THESIS UNDER JOINT SUPERVISION
UNIVERSITA' MEDITERRANEA DI REGGIO CALABRIA
TAMPERE UNIVERSITY

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE,
DELLE INFRASTRUTTURE E DELL'ENERGIA SOSTENIBILE (DIIES)

PHD IN
INFORMATION ENGINEERING

ciclo XXXV

DOCTORAL PROGRAMME IN DYNAMIC WEARABLE APPLICATIONS
WITH PRIVACY CONSTRAINTS

FACULTY OF INFORMATION TECHNOLOGY AND
COMMUNICATION SCIENCES

TOWARD DYNAMIC SOCIAL-AWARE NETWORKING BEYOND FIFTH GENERATION

CANDIDATE

OLGA CHUKHNO

SUPERVISORS

Prof. Antonio Iera (University of Calabria)

Prof. Antonella Molinaro (University Mediterranea of Reggio Calabria)

Assoc. Prof. Sergey Andreev (Tampere University)

Dr. Olga Galinina (Tampere University)

REGGIO CALABRIA, JANUARY 2023

Toward Dynamic Social-Aware Networking Beyond Fifth Generation

Doctoral thesis submitted by **OLGA CHUKHNO** in order to be eligible for a double doctoral degree awarded by the University Mediterranea of Reggio Calabria and Tampere University

Dipartimento di Ingegneria dell'Informazione, delle Infrastrutture e dell'Energia Sostenibile (DIIES), PhD Course in Information Engineering, XXXV CICLO

Doctoral Programme in Dynamic Wearable Applications with Privacy Constraints Tampere University

AUTHOR:

Olga Chukhno

SUPERVISORS:

Prof. Antonio Iera

Prof. Antonella Molinaro

Assoc. Prof. Sergey Andreev

Dr. Olga Galinina

Reggio Calabria (Italy), January 2023

This dissertation is funded by the European Union's Horizon 2020 Research and Innovation programme under the Marie Skłodowska-Curie grant agreements No. 813278 (A-WEAR: A network for dynamic wearable applications with privacy constraints, `http://www.a-wear.eu/`)

Toward Dynamic Social-Aware Networking Beyond Fifth Generation ©Copyright by Olga Chukhno 2023. This work is licensed under CC BY 4.0.

OLGA CHUKHNO

TOWARD DYNAMIC SOCIAL-AWARE NETWORKING BEYOND FIFTH GENERATION

The Teaching Staff of the PhD course in
INFORMATION ENGINEERING
consists of:

Antonio Iera (coordinator)

Pier Luigi Antonucci,
Giuseppe Araniti,
Francesco Buccafurri,
Claudia Campolo,
Giuseppe Coppola,
Mariantonia Cotronei,
Lorenzo Crocco,
Dominique Dallet,
Claudio De Capua,
Francesco Della Corte,
Giuliana Faggio,
Gioia Failla,
Fabio Filianoti,
Patrizia Frontera,
Sofia Giuffrè,
Giorgio Graditi,
Voicu Groza,
Tommaso Isernia,
Gianluca Lax,
Aime Lay Ekuakille,
Gaetano Licitra,
Antonella Molinaro,
Andrea Morabito,
Carlo Francesco Morabito,
Giacomo Morabito,
Rosario Morello,
Fortunato Pezzimenti,
Filippo Pratico',
Domenico Rosaci,
Giuseppe Ruggeri,
Mariateresa Russo,
Antonino Vitetta

To my family

Acknowledgments

This study was carried out under the financial support of the A-WEAR project funded by the European Union's Horizon 2020 Research and Innovation programme under the Marie Skłodowska-Curie grant agreements No. 813278.

I would like to express my thanks to my supervisors, *Prof. Antonio Iera*, *Prof. Antonella Molinaro*, *Assoc. Prof. Sergey Andreev*, and *Dr. Olga Galinina*, for their invaluable guidance, support, and mentorship. Their expertise, knowledge, and experience have been instrumental in helping me navigate the complexities of my research, and their feedback and suggestions have been extremely valuable in shaping the outcome. I am deeply appreciative to *Assist. Prof. Sara Pizzi Assoc. Prof. Claudia Campolo*, and *Assoc. Prof. Giuseppe Araniti*, for the time and energy they invested in providing me with the resources and support I needed to succeed. Their patience and encouragement have been a source of inspiration and motivation. I also could not have undertaken this journey without my defense committee, who generously provided knowledge and expertise.

I would like to extend my gratitude to my colleagues, *Simona Lohan*, *Aleksandr Ometov*, *Dmitri Moltchanov*, *Yuliya Gaidamaka*, *Konstantin Samouylov*, *Yevgeni Koucheryavy*, *Jari Nurmi*, *Viktoriia Shubina*, *Justyna Skibińska*, *Pavel Pascacio*, *Laura Flueratoru*, *Darwin Quezada Gaibor*, *Asad Ali*, *Asma Channa*, *Ekaterina Svertoka*, *Waleed Bin Qaim*, *Raúl Casanova-Marqués*, *Sylvia Holcer*, *Joaquín Torres-Sospedra*, *Sven Casteleyn*, *Radim Burget*, *Jiri Hosek*, who I had the pleasure of collaborating with during my doctoral study. I am honored to have had the opportunity to collaborate with such a talented and dedicated group of people.

