: of base layer, E: of enhancement layer, S: successful, F: fail).

For indirect BS services, a similar reasoning with direct BS services and corresponding energy consumption is valid. For all C_{Dev}^{cache}'s in the observation domain, $EPDC$ achieves an energy reduction of the BS than alternatives LRU, PDC and SXO

regarding the indirect BS services as shown in Figure 5.8. The improvement of indirect BS energy consumption of $EPDC$ over LRU (PDC) reaches its maximum rate 11.1% from 5.63 to 5.00 kJ (10.8% from 4.38 to 3.91 kJ) for $C_{Dev}^{cache} = 100$ (150) Mbits. We also monitor the highest improvement of $EPDC$ over SXO as 8.1% from 3.40 to 3.12 kJ at $C_{Dev}^{cache} = 200$ Mbits. $EPDC$ also achieves indirect BS energy results close to the OPT with at most 1.6% difference. Hence, it is practical for the energy consumption reduction in the BS.

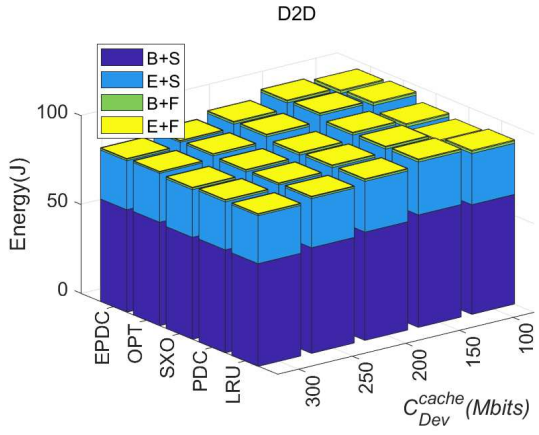

Figure 5.9. The energy consumed for retrievals in D2D mode (B: of base layer, E: of enhancement layer, S: successful, F: fail).

For the D2D mode analysis, the corresponding service capacity is not affected by varying device cache capacity C_{Dev}^{cache} evidently between different policies. In accordance with the D2D service, its energy consumption does not change by the varying C_{Dev}^{cache} for any caching algorithm (in Figure 5.9). The varying cache capacity essentially impacts the local hit results. When we compare $EPDC$ to the techniques LRU, PDC and SXO, we observe a slight increase in D2D service capacity for any C_{Dev}^{cache}. Accordingly, the D2D energy consumption for $EPDC$ is slightly greater (its performance is slightly worse) than the alternative approaches. In terms of D2D energy consumption, the highest rise of $EPDC$ (its performance degradation) compared to LRU (PDC) is 3.6% from 92.43 to 95.78 J (5.6% from 90.49 to 95.78 J) for $C_{Dev}^{cache} = 100$ Mbits (in Figure 5.9). $EPDC$ also performs 4.0% worse than SXO. The alternatives do not consider prospective energy consumption regarding different operation modes (local, D2D, etc.). On the contrary, our $EPDC$ algorithm utilizes them and we observe an increase in D2D service capacity and corresponding energy consumption in $EPDC$

compared to the alternatives. However, this increase in not as evident as expected. The large cache capacity of the BS and its high power level compared to devices are more dominating determining factors for the total energy consumption of $EPDC$. Therefore, $EPDC$ chokes BS selection rather that the D2D mode proliferation and thus no significant difference in D2D preference between $EPDC$ and other policies is observed.

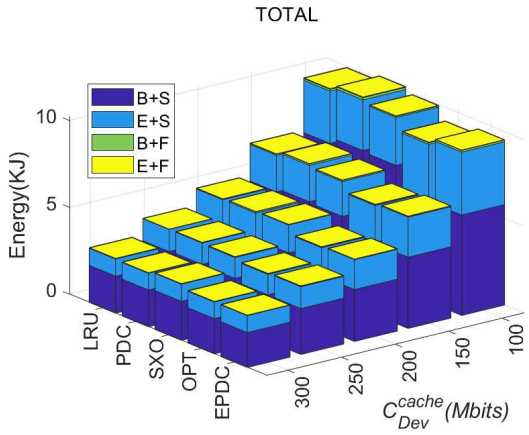

Figure 5.10. The total energy consumed for retrievals and local hits (B: of base layer, E: of enhancement layer, S: successful, F: fail).

Finally, we elaborate on the total service capacity and corresponding energy consumption of the network. As already mentioned, in different caching algorithms the D2D service capacity is not affected by the varying C_{Dev}^{cache}. However, with increasing C_{Dev}^{cache}, the local service capacity rises (in Figure 5.6) while the BS service capacity falls in all caching algorithms. These changes in service capacities balance each other and the total service capacity is not changed evidently for varying C_{Dev}^{cache} in each caching algorithm. As shown in Figure 5.10, for increasing C_{Dev}^{cache}, the total network energy consumption declines in all investigated caching algorithms. Thus, we deduce that with increasing device capacity the network-wide energy efficiency is improved.

Our proposal $EPDC$ falls back from the OPT at most 3.5% (for $C_{Dev}^{cache} = 100$ Mbits) in the investigated device cache capacity range. The total energy result of $EPDC$ converges to that of the optimal policy especially for large C_{Dev}^{cache}'s. The measured total network service capacity does not change significantly between different caching algorithms, for any C_{Dev}^{cache} values. Additionally, the observed total network

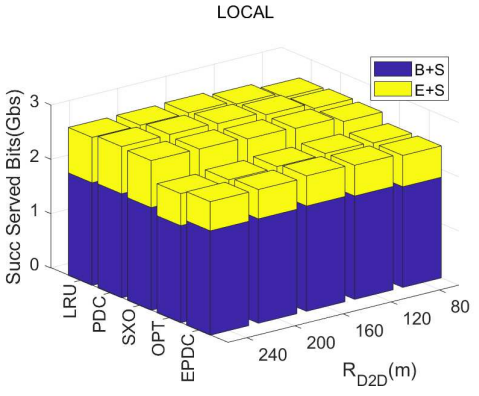

Figure 5.11. Locally served content units (B: of base layer, E: of enhancement layer, S: successful).

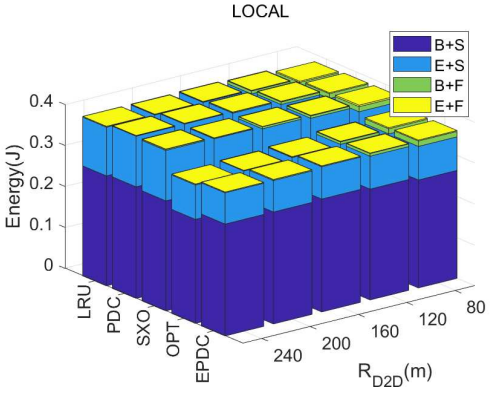

Figure 5.12. The energy consumed for local hits (B: of base layer, E: of enhancement layer, S: successful, F: fail).

energy of our $EPDC$ is lower and thus it is more energy-efficient than alternative approaches LRU, PDC and SXO (for any C_{Dev}^{cache}) as demonstrated in Figure 5.10. The greatest improvement of our proposal $EPDC$ compared to the LRU (SXO) is 6.7% from 5.08 to 4.74 kJ (8.4% from 5.18 to 4.74 kJ) for $C_{Dev}^{cache} = 200$ Mbits in terms of total network energy. Its improvement over PDC reaches the maximum at $C_{Dev}^{cache} = 150$ Mbits with 9.6% (from 7.15 to 6.46 kJ). These results present the beneficial nature of our proposed caching algorithm $EPDC$ for energy-sensitive network scenarios.