I would also like to express my sincere gratitude to my colleagues and friends, *Giuseppe Ruggeri*, *Marica Amadeo*, *Vincenzo Violi*, *Tomas Bravenec*, *Domenico Zappalà*, *Angelo Tropeano*, *Gianluca Brancati*, *Gurtaj Singh (Giorgio)*, *Giacomo Genovese*, *Pasquale Scopelliti*, *Giuseppe Marrara*, *Roman Klus*, and girls: *Chiara Suraci*, *Federica Rinaldi*, *Alessia Ferraro*, *Sabrina Zumbo*, *Giada Battaglia*, *Lucie Klus*, *Salwa Saafi*, *Olga Vikhrova*, *Daria Alekseeva*, and *Anna Gaydamaka*, for their support and contributions to my life. Their unwavering support, encouragement, and companionship have made this journey not just professional but also personal. Thank you for being a constant source of inspiration to me.

Last but not least, I am grateful for my closest friends and their families *(they are all in my heart)*. They have been there to listen to me when I needed someone to talk to, celebrate my successes, and pick me up when I was down. I am so grateful for the unwavering support and love that they have shown me. I am truly blessed to have them in my life, and I could not imagine going through life's journey without them all by my side. I want to take a moment to express my deep gratitude to my family for their unwavering support and love. I want to thank my parents, *Natalia Chukhno* and *Viktor Chukhno*, for always being there for me and teaching me the values of hard work and perseverance. I would also like to thank my sisters, *Nadia Chukhno* and *Marina Borisova* (and my lovely nephew *Egor*), for their love, support, and understanding. I am grateful for the bond we share. I want to take a moment to express my gratitude to my beloved pets, my source of unconditional love.

Olga Chukhno. January 25, 2023, Reggio Calabria, Italy.

Abstract

The rise of the intelligent information world presents significant challenges for the telecommunication industry in meeting the service-level requirements of future applications and incorporating societal and behavioral awareness into the Internet of Things (IoT) objects. Social Digital Twins (SDTs), or Digital Twins augmented with social capabilities, have the potential to revolutionize digital transformation and meet the connectivity, computing, and storage needs of IoT devices in dynamic Fifth-Generation (5G) and Beyond Fifth-Generation (B5G) networks.

This research focuses on enabling dynamic social-aware B5G networking. The main contributions of this work include *(i)* the design of a reference architecture for the orchestration of SDTs at the network edge to accelerate the service discovery procedure across the Social Internet of Things (SIoT); *(ii)* a methodology to evaluate the highly dynamic system performance considering jointly communication and computing resources; *(iii)* a set of practical conclusions and outcomes helpful in designing future digital twin-enabled B5G networks.

Specifically, we propose an orchestration for SDTs and an SIoT-Edge framework aligned with the Multi-access Edge Computing (MEC) architecture ratified by the European Telecommunications Standards Institute (ETSI). We formulate the optimal placement of SDTs as a Quadratic Assignment Problem (QAP) and propose a graph-based approximation scheme considering the different types of IoT devices, their social features, mobility patterns, and the limited computing resources of edge servers. We also study the appropriate intervals for re-optimizing the SDT deployment at the network edge. The results demonstrate that accounting for social features in SDT placement offers considerable improvements in the SIoT browsing procedure. Moreover, recent advancements in wireless communications, edge computing, and intelligent device technologies are expected to promote the growth of SIoT with pervasive sensing and computing capabilities, ensuring seamless connections among SIoT objects.

We then offer a performance evaluation methodology for eXtended Reality (XR) services in edge-assisted wireless networks and propose fluid approximations to characterize the XR content evolution. The approach captures the time and space dynamics of the content distribution process during its transient phase, including time-varying loads, which are affected by arrival, transition, and departure processes. We examine the effects of XR user mobility on both communication and computing patterns. The results demonstrate that communication and computing planes are the key barriers to meeting the requirement for real-time transmissions. Furthermore, due to the trend toward immersive, interactive, and contextualized experiences, new use cases affect user mobility patterns and, therefore, system performance.

Index terms: Beyond Fifth-Generation, Social Internet of Things, Digital Twining, Wireless Networks, Edge Computing.

Sommario

L'emergere del nuovo mondo dell'informazione intelligente crea sfide senza precedenti per l'industria delle telecomunicazioni per soddisfare i requisiti di servizio stringenti delle applicazioni future e gestire l'esigenza di incorporare una nuova consapevolezza sociale e comportamentale negli oggetti della Internet of Things (IoT). I Social Digital Twins (SDTs), ovvero Digital Twins potenziati con capacità sociali, sembrano essere il fattore chiave per consentire la trasformazione digitale e soddisfare i requisiti di connettività, elaborazione e archiviazione dei dispositivi IoT in reti di quinta generazione (5G) e evoluzioni future (B5G).

Questo lavoro di ricerca è finalizzato alla progettazione di soluzioni di networking per il supporto dei SDT nel contesto beyond 5G. I principali contributi della tesi includono: *(i)* progettazione di un'architettura di riferimento e di soluzioni per l'orchestrazione dei SDT alla periferia (edge) della rete finalizzate ad accelerare la procedura di scoperta dei servizi attraverso il paradigma di Social Internet of Things (SIoT); *(ii)* un insieme di metodologie per valutare le prestazioni delle soluzioni progettate che tengano conto sia delle risorse di comunicazione che di calcolo; *(iii)* una serie di conclusioni pratiche e risultati utili per la progettazione della futura evoluzione della rete 5G basata sull'uso esteso del concetto di digital twin.

Nello specifico, si propone un'orchestrazione per i SDTs e un framework SIoT-Edge allineato con l'architettura MEC standardizzata dall'European Telecommunications Standards Institute (ETSI). Si formula il posizionamento ottimale di SDTs come QAP e si propone uno schema di approssimazione basato su grafi considerando i diversi tipi di dispositivi IoT, le loro caratteristiche sociali, i modelli di mobilità e le limitate risorse di calcolo dei server perimetrali. Inoltre, si esplorano gli intervalli di ri-ottimizzazione per la distribuzione dei SDT all'edge della rete. I risultati dimostrano che tenere conto delle caratteristiche sociali nel posizionamento SDT possa offrire notevoli miglioramenti nella procedura di navigazione SIoT. Inoltre, i recenti progressi nelle comunicazioni wireless, nell'edge computing e nelle tecnologie dei dispositivi intelligenti promuoveranno la crescita di SIoT con capacità di rilevamento e calcolo pervasive, garantendo connessioni senza soluzione di continuità tra gli oggetti SIoT.

Si offre una metodologia di valutazione delle prestazioni per i servizi XR nelle reti wireless edge assistite e si propone un'approssimazione fluida per caratterizzare l'evoluzione del contenuto XR. L'approccio cattura le dinamiche temporali e spaziali del processo di distribuzione dei contenuti nella sua fase transitoria, che include carichi variabili nel tempo, cioè carichi che sono una funzione del tempo e dipendono dai processi di arrivo, transizione e partenza. Si indaga gli impatti della mobilità degli utenti XR dal punto di vista della comunicazione e dell'informatica. I risultati dimostrano che sia i piani di comunicazione che

quelli informatici sono le principali barriere per soddisfare il requisito della trasmissione in tempo reale dei servizi XR. Inoltre, a causa della tendenza verso esperienze immersive, interattive e contestualizzate, i nuovi casi d'uso influenzano la mobilità degli utenti e, quindi, le prestazioni del sistema.

Parole chiave: Reti di quinta generazione ed evoluzioni future, Social Internet of Things, Digital Twining, Reti Wireless, Edge Computing.

Contents

List of Figures

List of Tables

Abbreviations

3D	Three Dimensional
3DoF	Three Degrees of Freedom
3GPP	Third Generation Partnership Project
5G	Fifth-Generation
B5G	Beyond Fifth-Generation
6DoF	Six Degrees of Freedom
AI	Artificial Intelligence
AR	Augmented Reality
AP	Access Point
BIM	Building Information Modeling
BS	Base Station
CEP	Closest Edge Placement
C-LOR	Co-Location Object Relationship
CPU	Central Processing Unit
CSI	Channel State Information
C-WOR	Co-Work Object Relationship
D2D	Device-to-Device
DL	Deep Learning
DRL	Deep Reinforcement Learning
DT	Digital Twin
E2E	End-to-End
EHF	Extremely High Frequency
eSoCEP	enhanced Social-aware Closest Edge Placement
ETSI	European Telecommunications Standards Institute
HMD	Head-Mounted Display
I/O	Input/Output
IoT	Internet of Things
KPI	Key Performance Indicators
LB	Local Branching
LoS	Line-of-Sight

MAC Medium Access Control
ME Mobile Edge
MEC Multi-access Edge Computing
ML Machine Learning
MR Mixed Reality
mmWave Millimeter Wave
NFV Network Function Virtualization
NLoS Non-Line-of-Sight
NR New Radio
ODE Ordinary Differential Equation
OOR Ownership Object Relationship
POR Parental Object Relationship
QAP Quadratic Assignment Problem
RAM Random Access Memory
RAN Radio Access Network
RAT Radio Access Technology
RINS Relaxation Induced Neighborhood Search
RTT Round-Trip Time
RSSI Received Signal Strength Indicator
SDN Software-Defined Networking
SDT Social Digital Twin
SINR Signal-to-Interference-plus-Noise Ratio
SIoT Social Internet of Things
SOR Social Object Relationship
SoCEP Social-aware Closest Edge Placement
SVE Social Virtual Entity
SWIM Small World In Motion
THz Terahertz
UMi Urban Microcell
VE Virtual Entity
VM Virtual Machine
VR Virtual Reality
XR eXtended Reality

1

Introduction

This Chapter introduces the motivation of this research and its aims, defines central research objectives, and summarizes the contributions. The Chapter also covers the structure of the thesis.

1.1 Motivation

The Internet of Things (IoT) refers to a network of connected devices that exchange data over the Internet. One of the earliest examples of the IoT was a Coca-Cola machine at Carnegie Mellon University [6, 7]. The IoT had then evolved into a system operating with numerous technologies, i.e., the Internet, wireless communication, micro-electromechanical, and embedded systems [8]. As smartphones became an essential communication device, they also became a part of the IoT. In October 2021, vehicles also joined the IoT when May Mobility initiated a pilot program to test its autonomous driving software.

To date, smart environments employ a wide range of IoT devices resulting in a significant increase in mobile data traffic and the global rollout of the Fifth-Generation (5G) system. This global rollout of the 5G system opens up significant opportunities for IoT applications that share data with devices with minimal processing and storage capabilities. According to the most recent study by Strategy Analytics, the number of internet-connected devices is projected to reach 50 billion by the end of 2030, while 59% of IoT data processing will occur at the edge by 2025 [9].