5.4.2. Impact of the radius for the reception range of a requester in D2D mode R_{D2D}

We elaborate on the impact of different D2D operations settings. In that regard, we vary the radius for the reception range of a requester in D2D mode R_{D2D}.

For different caching algorithms, we investigate the local hit service capacity and its energy consumption. In Figure 5.11, in any caching technique the local hit service capacity changes negligibly with varying R_{D2D}. When a local hit occurs, other network resources are not exploited and hence the radius component R_{D2D} of the D2D mechanism does not affect the local hit service capacity. Similarly, varying R_{D2D} does

not impact the local hit energy consumption in any caching policy (Figure 5.12).

When we monitor the results of our proposal $EPDC$ and alternative algorithms LRU, PDC and SXO, for any fixed R_{D2D} in the inspection range, $EPDC$ attains lower local service capacity and accordingly it has improvement over the alternatives in local energy consumption. The greatest local energy improvement of our proposal $EPDC$ over LRU (PDC) reaches 9.8% from 0.39 to 0.35 J (11.6% from 0.40 to 0.35 J) at $R_{D2D} = 240$ m (in Figure 5.12). The maximum local energy improvement of $EPDC$ over SXO is 11.1% from 0.39 to 0.35 J. $EPDC$ achieves close results to OPT. Even its worst performance degradation compared to the optimal caching OPT is 3.3% only (from 0.35 to 0.34 J) at $R_{D2D} = 240$ m in terms of local energy consumption. Consequently, the local energy expenditure improvement of $EPDC$ over LRU, PDC and SXO techniques indicates its beneficial aspect in terms of local services.

In our analysis on the direct BS services, with increasing R_{D2D} the energy consumption declines. The reason is that the probability of finding requested units in neighbour devices rises and subsequently the D2D service allocations are employed rather than BS services. Therefore, the direct BS service capacity and its energy consumption reduces in any caching policy. Similarly a decrease for indirect BS service capacity and its energy consumption (in Figure 5.13) are observed with increasing R_{D2D} at all caching techniques.

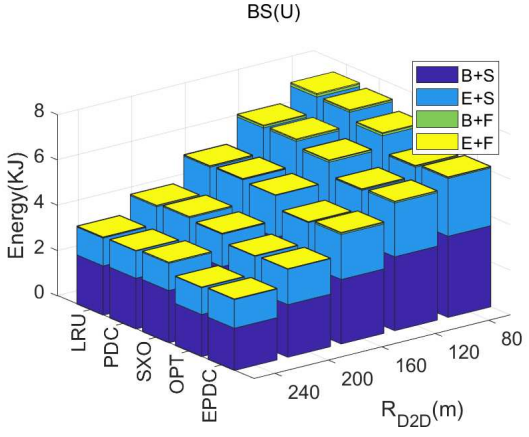

Figure 5.13. The energy consumed for retrievals in BS(U) mode (B: of base layer, E: of enhancement layer, S: successful, F: fail).

For any R_{D2D}, $EPDC$ hits lower rate than the alternatives LRU, PDC and SXO for the services supplied from the universal source across the BS (indirect BS services). It also achieves improvement in indirect BS energy consumption over LRU, PDC and SXO techniques as shown in Figure 5.13. The reason for this is that the optimal policy and our proposal $EPDC$ are designed for reducing the total network energy consumption and for this purpose they omit long routes of large power (from the universal source across the BS to requesters). For direct BS service energy, at the largest $R_{D2D} = 240$ m the improvement of $EPDC$ over LRU (SXO) is 12.2% (10.2%). Moreover, this energy improvement over PDC raises up to 19.4% at the largest $R_{D2D} =$ 240 m. With decreasing R_{D2D}, our proposal $EPDC$ consumes more direct BS energy than LRU, PDC and SXO algorithms. The reason is that for $EPDC$, with decreasing R_{D2D} D2D service capacity decreases and this enforces BS mode selection and thus BS energy consumption elevates.

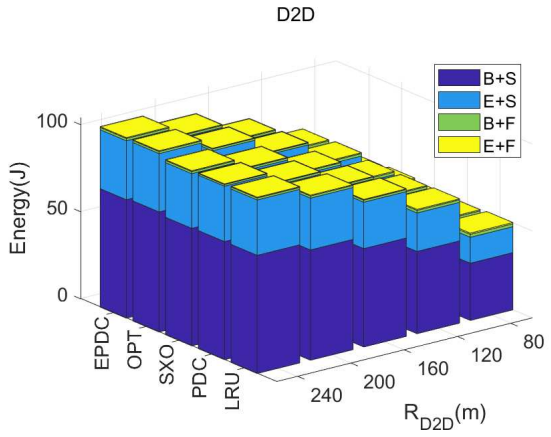

Figure 5.14. The energy consumed for retrievals in D2D mode (B: of base layer, E: of enhancement layer, S: successful, F: fail).

For all caching algorithms, the D2D service capacity and its energy consumption rise with increasing R_{D2D} as illustrated in Figure 5.14. The reason for the D2D service capacity improvement is that with increasing R_{D2D} the number of devices in reception range of requesters increases and consequently a higher probability of finding requested content unit(s) in a neighbour device is achieved according to Equation 5.7. Our algorithm $EPDC$ achieves only slightly higher D2D rate and slightly larger D2D energy consumption compared to alternatives (in Figure 5.14). As explained in the previous subsection, the high power of the BS is instrumental for determining the total network

energy and our proposal $EPDC$ algorithm chokes the BS selection rather than boosting the D2D mode selection. As a result, just slight change in D2D usage is observed between different algorithms.

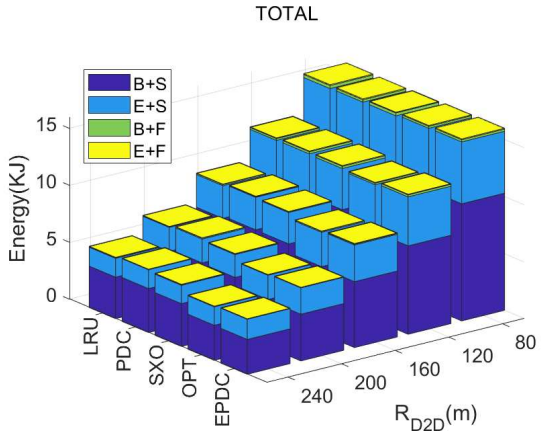

Figure 5.15. The total energy consumed for retrievals and local hits (B: of base layer, E: of enhancement layer, S: successful, F: fail).

We interrogate how the total service capacity and energy consumption of the network are affected by varying D2D radius R_{D2D}. In all caching algorithms, the total network service capacity rises with increasing R_{D2D} from 80 m to 120 m but it becomes saturated with further R_{D2D} increase. The initial service capacity rise is observed due to the D2D service capacity improvement that is not canceled by the BS service capacity drop. However, with further rise in R_{D2D} D2D is canceled due to its slower improvement. With increasing R_{D2D}, the BS energy expenditure decreases, the D2D energy consumption increases and the local hit expenditure is stagnant in any caching algorithm. The BS services take longer service time with higher power than D2D mechanism. Thus, the BS is instrumental for determining the total network energy consumption with its great energy consumption and hence with increasing R_{D2D}'s, the total network energy decreases in any caching algorithm as illustrated in Figure 5.15.