The advancement of the IoT has led to the integration of social networking principles, known as the Social Internet of Things (SIoT) paradigm, to connect individuals and devices on a global scale and facilitate the exchange of information between heterogeneous IoT objects, taking into account both physical and social behavior and providing several benefits. First, the SIoT facilitates service discovery utilizing social network techniques. Second, it enables the interchange of information linked with or created by devices on a social basis, including novel group-based communication methods such as *sociocast* [10]. Then, it guarantees scalability via social collaboration between nodes. An additional benefit is the ability to establish social connections between devices that use different technologies, which facili-

tates interoperability across various IoT platforms. Finally, SIoT may exploit the degree of interaction between objects to offer secure connections between devices that have a social relationship.

As the number of IoT devices and social connections continues to grow, digital twinning becomes increasingly essential in processes such as service discovery. The introduction of so-called Social Digital Twins (SDTs) at the edge [1,2], utilized to expose the resource/services on behalf of the physical objects and keep track of the social relationships dynamically established among their physical replicas, may facilitate their discovery by traversing the social network graph. In this scenario, placement of SDTs at the edge might be adopted to make social network browsing faster and more efficient. To the best of our knowledge, there is no work so far addressing this problem, mainly due to challenges posed by the social capabilities of the IoT devices. Since most IoT devices are resource-constrained, creating and managing social interactions would further complicate their design. Hence, an efficient operation of an IoT network, which is even less likely when the social component is considered, is only possible with the adoption of edge computing [11].

In the future, 5G and Beyond Fifth-Generation (B5G) applications and services will likely require high-speed connectivity to meet demands for sufficient bandwidth and data rates, dynamic mobility, resilience, and other requirements [12]. There will be a need to provide connectivity with extremely low end-to-end latency and high reliability, for example, for immersive services [13, 14]. However, some services can already meet ever-increasing expectations, thereby drawing the contours of new use cases. With the rapid development of mobile and display technologies, the use of eXtended Reality (XR) has brought revolutionary improvements. Still, full XR adoption remains hindered for various reasons, e.g., hardware and source device computing costs, low resolution and visual quality, and usability of devices due to mobility [15].

The emergence of problems related to user mobility in IoT is linked to usage-specific movements [5] (one may observe different mobility patterns of users involved in message writing and audio recording while walking) [5]. Even though most studies consider the specifics inherent to the conventional mobile phone application user, primarily due to the unavailability of affordable head-worn devices that provide a satisfactory level of user experience, we are observing a paradigm shift from smartphones to wearable devices [16]. Head-worn devices are becoming integral to future systems and potentially may lead to the implementation of B5G use cases. In this regard, a significant concern is understanding the user context based on behavior analysis and the real-time response in the on-premises environment, i.e., seamless network connections, bandwidth availability, data transfer and application execution requirements, and data migration management. As applications impact movement, there is a need for a profound transformation in communication and computing patterns.

1.2 Objectives and Methodology

In the following, the **research objectives** (ROx) are formulated to address the identified challenges and research gaps to support *Dynamic Social-Aware Networking Beyond 5G*. We divide the challenges and research gaps into two sets when considering dynamic *(i)* social- and *(ii)* behavior-aware interactions in edge-aided IoT.

The proliferation of always-connected SIoT systems, devices, and sensors for digital automation in 5G and B5G networks demands low latency and high reliability and resilience. To meet these requirements, new orchestration policies and practical methodologies are needed for the optimal deployment of SDTs at the network edge to ensure timely and efficient service discovery. Past research does not cover the placement of Digital Twins (DTs), augmented with a social dimension, at the network edge. Moreover, unpredictable SIoT device movement might cause a non-negligible deviation from the optimal deployment of SDTs at the network edge, requiring their re-deployment. However, current literature only considers constant re-deployment intervals. This might lead to the following situations: *(i)* the system's time-averaged behavior may deviate from the optimum due to unexpected SIoT device mobility; *(ii)* the system's time-averaged behavior may trigger irrational network resource usage, e.g., when the devices are static, i.e., during nighttime.

To bridge these gaps and accelerate service discovery procedure across SIoT, we:

- **RO1. Propose a framework for SDT placement at the network edge.**
- **RO2. Propose orchestration policies for SDTs at the network edge under mobility conditions.**
- **RO3. Explore SIoT networks from the perspective of IoT object motion and propose orchestration policies for SDTs at the network edge under heterogeneous mobility conditions.**

The second set of research challenges and gaps centers around behavior-aware interactions in edge-aided IoT when considering ubiquitous contextualized experiences. The technology evolves along with ever-increasing proliferation, interest in, and demand for new applications and services, e.g., emerging XR services that submerge users into a virtual universe. This unlocks XR's freedom of mobility and interaction and triggers a new stage of XR technology adoption, bringing new challenging and technical problems to be addressed, such as dynamicity in content distribution evolution.

Current literature on XR systems primarily focuses on their steady-state operation. However, due to XR interaction freedom, state-of-the-art solutions may not be efficient in real-world applications. The challenge is to develop methods that can account for the dynamic and non-stationary nature of immersive reality interactions and provide network planners with a means to evaluate system performance. Furthermore, immersive interactions can distract users from the physical world, alter their behavior and motion, and thereby impact the operation of communication networks.

To bridge these gaps, we:

- **RO4. Propose a practical methodology for evaluating the performance of dynamic XR systems, with a focus on the computing plane.**
- **RO5. Propose a practical methodology for evaluating the performance of highly dynamic XR systems with periodic arrival rates, with a focus on the communication and computing plane.**
- **RO6. Explore XR communication networks from the perspective of user interaction patterns, highlight the entailed challenges, and propose an assessment methodology for XR services.**

To address the challenges and research gaps formulated above, a general research approach was adopted in this work. It consists of several stages defined as follows.

- **Study state-of-the-art technologies and research methods.** This stage aims at acquiring knowledge of the system of interest necessary to develop the frameworks and their subsequent cross-validation.
- **Use case analysis and problem statement.** Focusing on the challenging use cases and solutions available in the literature, different research gaps are identified at this stage and translated into a subset of problems.
- **Solution design**. Indicated challenges and problems are investigated to formulate solutions for further testing and verification.
- **Model definition**. Each identified problem is addressed in the model design that provides relevant and illustrative Key Performance Indicators (KPI) for assessing and validating the proposed solutions. The solutions are evaluated and validated via advanced simulators.
- **Evaluation.** At this stage, the obtained results are evaluated and discussed to highlight the root cause of the addressed problem.

1.3 Contributions

1.3.1 Research Outputs

The detailed contributions and relevant publications are organized in 6 blocks (3 per each set of research challenges and gaps), each corresponding to a research objective from the list presented in Section 1.2. Contributions C1, C2, and C3 focus on dynamic social-aware interactions in edge-aided IoT and are included in Chapter 2, while Chapter 3 contains contributions C4, C5, and C6, focusing on dynamic behavior-aware interactions in edge-aided IoT.

C1. SDT placement framework.

The detailed contributions are summarized as follows:

- The design of a framework for the static Social-aware Closest Edge Placement (SoCEP) of DTs;

- The formulation of the SDT placement by accounting for the limited computing resources of edge servers, social relationships among IoT devices, and constraints on the latency in the connectivity between a physical device and the corresponding DT and in the inter-DT connectivity;
- The evaluation of the performance of the SoCEP against a baseline solution under different settings in terms of storage constraints on edge servers and latency demands.

These contributions have been included in the publication:

***Chukhno, O.**, Chukhno, N., Araniti, G., Campolo, C., Iera, A. and Molinaro, A., 2020. Optimal placement of social digital twins in edge IoT networks. Sensors, 20(21), p.6181.*

C2. Orchestration for SDTs under mobility.

The detailed contributions are summarized as follows:

- The design of an SIoT-edge framework, named enhanced Social-aware Closest Edge Placement (eSoCEP) ratified as a functionality of the European Telecommunications Standards Institute (ETSI) Mobile Edge (ME) orchestrator;
- The formulation of the optimal placement of SDTs by accounting for different types of IoT devices, their social features, mobility patterns, and the limited computing resources of edge servers;
- The design of an approximation scheme to find near-optimal solutions and the application of approximation techniques;
- The evaluation of the performance of the proposed algorithm against the formulated optimal solution and benchmark schemes;
- The analysis of the time-dependent behavior of the SIoT-edge system under conditions of device mobility and the time interval duration selection.

These contributions have been included in the publication:

***Chukhno, O.**, Chukhno, N., Araniti, G., Campolo, C., Iera, A. and Molinaro, A., 2022. Placement of Social Digital Twins at the Edge for Beyond 5G IoT Networks. IEEE Internet of Things Journal (Early Access).*

C3. Assessment methodology and orchestration for SDTs under mobility.

The detailed contributions are summarized as follows:

- The review of device motion patterns that might depend on the time, the device type, space, and the scenario;
- The design of social-aware orchestration comprising Network Function Virtualization (NFV), Software-Defined Networking (SDN), edge/fog and cloud computing, Deep Learning (DL)-based user activity prediction, and sensing and tracking technologies;
- The evaluation of the performance of the proposed orchestration;
- The analysis of the re-optimization time interval concerning the impact on the service discovery latency for traditional system design and co-design of localization, sensing, and Artificial Intelligence (AI)-driven communication and computation.

These contributions have been included in the publication:

***Chukhno, O.**, Chukhno, N., Araniti, G., Campolo, C., Iera, A. and Molinaro, A., 2023. Social-Aware Orchestration in 5G+/6G-IoT Ecosystems. (Ready for submission).*

C4. XR System performance evaluation methodology.

The detailed contributions are summarized as follows:

- The practical methodology to characterize XR content evolution in dynamic networks as continuous fluids considering the computing plane;
- The evaluation of the performance of the proposed methodology based on a fluid approximation;
- The assessment of the XR system characteristics under different network settings;
- The practical conclusions for designing XR networks considering the computing plane.

These contributions have been included in the publication:

***Chukhno, O.**, Galinina, O., Andreev, S., Molinaro, A. and Iera, A., 2023. Content Distribution Dynamics of Edge-Aided Immersive Reality Services. (Ready for submission).*

C5. Joint communication and computing performance evaluation methodology for XR system.

The detailed contributions are summarized as follows:

- The practical methodology to evaluate joint communication and computing system performance with periodic arrival rates;
- The characterization of the content evolution by capturing the time and space dynamics of the content distribution process;
- The validation of the proposed methodology performance based on a fluid approximation through Monte-Carlo simulations;
- The system performance assessment under different network settings;
- The practical conclusions for designing XR wireless communication networks considering joint communication and compute with periodic arrival processes.