We compare our $EPDC$ algorithm to LRU, PDC, SXO and the optimal OPT. For any fixed R_{D2D}, the total network service capacity does not change evidently between caching algorithms but the corresponding energy for $EPDC$ is less than LRU, PDC and SXO (in Figure 5.15). This total network energy improvement of $EPDC$ over alternatives LRU, PDC and SXO rises with increasing R_{D2D} and the greatest

improvement of $EPDC$ over LRU (SXO) is observed at $R_{D2D} = 240$ m with 9.5% from 5.38 to 4.87 kJ (with 10.6% 5.45 to 4.87 kJ). The maximum achieved improvement over PDC is 12.4%. $EPDC$ slightly performs worse than OPT with at most 2.4% performance degradation at $R_{D2D} = 240$ m. Despite this fall back, we observe improvement of $EPDC$ over the comparison algorithms LRU, PDC, SXO for $R_{D2D} \geq 120$ m. No evident change in total network service capacity is observed between different caching algorithms. Hence, especially for large R_{D2D}'s, our proposed $EPDC$ algorithm becomes more energy efficient than the alternative algorithms.

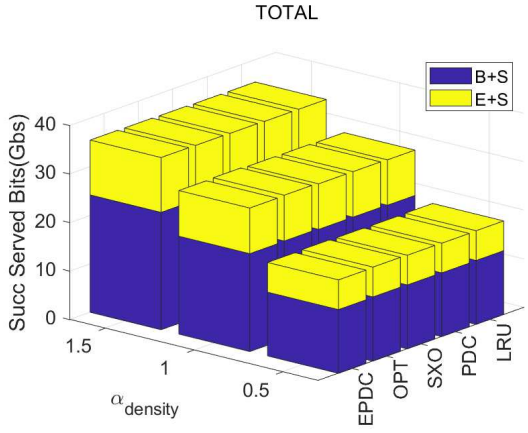

Figure 5.16. The total served content units (B: of base layer, E: of enhancement layer, S: successful).

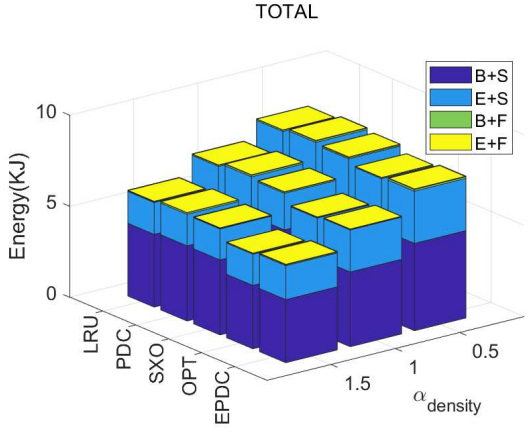

Figure 5.17. The total energy consumed for retrievals and local hits (B: of base layer, E: of enhancement layer, S: successful, F: fail).

5.4.3. Impact of the device density factor $\alpha_{density}$

The density of devices spread across the cell is a key determinant for the D2D operation mode selection. In that regard, we interplay with the number of devices in the cell by multiplying the mean density λ_{users} with density factor $\alpha_{density}$ and we investigate the total network service capacity and its energy consumption.

With increasing the device density by factor $\alpha_{density}$, in each caching algorithm we observe a rise in the number of requesters and hence increase in total network service as shown in Figure 5.16. For any $\alpha_{density}$, this total network service capacity does not change much between different caching algorithms. When we monitor the total network energy results in Figure 5.17, we see that with increasing device density the total network energy consumption decreases due to the increased D2D mode and decreased BS mode operations. With increasing device density by multiplying with factor $\alpha_{density}$ (from 0.5 to 1.5), it is shown that the improvement in total network energy consumption of $EPDC$ over the alternative approach LRU (SXO) increases from 2.3% to 9.2% (from 2.9% to 8.1%). We observe that $EPDC$ improvement over PDC algorithm is 9.6% (7.15 to 6.46 kJ) for $\alpha_{density}$ =1 and 9.2% (5.95 to 5.40 kJ) for $\alpha_{density}$ =1.5. For increasing device density, we observe the improvement of our proposal $EPDC$ in terms of total network energy over alternative algorithms expect for PDC. Besides, no evident change in service capacity between algorithms is observed and hence improved energy efficiency is achieved with increasing device density.

5.4.4. Discussion

In this chapter, we have formulated and solved the energy consumption minimizing optimal caching in cellular D2D edge networks regarding the multimedia delivery. Due to the feasibility concerns, we have proposed an energy prioritized D2D caching ($EPDC$) algorithm as well. $EPDC$ outperforms LRU, PDC and SXO algorithms in terms of network-wide EE especially for larger D2D service ranges. Moreover, the EE of $EPDC$ over LRU, SXO is promoted by increasing device density. In networking systems, the devices are battery operated while the BS is usually not. Accordingly,

a device and BS differentiating energy efficiency model can be developed as a future work. Our model assumptions are deterministic. However, in real system scenarios the content model, the traffic model and even the network model are not deterministic. In that regard, online model prediction approaches can be utilized to serve multimedia services in a more realistic manner owing to the fact that real systems are not static but rather dynamic.

6. CONCLUSIONS

In the future, 6G is envisioned to service extremely-high-definition multimedia content by high capacities [26]. The expected supported peak data rate for 6G is Tbps [131]. 6G is envisioned to serve ultra-low latency, improved spectral efficiency and EE as well. Regarding the network system, 6G is expected to operate for ultra heterogeneous networks. In that regard, mega satellite constellations, UAV systems and dense D2D communications will co-exist together and the heterogeneity in such networks is a motivator for the performance improvement. Motivated by this vision of the 6G, we explore satellite integrated content-centric and spectrum sharing Heterogeneous Networks (HetNets) that are high multimedia generating sources leading to intense burden on the network in terms of energy efficiency. This thesis provides an understanding of such networks in terms of modeling, resource allocation and caching.

For multimedia services, we modeled an in-network content caching integrated spectrum sharing HetNet with satellite and terrestrial links as a Continuous Time Markov Chain. For a more realistic network setup, universal source concept is introduced that allows access to contents that are out of the network range and it additionally preserves the system EE under highly loaded networks. We investigated the trade-off factors in our multi-mode HetNet and showed the EE improvement by increasing D2D mode weight. Furthermore, we have enabled D2D overlaying and observed the performance improvement in terms of both EE and goodput. We studied the connectivity mode assignment and formulated the EE maximizing mode assignment problem for the multi-mode HetNet architecture. Due to the feasibility, we developed a sub-optimal connectivity mode assignment that enables large D2D mode selection leading to an improved EE. In a nutshell, D2D is an instrumental communication method for improving EE performance for our multi-mode operating HetNet architecture and additional gains can be achieved by enabling overlaying in D2D services and exercising a sub-optimal connectivity mode assignments for the multi-mode HetNet.