These contributions have been included in the publication:

***Chukhno, O.**, Galinina, O., Andreev, S., Molinaro, A. and Iera, A., 2023. Content Distribution Dynamics of Edge-Aided Immersive Reality Services. (Ready for submission).*

C6. Joint user behavior, communication, and computing assessment methodology for XR system.

The detailed contributions are summarized as follows:

- The analysis of user behavior patterns that confirms use case-dependent changes in gait characteristics, such as direction, velocity, stride length, step width, and stance time;

- The sources of evidence of the user motion impact on the network operation;
- The case study for mobile XR that characterizes the system performance with respect to user motion, communication, and computing;
- The results confirm the uniqueness of XR applications in terms of user behavior patterns, which demands for the development of innovative application-centric algorithms, protocols, and mechanisms to support a high-performance connection in response to stringent XR requirements.

These contributions have been included in the publication:

***Chukhno, O.**, Galinina, O., Andreev, S., Molinaro, A. and Iera, A., 2022. Interplay of User Behavior, Communication, and Computing in Immersive Reality 6G Applications. IEEE Communications Magazine, 60(12), 28-34.*

1.3.2 Publications

The list of the author's publications produced during the Ph.D. period includes 5 articles related to the subject of the thesis and mentioned in Chapter 2 and Chapter 3.

Articles related to the subject of the thesis are:

1. ***Chukhno, O.**, Chukhno, N., Araniti, G., Campolo, C., Iera, A. and Molinaro, A., 2020. Optimal placement of social digital twins in edge IoT networks. Sensors, 20(21), p.6181.*
2. ***Chukhno, O.**, Chukhno, N., Araniti, G., Campolo, C., Iera, A. and Molinaro, A., 2022. Placement of Social Digital Twins at the Edge for Beyond 5G IoT Networks. IEEE Internet of Things Journal (Early Access).*
3. ***Chukhno, O.**, Galinina, O., Andreev, S., Molinaro, A. and Iera, A., 2022. Interplay of User Behavior, Communication, and Computing in Immersive Reality 6G Applications. IEEE Communications Magazin, 60(12), 28-34.*
4. ***Chukhno, O.**, Chukhno, N., Araniti, G., Campolo, C., Iera, A. and Molinaro, A., 2023. Social-Aware Orchestration in 5G+/6G-IoT Ecosystems. (Ready for submission).*
5. ***Chukhno, O.**, Galinina, O., Andreev, S., Molinaro, A. and Iera, A., 2023. Content Distribution Dynamics of Edge-Aided Immersive Reality Services. (Ready for submission).*

1.4 Thesis Outline

The thesis is organized into 4 Chapters, their content is briefly described below.

- **Chapter 1** contains the motivation, objectives, contributions, and structure of this work.
- **Chapter 2** focuses on dynamic social-aware interactions in edge-aided IoT.
- **Chapter 3** focuses on dynamic behavior-aware interactions in edge-aided IoT.
- **Chapter 4** includes the summary of research outcomes and the discussion of future research.

The final part of the thesis includes the bibliography.

2

Dynamic Social-Aware Interactions in Edge-Aided IoT

This chapter focuses on dynamic social-aware interactions in edge-aided IoT.

In Section 2.1, we propose an SDT placement framework. We address the optimal DT placement problem at the network edge to reduce the latency between physical devices and the corresponding DTs and among DTs of friend devices to ensure efficient data exchange and accelerate the service discovery across the SIoT network. Specifically, we present a framework for the SoCEP of DTs based on edge computing and SIoT for a static IoT object deployment scenario.

In Section 2.2, we propose an orchestration for SDTs. We present the design of an SIoT-edge framework that is aligned with the Multi-access Edge Computing (MEC) architecture standardized by the ETSI. We formulate the optimal placement of SDTs considering the different types of IoT devices, their social features, mobility patterns, and the limited computing resources of edge servers. We propose an approximation scheme for finding near-optimal solutions.

In Section 2.3, we present an assessment methodology and propose orchestration policies. We investigate time-, device-, space-, and scenario-depending network re-optimization intervals and their enablers for the deployment of SDTs at the network edge. We offer a review of device motion patterns that might depend on the time (e.g., morning, day, evening, or night hours), the device type (e.g., static sensors, cars, wearable devices, etc.), space (i.e., environment), and the scenario (e.g., critical situation or everyday routine actions). We offer social-aware orchestration comprising NFV, SDN, edge/fog and cloud computing, deep learning-based user activity prediction, and sensing/tracking technologies.

2.1 Social Digital Twin Placement Framework

Smart environments comprise numerous IoT objects, including personal and measuring devices, sensors, and actuators. Such a remarkable rise in connected devices has led to a significant increase in mobile data traffic [17], contributing to the advent of the 5G and B5G systems. This creates tremendous development prospects for IoT applications leveraging the opportunity of exchanging data with every item while requiring minimal storage and computing resources. According to the recent forecast in [18], the number of IoT connections will reach 24 billion by the end of 2025.

The SIoT paradigm [19, 20] has gained ground in recent literature as a valuable tool to integrate humans and machines into one global social network. The integration of social networking concepts with the IoT provides several advantages. First, SIoT facilitates effective service discovery based on typical social network techniques. Second, it enables the exchange on a social basis of information associated with/generated by devices [10, 21]. Third, it guarantees scalability through social collaboration among nodes. Then, SIoT allows establishing social relationships between objects that use different technologies; this allows device interoperability across different IoT platforms. Finally, SIoT can leverage the degree of interaction between objects to guarantee trusted connections between friend devices [22].

However, IoT devices are typically resource-constrained; creating and managing social relationships would provide additional design challenges. Widespread usage of IoT, even more, when the social dimension is considered, is unlikely without adopting the concept of cloud computing. Indeed, the cloud may augment resource-constrained IoT devices by providing extra computation, storage, and context-awareness capabilities. In particular, the DT [23, 24] concept has gained momentum as a concrete means to implement such a vision by bridging the physical world with the digital one. Physical objects are augmented with a digital footprint hosted in the remote cloud and, thus, enabled to perceive the environment better and understand their role in the context in which they are immersed.

Unlike the existing literature [25–27], we align the DT concept with the edge computing paradigm [16], among the most influential technical developments that dominate the IoT industry in 2020 [28]. Edge computing [29–31] aims to localize processing resources closer to end-devices rather than a centralized cloud computing environment. Since data do not traverse over a network to the cloud to be processed, the network load and the data latency are significantly reduced.

As the further step related to the state-of-the-art development [32, 33], we consider DTs of physical objects augmented with social capabilities, i.e., SDTs. The concept of Social Virtual Entity (SVE) can be traced back to early research, such as in in [34, 35], where the notion of SVE had been preliminarily introduced. However, hosting applications and SDTs presents several issues, particularly at the edge. Some of these challenges are inherited from the literature on deploying virtualized applications at the edge [36–38]. Other challenges are specifically unique to the SIoT environment. First, the available edge servers colocated with

Access Points (APs) or/and Base Stations (BSs) have unequal distribution of the limited network edge resources. Next, it may be necessary to ensure low-latency communication between the physical and virtual counterparts to achieve the DT objectives. This is especially true when considering interactive applications. Finally, SDTs may also need to be interconnected given the quick navigation of the social network of IoT devices whenever the resources or capabilities of friend devices must be discovered and/or chained.

In such a context, we aim to offer the following contributions:

- we propose a placement framework for SDTs, SoCEP;
- we formulate the placement problem by accounting for the limited edge server computing resources, social relationships among IoT devices, and latency constraints;
- we apply a linear relaxation for the formulated optimization problem;
- we assess the performance of the proposed framework under different settings.

The rest of this section is structured as follows. The motivation behind the investigated topic and an overview of background materials are presented in Subsection 2.1.1. The proposed SIoT-edge framework is outlined in Subsection 2.1.2, whereas the problem formulation for SDT placement is characterized in Subsection 2.1.3. In Subsection 2.1.4, simulation experiments are presented. Discussions are drawn in Subsection 2.1.5.

2.1.1 Motivation

As most IoT objects are resource-hungry, the gap between required and locally accessible resources increases. In this context, cloud computing is a valuable solution to the problem. However, specific IoT applications require short response times (e.g., automation control in an intelligent factory [39]), while others generate massive amounts of data that must be analyzed (e.g., Augmented Reality (AR), Virtual Reality (VR), Mixed Reality (MR), and XR [40]). In contrast, some may demand security guarantees (e.g., surveillance in a public place [41]). Cloud computing cannot meet these requirements and cannot support these types of IoT applications. Edge computing can address the challenges mentioned above by hosting storage and processing resources and applications closer to end users [42].

In recent years, there has been considerable interest in edge computing for IoT [43, 44], also fueled by the ETSI activities, which refer to such a paradigm as MEC [31, 45]. A large body of literature addresses computation offloading [46,47] in view of allowing IoT devices to speed up data processing, thus reducing energy consumption. In [48], the problem of selecting the appropriate offloading routes for IoT services has been addressed. The optimization of the response time of IoT applications has been done in [49,50], where computation offloading methods improve latency experienced by users.

Further, edge computing offers new opportunities in the field of IoT that would not be possible by leveraging traditional cloud-based systems. For instance, DTs, which have already been shifted from concept to reality [33], can be effectively implemented at the edge. The idea of DTs was first introduced in [51] and later formalized in [25], where the main elements

of the DT concept are identified, namely, a real space (physical objects), a virtual space (virtual objects), and the link for the data flow between real and virtual domains. Virtual objects deliver the semantic description of the related physical objects and their resources and capabilities, which are abstracted into attributes. This abstraction allows performing an effective search of the capabilities/resources needed for creating and composing IoT services at the application layer [34].

The virtualization layer has evolved into a vital element of many reference IoT platforms [52, 53] and commercial implementations [34], and the interest in it has been further sparked by the emerging DTs concept [54]. However, the communication processes between a physical object and its virtual counterpart and among different DTs are still open issues that attract growing interest [33]. Another challenge is the placement of such virtual counterparts at the edge in a way to accommodate the distributed and limited nature of computing and storage resources of edge servers.

A growing body of literature has investigated the possibilities of edge networks to satisfactorily meet the latency constraints on pairing a physical device and its DT [35, 55]. A cost-aware cloudlet placement strategy accounting for the cost of deploying edge servers and the end-to-end latency between physical objects and their avatars is proposed in [38]. In [56], the placement problem is considered as a generalized assignment problem to reduce latency. Further related works focus on the virtual machine replica placement problem [57, 58], the service entity problem [37,59], and the joint service placement [60,61] with additional focuses on, e.g., request routing/scheduling [36, 62].