According to the findings that reveal the importance of the D2D communication, we elaborated on the D2D edge network. For this network, we created a multi-dimensional multimedia model based on *i)* popularity, *ii)* chunking and *iii)* layering attributes. We proposed attribute-based cache replacement techniques with varying prioritization schemes. Based on the simulation results for different caching proposals, the chunking dimension is instrumental to alleviate the system energy burden while layering dimension has higher system goodput boosting capability. Moreover, both of our content attribute-based caching algorithms outperform the LRU in terms of EE. After the utilization of content attributes, the cooperation regarding the caching mechanism is another motivator for improved network performance. Therefore, we proposed availability/distance-based cooperative cache replacement techniques for the D2D edge network. The distance-based proposal has the greatest system goodput performance compared to the LRU, $MIN\text{-}ACC$, SXO and $CPPC$ algorithms with a trade-off factor increased energy consumption. Since multimedia services is the fundamental component for the HetNet and as a subcomponent D2D edge network, we utilize another multimedia attribute, namely scene change dynamics for alleviating the energy burden. At highly dynamic scene change regime, our availability-based proposal can be utilized that has much the same energy consumption as LRU (less energy consumption than $MIN\text{-}ACC$). To sum up, the content attributes are determining factors regarding the network performance for caching. By different attribute orders in prioritized caching mechanisms, we observe that chunking reduces energy consumption while layering improves goodput. For the cooperative caching, even though the scene change dynamic is not used at the caching decision phase, it is utilized as a motivator factor in terms of energy performance gain.

Content consumption is the building block of the HetNet. Therefore, we elaborated on the optimization for content services regarding the energy consumption minimization. We enhanced the D2D edge with cellular backhaul support for allowing access out of the HetNet for enabling a more realistic system setup. We constructed a comprehensive service-mode based energy model and defined content energy expenditures based on the prospective consumptions. The energy minimization problem is solved by the dynamic programming. Due to the feasibility concerns, we also proposed

an energy-cost model based *Energy Prioritized D2D Caching (EPDC)* algorithm. According to the simulations, $EPDC$ outperforms LRU, PDC and SXO algorithms in terms of EE especially for larger D2D service ranges. Besides, the EE of $EPDC$ over LRU and SXO is improved with increasing device density. Overall, the prioritization of content unit energy consumptions in $EPDC$ is instrumental for improving EE. Further EE gain is achievable under large D2D range and high device density scenarios.

6.1. Future Directions

For our holistic satellite integrated D2D enabled HetNet model, the services cease operation in case of a transmission failure or busy channel and this leads to decreased service capacity. To overcome this problem, a queuing based retrial mechanism can be integrated with an acceptable waiting time in a queue. This integration allows for improved modeling of system dynamics. For the cache replacement algorithms that we have proposed, the scene dynamics can be integrated at decision process as a future extension. Apart from this, we can elaborate on the proactive cache placement problem according to content attributes. Moreover, we have a simple fixed power allocation scheme. We can adopt to a dynamic power allocation to realize the potential of cellular D2D networks in terms of energy efficiency.

The dynamics of the HetNet environment is assumed to be static in our model. However, such environments are dynamic in real life and envisioned to be highly dynamic with mega satellite constellations, UAV systems and dense D2D deployments in the future with the arrival of 6G. Artificial intelligence (AI) is visioned to be an instrumental aspect of the 6G [132]. In that regard, the connectivity mode management can be handled by detecting the environment via AI and acting accordingly. In our cache management procedures, we have taken the content attributes quasi-stationary. For instance, the content popularity follows the Zipf distribution. However, the multimedia trends vary over time. Moreover, by the 6G vision extremely-high-definiton multimedia contents will be served in the future [26]. Owing to these facts, for adapting to more dynamic contents in time-domain and tremendously large multimedia services that will arise in the 6G era, online caching algorithms with reinforcement learning can

be exercised.

REFERENCES

1. Sinem Kafıloğlu, S., G. Gür, and F. Alagöz, "A Markovian model for satellite integrated cognitive and D2D HetNets", *Computer Networks*, Vol. 169, p. 107083, 2020.

2. Kafıloğlu, S. S., G. Gür, and F. Alagöz, "Analysis of Content-Oriented Heterogeneous Networks with D2D and Cognitive Communications", *CoRR*, Vol. abs/1808.01021, 2018.

3. Kafıloğlu, S. S., G. Gür, and F. Alagöz, "Connectivity Mode Management for User Devices in Heterogeneous D2D Networks", *IEEE Wireless Communications Letters*, Vol. 10, No. 1, pp. 194–198, 2021.

4. Kafıloğlu, S. S., G. Gür, and F. Alagöz, "Multidimensional Content Modeling and Caching in D2D Edge Networks", *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1–6, Sep. 2019.

5. Kafıloğlu, S. S., G. Gür, and F. Alagöz, "Cooperative Caching and Video Characteristics in D2D Edge Networks", *IEEE Communications Letters*, Vol. 24, No. 11, pp. 2647–2651, 2020.

6. Seeling, P., M. Reisslein, and B. Kulapala, "Network performance evaluation using frame size and quality traces of single-layer and two-layer video: A tutorial", *IEEE Commun. Surveys Tuts.*, Vol. 6, No. 3, pp. 58–78, Third 2004.

7. "Cisco Annual Internet Report (2018–2023)", *White Paper*, 2020.

8. Clement, J., "Number of monthly logged-in YouTube viewers worldwide as of May 2019", May 2019, `https://www.statista.com/statistics/859829/logged-in-youtube-`

viewers-worldwide/, accessed in September 2020.

9. Watson, A., "Number of Netflix paying streaming subscribers worldwide from 3rd quarter 2011 to 1st quarter 2020", April 2020, https://www.statista.com/statistics/250934/quarterly-number-of-netflix-streaming-subscribers-worldwide/, accessed in September 2020.

10. Klügel, M. and W. Kellerer, "Optimal Mode Selection by Cross-Layer Decomposition in D2D Cellular Networks", *IEEE Transactions on Wireless Communications*, Vol. 19, No. 4, pp. 2528–2542, 2020.

11. Han, L., Y. Zhang, Y. Li, and X. Zhang, "Spectrum-Efficient Transmission Mode Selection for Full-Duplex-Enabled Two-Way D2D Communications", *IEEE Access*, Vol. 8, pp. 115982–115991, 2020.

12. Sambo, Y., M. Shakir, K. Qaraqe, E. Serpedin, M. Imran, and B. Ahmed, "Energy efficiency improvements in HetNets by exploiting device-to-device communications", *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, pp. 151–155, Sept 2014.

13. Amjad, M., M. H. Rehmani, and S. Mao, "Wireless Multimedia Cognitive Radio Networks: A Comprehensive Survey", *IEEE Communications Surveys Tutorials*, Vol. 20, No. 2, pp. 1056–1103, 2018.

14. Mukherjee, A., S. Choudhury, P. Goswami, G. A. Bayessa, and S. K. S. Tyagi, "A novel approach of power allocation for secondary users in cognitive radio networks", *Computers and Electrical Engineering*, Vol. 75, pp. 301 – 308, 2019.

15. Adigun, O., M. Pirmoradian, and C. Politis, *Cognitive Radio for 5G Wireless Networks*, pp. 149–163, John Wiley & Sons, Ltd, 2015.

16. Mokhtarzadeh, H., A. Taherpour, A. Taherpour, and S. Gazor, "Throughput Maximization in Energy Limited Full-Duplex Cognitive Radio Networks", *IEEE*

Transactions on Communications, Vol. 67, No. 8, pp. 5287–5296, 2019.

17. Ren, J., Y. Zhang, N. Zhang, D. Zhang, and X. Shen, "Dynamic Channel Access to Improve Energy Efficiency in Cognitive Radio Sensor Networks", *IEEE Transactions on Wireless Communications*, Vol. 15, No. 5, pp. 3143–3156, 2016.

18. Liu, Y., X. Yang, K. S. Chou, and L. Cuthbert, "Cognitive radio using spectrum-sharing and power minimisation", *2017 IEEE 18th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pp. 1–6, 2017.

19. Kaur, A., S. Sharma, and A. Mishra, "A Novel Jaya-BAT Algorithm Based Power Consumption Minimization in Cognitive Radio Network", *Wireless Personal Communications*, Vol. 108, 05 2019.

20. Agiwal, M., A. Roy, and N. Saxena, "Next Generation 5G Wireless Networks: A Comprehensive Survey", *IEEE Communications Surveys Tutorials*, Vol. 18, No. 3, pp. 1617–1655, 2016.

21. Tullberg, H., P. Popovski, Z. Li, M. A. Uusitalo, A. Hoglund, O. Bulakci, M. Fallgren, and J. F. Monserrat, "The METIS 5G System Concept: Meeting the 5G Requirements", *IEEE Communications Magazine*, Vol. 54, No. 12, pp. 132–139, 2016.

22. "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022", *White Paper*, February 2019.

23. Gao, H., M. Wang, and T. Lv, "Energy Efficiency and Spectrum Efficiency Tradeoff in the D2D-Enabled HetNet", *IEEE Transactions on Vehicular Technology*, Vol. 66, No. 11, pp. 10583–10587, 2017.

24. Gür, G., "Spectrum Sharing and Content-Centric Operation for 5G Hybrid Satellite Networks: Prospects and Challenges for Space-Terrestrial System Integra-

tion", *IEEE Vehicular Technology Magazine*, Vol. 14, No. 4, pp. 38–48, 2019.

25. Zhang, J., X. Zhang, M. A. Imran, B. Evans, Y. Zhang, and W. Wang, "Energy efficient hybrid satellite terrestrial 5G networks with software defined features", *Journal of Communications and Networks*, Vol. 19, No. 2, pp. 147–161, 2017.

26. Zhang, Z., Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G Wireless Networks: Vision, Requirements, Architecture, and Key Technologies", *IEEE Vehicular Technology Magazine*, Vol. 14, No. 3, pp. 28–41, 2019.

27. Yang, P., Y. Xiao, M. Xiao, and S. Li, "6G Wireless Communications: Vision and Potential Techniques", *IEEE Network*, Vol. 33, No. 4, pp. 70–75, 2019.

28. Giordani, M., M. Polese, M. Mezzavilla, S. Rangan, and M. Zorzi, "Toward 6G Networks: Use Cases and Technologies", *IEEE Communications Magazine*, Vol. 58, No. 3, pp. 55–61, 2020.

29. Handley, M., "Using Ground Relays for Low-Latency Wide-Area Routing in Megaconstellations", *Proceedings of the 18th ACM Workshop on Hot Topics in Networks*, HotNets '19, p. 125–132, Association for Computing Machinery, New York, NY, USA, 2019, https://doi.org/10.1145/3365609.3365859.

30. Henry, C., "SpaceX submits paperwork for 30,000 more Starlink satellites", October 2019, https://spacenews.com/spacex-submits-paperwork-for-30000-more-starlink-satellites/, accessed in December 2020.

31. Sheetz, M., "SpaceX is manufacturing 120 Starlink internet satellites per month", August 2020, https://www.cnbc.com/2020/08/10/spacex-starlink-satellte-production-now-120-per-month.html, accessed in December 2020.

32. Henry, C., "Amazon planning 3,236-satellite constellation for internet connectivity", April 2019, https://spacenews.com/amazon-planning-3236-satellite-constellation-for-internet-connectivity/, accessed in December 2020.

33. Xu, L., C. Jiang, Y. Shen, T. Q. S. Quek, Z. Han, and Y. Ren, "Energy Efficient D2D Communications: A Perspective of Mechanism Design", *IEEE Transactions on Wireless Communications*, Vol. 15, No. 11, pp. 7272–7285, Nov 2016.

34. Gür, G., S. Bayhan, and F. Alagöz, "Cognitive femtocell networks: an overlay architecture for localized dynamic spectrum access [Dynamic Spectrum Management]", *IEEE Wireless Communications*, Vol. 17, No. 4, pp. 62–70, 2010.

35. Yao, H., C. Fang, C. Qiu, C. Zhao, and Y. Liu, "A novel energy efficiency algorithm in green mobile networks with cache", *EURASIP J. on Wireless Communications and Networking*, Vol. 2015, No. 1, p. 139, May 2015.

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

37. Li, L., G. Zhao, and R. S. Blum, "A Survey of Caching Techniques in Cellular Networks: Research Issues and Challenges in Content Placement and Delivery Strategies", *IEEE Communications Surveys and Tutorials*, pp. 1–1, 2018.

38. Lee, M. C. and A. F. Molisch, "Individual Preference Aware Caching Policy Design for Energy-Efficient Wireless D2D Communications", *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pp. 1–7, Dec 2017.

39. Xu, Y. and F. Liu, "QoS Provisionings for Device-to-Device Content Delivery in Cellular Networks", *IEEE Transactions on Multimedia*, Vol. 19, No. 11, pp. 2597–2608, Nov 2017.

40. Alagoz, F. and G. Gur, "Energy Efficiency and Satellite Networking: A Holistic Overview", *Proceedings of the IEEE*, Vol. 99, No. 11, pp. 1954–1979, Nov 2011.

41. Brückner, M., P. Drieß, M. Osdoba, and A. Mitschele-Thiel, "A dependency-aware QoS system for mobile satellite communication", *2016 IEEE Wireless Communications and Networking Conference*, pp. 1–6, April 2016.

42. Narottama, B., A. Fahmi, B. Syihabuddin, and A. J. Isa, "Cluster head rotation: a proposed method for energy efficiency in D2D communication", *2015 IEEE International Conference on Communication, Networks and Satellite (COMNESTAT)*, pp. 89–90, Dec. 2015.

43. Wang, Z., H. Shah-Mansouri, and V. W. S. Wong, "How to Download More Data from Neighbors?A Metric for D2D Data Offloading Opportunity", *IEEE Transactions on Mobile Computing*, Vol. 16, No. 6, pp. 1658–1675, 2017.

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

45. Schmidt, J. F., M. K. Atiq, U. Schilcher, and C. Bettstetter, "Encouraging Device-to-Device Communications to Improve Energy Efficiency in Cellular Systems", *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, pp. 1–5, May 2016.

46. Asheralieva, A. and Y. Miyanaga, "QoS-Oriented Mode, Spectrum, and Power Allocation for D2D Communication Underlaying LTE-A Network", *IEEE Transactions on Vehicular Technology*, Vol. 65, No. 12, pp. 9787–9800, Dec 2016.

47. Su, H. and X. Zhang, "Cross-Layer Based Opportunistic MAC Protocols for QoS Provisionings Over Cognitive Radio Wireless Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 26, No. 1, pp. 118–129, Jan 2008.

48. Najimi, M., "Energy-efficient resource allocation in D2D communications for energy harvesting–enabled NOMA-based cellular networks", *International Journal of Communication Systems*, Vol. 33, No. 2, p. e4184, 2020.

49. Panahi, F. H., F. H. Panahi, and T. Ohtsuki, "Energy Efficiency Analysis in Cache-Enabled D2D-Aided Heterogeneous Cellular Networks", *IEEE Access*, Vol. 8, pp. 19540–19554, 2020.

50. Liolis, K., A. Geurtz, R. Sperber, D. Schulz, S. Watts, G. Poziopoulou, B. Evans, N. Wang, O. Vidal, B. Tiomela Jou, M. Fitch, S. Diaz Sendra, P. Sayyad Khodashenas, and N. Chuberre, "Use cases and scenarios of 5G integrated satellite-terrestrial networks for enhanced mobile broadband: The SaT5G approach", *International Journal of Satellite Communications and Networking*, Vol. 37, No. 2, pp. 91–112, 2019.

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

52. Plimon, K., J. Ebert, K. Plimon, W. Gappmair, M. Angelone, and A. Ginesi, "Interference-Dependent Performance of Multi-User Detection in High Throughput Satellite Systems", *2018 11th International Symposium on Communication Systems, Networks Digital Signal Processing (CSNDSP)*, pp. 1–6, July 2018.

53. Zhao, G., S. Chen, L. Qi, L. Zhao, and L. Hanzo, "Mobile-Traffic-Aware Offloading for Energy- and Spectral-Efficient Large-Scale D2D-Enabled Cellular Networks", *IEEE Transactions on Wireless Communications*, Vol. 18, No. 6, pp. 3251–3264, June 2019.

54. Golrezaei, N., A. G. Dimakis, and A. F. Molisch, "Wireless device-to-device communications with distributed caching", *2012 IEEE International Symposium on Information Theory Proceedings*, pp. 2781–2785, July 2012.

55. Hachem, J., N. Karamchandani, S. Moharir, and S. Diggavi, "Caching with partial matching under Zipf demands", *2017 IEEE Information Theory Workshop (ITW)*, pp. 61–65, November 2017.

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

57. Hwang, K. W., D. Applegate, A. Archer, V. Gopalakrishnan, S. Lee, V. Misra, K. K. Ramakrishnan, and D. F. Swayne, "Leveraging Video Viewing Patterns for Optimal Content Placement", Bestak, R., L. Kencl, L. E. Li, J. Widmer, and H. Yin (editors), *NETWORKING 2012*, pp. 44–58, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

58. Fu, B., D. Staehle, G. Kunzmann, E. Steinbach, and W. Kellerer, "QoE-Based SVC Layer Dropping in LTE Networks Using Content-Aware Layer Priorities", *ACM Trans. Multimedia Comput. Commun. Appl.*, Vol. 12, No. 1, pp. 7:1–7:23, August 2015.

59. Chau, P., Y. Lee, T. D. Bui, J. Shin, and J. P. Jeong, "An Efficient Resource Allocation Scheme for Scalable Video Multicast in LTE-Advanced Networks", *2017 11th International Conference on Ubiquitous Information Management and Communication*, pp. 1–8, Jan. 2017.

60. Kim, S., E. Go, Y. Song, H. Cho, M. Rim, and C. G. Kang, "A Study on D2D Caching Systems with Mobile Helpers", *2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN)*, pp. 630–633, July 2018.

61. Xu, C., M. Wang, X. Chen, L. Zhong, and L. A. Grieco, "Optimal Information Centric Caching in 5G Device-to-Device Communications", *IEEE Trans. Mobile Comput.*, Vol. 17, No. 9, pp. 2114–2126, Sept. 2018.

62. Ramzan, N., E. Quacchio, T. Zgaljic, S. Asioli, L. Celetto, E. Izquierdo, and F. Rovati, "Peer-to-peer streaming of scalable video in future Internet applications", *IEEE Communications Magazine*, Vol. 49, No. 3, pp. 128–135, March 2011.

63. Zhan, C. and G. Yao, "SVC-based caching and transmission strategy in wireless device-to-device networks", *2018 16th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pp. 1–6, May 2018.

64. Zhan, C. and Z. Wen, "Content Cache Placement for Scalable Video in Heterogeneous Wireless Network", *IEEE Commun. Lett.*, Vol. 21, No. 12, pp. 2714–2717, Dec. 2017.

65. Hong, D., D. De Vleeschauwer, and F. Baccelli, "A chunk-based caching algorithm for streaming video", *NET-COOP 2010 - 4th Workshop on Network Control and Optimization*, Gent, Belgium, November 2010, session 05 : Streaming applications.

66. Cho, K., M. Lee, K. Park, T. T. Kwon, Y. Choi, and S. Pack, "WAVE: Popularity-based and collaborative in-network caching for content-oriented networks", *2012 Proceedings IEEE INFOCOM Workshops*, pp. 316–321, March 2012.

67. Suksomboon, K., S. Tarnoi, Y. Ji, M. Koibuchi, K. Fukuda, S. Abe, N. Motonori, M. Aoki, S. Urushidani, and S. Yamada, "PopCache: Cache more or less based on content popularity for information-centric networking", *38th Annual IEEE Conference on Local Computer Networks*, pp. 236–243, Oct 2013.

68. Gregori, M., J. Gómez-Vilardebó, J. Matamoros, and D. Gündüz, "Wireless Content Caching for Small Cell and D2D Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 34, No. 5, pp. 1222–1234, May 2016.

69. Chen, B., C. Yang, and Z. Xiong, "Optimal Caching and Scheduling for Cache-

Enabled D2D Communications", *IEEE Communications Letters*, Vol. 21, No. 5, pp. 1155–1158, May 2017.

70. Ji, M., R. Chen, G. Caire, and A. F. Molisch, "Fundamental limits of distributed caching in multihop D2D wireless networks", *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 2950–2954, June 2017.

71. Pedersen, H. A. and S. Dey, "Mobile Device Video Caching to Improve Video QoE and Cellular Network Capacity", *Proceedings of the 17th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, MSWiM '14, pp. 103–107, ACM, New York, NY, USA, 2014.

72. Zhang, K., S. Leng, Y. He, S. Maharjan, and Y. Zhang, "Cooperative Content Caching in 5G Networks with Mobile Edge Computing", *IEEE Wireless Communications*, Vol. 25, No. 3, pp. 80–87, JUNE 2018.

73. Bok, K., J. Kim, and J. Yoo, "Cooperative caching for multimedia data in mobile P2P networks.", *Multimed Tools Appl*, Vol. 78, p. 5193–5216, 2019.

74. Ghandeharizadeh, S. and S. Shayandeh, "Cooperative Caching Techniques for Continuous Media in Wireless Home Networks", *Proc. of the 1st Int. Conference on Ambient Media and Systems*, Ambi-Sys '08, 2008.

75. Wu, D., L. Zhou, Y. Cai, and Y. Qian, "Collaborative Caching and Matching for D2D Content Sharing", *IEEE Wireless Communications*, Vol. 25, No. 3, pp. 43–49, 2018.

76. Liangzhong Yin and Guohong Cao, "Supporting cooperative caching in ad hoc networks", *IEEE Transactions on Mobile Computing*, Vol. 5, No. 1, pp. 77–89, 2006.

77. Ghandeharizadeh, S. and S. Shayandeh, "Domical Cooperative Caching for Streaming Media in Wireless Home Networks", *ACM Trans. Multimedia Com-*

put. Commun. Appl., Vol. 7, No. 4, pp. 40:1–40:17, December 2011.

78. Gabry, F., V. Bioglio, and I. Land, "On Energy-Efficient Edge Caching in Heterogeneous Networks", *IEEE Journal on Selected Areas in Communications*, Vol. 34, No. 12, pp. 3288–3298, 2016.

79. Vu, T. X., L. Lei, S. Chatzinotas, B. Ottersten, and A. V. Trinh, "On the Successful Delivery Probability of Full-Duplex-Enabled Mobile Edge Caching", *IEEE Communications Letters*, Vol. 23, No. 6, pp. 1016–1020, 2019.

80. Chen, B. and C. Yang, "Caching Policy for Cache-Enabled D2D Communications by Learning User Preference", *IEEE Transactions on Communications*, Vol. 66, No. 12, pp. 6586–6601, Dec 2018.

81. Lee, M.-C., M. Ji, A. Molisch, and N. Sastry, "Throughput–Outage Analysis and Evaluation of Cache-Aided D2D Networks With Measured Popularity Distributions", *IEEE Transactions on Wireless Communications*, Vol. PP, pp. 1–1, 08 2019.

82. Li, Z., J. Chen, and Z. Zhang, "Socially Aware Caching in D2D Enabled Fog Radio Access Networks", *IEEE Access*, Vol. 7, pp. 84293–84303, 2019.

83. Wu, Q., G. Y. Li, W. Chen, D. W. K. Ng, and R. Schober, "An Overview of Sustainable Green 5G Networks", *IEEE Wireless Communications*, Vol. 24, No. 4, pp. 72–80, Aug 2017.

84. Yang, S., X. Qiu, H. Xie, J. Guan, Y. Liu, and C. Xu, "GDSoC: Green dynamic self-optimizing content caching in ICN-based 5G network", *Transactions on Emerging Telecommunications Technologies*, Vol. 29, No. 1, p. e3221, 2018.

85. Vu, T. X., S. Chatzinotas, B. Ottersten, and T. Q. Duong, "Energy Minimization for Cache-Assisted Content Delivery Networks With Wireless Backhaul", *IEEE Wireless Communications Letters*, Vol. 7, No. 3, pp. 332–335, June 2018.

86. Lin, S., D. Cheng, G. Zhao, and Z. Chen, "Energy-Efficient Wireless Caching in Device-to-Device Cooperative Networks", *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, pp. 1–5, June 2017.

87. Chen, M., Y. Hao, L. Hu, K. Huang, and V. K. N. Lau, "Green and Mobility-Aware Caching in 5G Networks", *IEEE Transactions on Wireless Communications*, Vol. 16, No. 12, pp. 8347–8361, Dec 2017.

88. Chai, R., Y. Li, and Q. Chen, "Joint Cache Partitioning, Content Placement, and User Association for D2D-Enabled Heterogeneous Cellular Networks", *IEEE Access*, Vol. 7, pp. 56642–56655, 2019.

89. Choi, M., J. Kim, and J. Moon, "Wireless Video Caching and Dynamic Streaming Under Differentiated Quality Requirements", *IEEE Journal on Selected Areas in Communications*, Vol. 36, No. 6, pp. 1245–1257, June 2018.

90. Kawamoto, Y., Z. M. Fadlullah, H. Nishiyama, N. Kato, and M. Toyoshima, "Prospects and challenges of context-aware multimedia content delivery in cooperative satellite and terrestrial networks", *IEEE Communications Magazine*, Vol. 52, No. 6, pp. 55–61, June 2014.

91. Aman, T., T. Yamazato, and M. Katayama, "Traffic prediction scheme for resource assignment of satellite/terrestrial frequency sharing mobile communication system", *2009 International Workshop on Satellite and Space Communications*, pp. 40–44, Sep. 2009.

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

93. Abdelkrim, E. B., M. A. Salahuddin, H. Elbiaze, and R. Glitho, "A Hybrid Regression Model for Video Popularity-Based Cache Replacement in Content Delivery Networks", *2016 IEEE Global Communications Conference (GLOBECOM)*, pp.

1–7, Dec 2016.

94. Feng, B., H. Zhou, H. Zhang, J. Jiang, and S. Yu, "A Popularity-Based Cache Consistency Mechanism for Information-Centric Networking", *2016 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, Dec 2016.

95. Miao, W., G. Min, Y. Wu, H. Wang, and J. Hu, "Performance Modelling and Analysis of Software-Defined Networking Under Bursty Multimedia Traffic", *ACM Trans. Multimedia Comput. Commun. Appl.*, Vol. 12, No. 5s, pp. 77:1–77:19, September 2016.

96. Jiang, T., H. Wang, and 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*, Vol. 30, No. 7, pp. 1215–1224, August 2012.

97. Gür, G. and S. Kafiloğlu, "Layered Content Delivery Over Satellite Integrated Cognitive Radio Networks", *IEEE Wireless Communications Letters*, Vol. 6, No. 3, pp. 390–393, June 2017.

98. Jiang, T., H. Wang, and Y. Zhang, "Modeling Channel Allocation for Multimedia Transmission Over Infrastructure Based Cognitive Radio Networks", *IEEE Systems Journal*, Vol. 5, No. 3, pp. 417–426, Sep. 2011.

99. Vo, N., T. Q. Duong, H. Zepernick, and M. Fiedler, "A Cross-Layer Optimized Scheme and Its Application in Mobile Multimedia Networks With QoS Provision", *IEEE Systems Journal*, Vol. 10, No. 2, pp. 817–830, 2016.

100. Kang, H. J. and C. G. Kang, "Mobile device-to-device (D2D) content delivery networking: A design and optimization framework", *Journal of Communications and Networks*, Vol. 16, No. 5, pp. 568–577, Oct 2014.

101. Liu, C. and B. Natarajan, "Average achievable throughput in D2D underlay net-

works", *2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 118–123, April 2016.

102. Rios, L. M. and N. V. Sahinidis, "Derivative-free optimization: a review of algorithms and comparison of software implementations", *Journal of Global Optimization*, Vol. 56, No. 3, pp. 1247–1293, Jul 2013.

103. Lewis, R. and V. Torczon, "Pattern Search Algorithms for Bound Constrained Minimization", *SIAM Journal on Optimization*, Vol. 9, No. 4, pp. 1082–1099, 1999.

104. Bouttier, E., R. Dhaou, F. Arnal, C. Baudoin, E. Dubois, and A. Beylot, "Analysis of Content Size Based Routing Schemes in Hybrid Satellite / Terrestrial Networks", *2016 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, 2016.

105. ElSawy, H., E. Hossain, and M. Alouini, "Analytical Modeling of Mode Selection and Power Control for Underlay D2D Communication in Cellular Networks", *IEEE Transactions on Communications*, Vol. 62, No. 11, pp. 4147–4161, Nov 2014.

106. Ji, M., G. Caire, and A. F. Molisch, "Fundamental Limits of Caching in Wireless D2D Networks", *IEEE Trans. Inf. Theory*, Vol. 62, No. 2, pp. 849–869, Feb. 2016.

107. Kusaladharma, S. and C. Tellambura, "Performance characterization of spatially random energy harvesting underlay D2D networks with primary user power control", *2017 IEEE International Conference on Communications (ICC)*, pp. 1–7, May 2017.

108. Emara, M., H. ElSawy, S. Sorour, S. Al-Ghadhban, M. Alouini, and T. Y. Al-Naffouri, "Optimal Caching in Multicast 5G Networks with Opportunistic Spectrum Access", *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pp. 1–7, Dec. 2017.

109. Wang, L., S. Bayhan, and J. Kangasharju, "Optimal chunking and partial caching in information-centric networks", *Computer Communications*, Vol. 61, pp. 48 – 57, 2015.

110. Siekkinen, M., E. Masala, and J. K. Nurminen, "Optimized Upload Strategies for Live Scalable Video Transmission from Mobile Devices", *IEEE Trans. Mobile Comput.*, Vol. 16, No. 4, pp. 1059–1072, April 2017.

111. Ohm, J.-R., "Advances in Scalable Video Coding", *Proc. IEEE*, Vol. 93, No. 1, pp. 42–56, Jan. 2005.

112. Reisslein, M., J. Lassetter, S. Ratnam, O. Lotfallah, F. Fitzek, and S. Panchanathan, "Traffic and quality characterization of scalable encoded video: a large-scale trace-based study, part 1: Overview and definitions", *Arizona State Uni. Telecommunications Research Center, Tech. Rep*, 2002.

113. Wu, T., K. D. Schepper, W. V. Leekwijck, and D. D. Vleeschauwer, "Reuse time based caching policy for video streaming", *2012 IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 89–93, Jan. 2012.

114. Ullah, S., T. LeAnh, A. Ndikumana, M. G. R. Alam, and C. S. Hong, "Layered video communication in ICN enabled cellular network with D2D communication", *2017 19th Asia-Pacific Network Operations and Management Symposium (APNOMS)*, pp. 199–204, Sept. 2017.

115. Lim, S., Y. Ko, G. Jung, J. Kim, and M. Jang, "Inter-Chunk Popularity-Based Edge-First Caching in Content-Centric Networking", *IEEE Commun. Lett.*, Vol. 18, No. 8, pp. 1331–1334, August 2014.

116. Yu, J., C. T. Chou, X. Du, and T. Wang, "Internal popularity of streaming video and its implication on caching", *20th International Conference on Advanced Information Networking and Applications - Volume 1 (AINA'06)*, Vol. 1, pp. 6 pp.–40, April 2006.

117. Buzzi, S., C. I, T. E. Klein, H. V. Poor, C. Yang, and A. Zappone, "A Survey of Energy-Efficient Techniques for 5G Networks and Challenges Ahead", *IEEE Journal on Selected Areas in Communications*, Vol. 34, No. 4, pp. 697–709, April 2016.

118. Luo, Y., P. Hong, and R. Su, "Energy-Efficient Scheduling and Power Allocation for Energy Harvesting-Based D2D Communication", *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, pp. 1–6, Dec 2017.

119. Narottama, B., A. Fahmi, and B. Syihabuddin, "Impact of number of devices and data rate variation in clustering method on device-to-device communication", *2015 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)*, pp. 233–238, Aug 2015.

120. Bhadauria, S. and S. Vishwakarma, "Energy Efficient D2D Application for Increasing Battery Usage of Smartphones", *International Journal of Hybrid Information Technology*, Vol. 9, pp. 311–328, 02 2016.

121. Khan, A., L. Sun, E. Ifeachor, J. O. Fajardo, F. Liberal, and H. Koumaras, "Video Quality Prediction Models Based on Video Content Dynamics for H.264 Video over UMTS Networks", *International Journal of Digital Multimedia Broadcasting*, Vol. 2010, 04 2010.

122. "Trace Files and Statistics: H.264/SVC Video Trace Library", `http://trace.eas.asu.edu/h264svc/`, accessed in September 2020.

123. Jin, S. and L. Wang, "Content and Service Replication Strategies in Multi-Hop Wireless Mesh Networks", *Proceedings of the 8th ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM'05)*, p. 79–86, ACM, 2005.

124. Zhang, P., J. Lu, Y. Wang, and Q. Wang, "Cooperative localization in 5G networks: A survey", *ICT Express*, Vol. 3, No. 1, pp. 27 – 32, 2017.

125. Malak, D. and M. Al-Shalash, "Optimal caching for device-to-device content distribution in 5G networks", *2014 IEEE Globecom Workshops (GC Wkshps)*, pp. 863–868, Dec. 2014.

126. Wei, X., P. Ding, L. Zhou, and Y. Qian, "QoE Oriented Chunk Scheduling in P2P-VoD Streaming System", *IEEE Transactions on Vehicular Technology*, Vol. 68, No. 8, pp. 8012–8025, Aug 2019.

127. Gür, G., "Energy-aware cache management at the wireless network edge for information-centric operation", *Journal of Network and Computer Applications*, Vol. 57, pp. 33 – 42, 2015.

128. Pan, X. and T. Zhang, "Comparison and Analysis of Algorithms for the 0/1 Knapsack Problem", *Journal of Physics: Conference Series*, Vol. 1069, p. 012024, IOP Publishing, 2018.

129. Yang, J., M. Ding, G. Mao, and Z. Lin, "Interference Management in In-Band D2D Underlaid Cellular Networks", *IEEE Transactions on Cognitive Communications and Networking*, Vol. 5, No. 4, pp. 873–885, 2019.

130. Sakr, A. H. and E. Hossain, "Cognitive and Energy Harvesting-Based D2D Communication in Cellular Networks: Stochastic Geometry Modeling and Analysis", *IEEE Transactions on Communications*, Vol. 63, No. 5, pp. 1867–1880, 2015.

131. Chen, S., Y. Liang, S. Sun, S. Kang, W. Cheng, and M. Peng, "Vision, Requirements, and Technology Trend of 6G: How to Tackle the Challenges of System Coverage, Capacity, User Data-Rate and Movement Speed", *IEEE Wireless Communications*, Vol. 27, No. 2, pp. 218–228, 2020.

132. Kato, N., B. Mao, F. Tang, Y. Kawamoto, and J. Liu, "Ten Challenges in Advancing Machine Learning Technologies toward 6G", *IEEE Wireless Communications*, Vol. 27, No. 3, pp. 96–103, 2020.

APPENDIX A: COPYRIGHT PERMISSION GRANTS

For [1], as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights

In arXiv for all licenses, with the exception of CC Zero, the original copyright holder retains ownership after posting. In [2], arXiv.org perpetual, non-exclusive license is granted. In that regard, I as the first author and my co-authors have the ownership right for [2].

For [3–5], in reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Boğaziçi University's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

©2019 IEEE. Reprinted, with permission, from S. S. Kafıloğlu, G. Gür and F. Alagöz, "Multidimensional Content Modeling and Caching in D2D Edge Networks," 2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), September 2019.

©2020 IEEE. Reprinted, with permission, from S. S. Kafıloğlu, G. Gür and F. Alagöz, "Cooperative Caching and Video Characteristics in D2D Edge Networks," IEEE Communications Letters, November 2020.