In the past years, the concept of using elements of social networks in IoT has attracted unprecedented attention from the research community [63–67]. The synergy of social networking and IoT paradigms can offer several benefits and allow the devices to create relationships autonomously. The "social network of intelligent objects" paradigm has been proposed in [19]. SIoT aims to simplify the navigability of a network of billions of devices and enhance their trustworthiness. Social relationships can be created, for instance, between objects belonging to the same owner, between fixed devices located in the same place, between objects carried by people who frequently meet, and between objects of the same model, vendor, and production batch. SIoT supports many novel applications and services for the IoT in a more robust and productive [68] way by facilitating the interaction between physical objects through the digital world. In particular, the exploitation of social network principles in the IoT domain has proven to foster resource visibility, enhance device and service discovery, and enable practical object reputation assessment, service composition, and source crowding [69, 70].

Researchers have followed various approaches to design an SIoT architecture and implement a platform for constructing the SIoT service environment and the virtualization of applications [68, 71]. In [72], a technique for the implementation of a Virtual Entity (VE) (the virtual equivalent of the physical object [34]) is addressed. In [73], a cloud-based social IoT solution, wherein each physical device has a virtual counterpart, is developed. The platform has four major features, i.e., social agents, Platform as a Service model, reusability, and

cloud storage. An analogous approach is presented in [74], where virtual objects of physical devices hosted at the edge are enabled to browse the social network of devices.

Indeed, IoT applications may be developed to use data and resources provided by friend devices. Some applications, for instance, may need to push (query) data to (from) categories of friends [35]. Additionally, a composite service may be constructed through the chain of resources (e.g., cached data) provided by the SDTs of friend devices. In these particular circumstances endowed by social flavor, the discovery of resources/capabilities of friend devices over the social network may be promoted by the existence of a social counterpart at the edge, which exposes them on behalf of the physical objects. Therefore, applications that may need to fast browse the social network by visiting SDTs of friend devices could benefit from finding SDTs of friends either in the same or nearest edge servers. This would lead to latency reduction and a decrease in the amount of data traversing the edge infrastructure.

Table 2.1: DT placement solutions.

Ref.	**Social features**	**Optimization function**	**Device-DT delay constraint**	**Edge capacity constraints**	**Device heterogeneity**	**Edge heterogeneity**	**Solution**
[38]	No	Sum of cloudlet cost and device-DT latency	No	Yes	No	Yes	Lagrangian heuristic algorithm
[75]	No	Sum of DT initialization (device-DT delay) and synchronization delay	Yes	Yes	Yes	Yes	Deep Reinforcement Learning (DRL)-based algorithm
[37]	Twitter social graph	Sum of activation, placement, proximity (device-DT delay), and colocation cost	No	No	Yes	Yes	Iterative solution of a series of minimum graph cuts
[76]	No	Sum of computing and communication delay (device-DT latency)	No	No	Yes	Yes	Distributed approximation scheme
[77]	No	Sum of access, switching and communication (device-DT) delay	No	Yes	Yes	Yes	Iteration-based algorithm

The problem of service placement, both in general and in the specific case of DTs, has been discussed in the scientific literature, and numerous solutions have appeared. Table 2.1 summarizes the main features of the closest works that address the (social) digital twin placement problem. Past literature does not cover the complete convergence of virtualiza-

tion and socialization capabilities of future IoT devices and applications by accounting for heterogeneous and resource-limited edge computing environments.

To fill the identified gap, we offer a framework for the *social-aware placement* of DTs at the network edge. We address the placement problem by taking into account the common proximity-driven approach for pairing physical objects and the corresponding virtual counterparts [38]. We target to minimize the latency experienced between edge servers hosting DTs, for which the corresponding physical devices have a social relationship. This is intended to guarantee that SIoT devices can quickly communicate and discover services querying their social relationship network on the virtualization layer hosting DTs.

2.1.2 System Model

This subsection introduces the reference system model and summarizes our modeling assumptions. Table 2.2 gathers the basic notations used throughout this Chapter.

Table 2.2: Notations related to Chapter 2.

Notation	**Description**
N	IoT devices
M	Edge servers
$G_P = (V_P, E_P)$	Graph of IoT devices
p_{ij}	Data exchange intensiveness between IoT devices $i, j \in V_P$
Γ	Edge server capacity
$aCPU_k$	Central Processing Unit (CPU) capability of edge server $k \in V_S$
aD_k	Disk capability of edge server $k \in V_S$
$aRAM_k$	Memory capability of edge server $k \in V_S$
CPU_i	CPU requirement to execute the SDT of physical IoT device $i \in V_P$
D_i	Disk requirement to execute the SDT of physical IoT device $i \in V_P$
RAM_i	Memory requirement to execute the SDT of physical IoT device $i \in V_P$
$G_S(V_S, E_S)$	Weighted undirected graph of edge servers
L_{ik}	Latency between device $i \in V_P$ and its SDT placed at edge server $k \in V_S$
L_{kl}	Latency between edge servers $k, l \in V_S$
d_{ik}	Physical distance between IoT device $i \in V_P$ and edge server $k \in V_S$
d_{kl}	Physical distance between SDTs deployed at edge servers $k, l \in V_S$
x_{ik}	Binary variable taking the value 1 if SDT of device $i \in V_P$ is mapped to edge server $k \in V_S$
$L_{\max_i}$	Maximum latency between physical device $i \in V_P$ and its SDT deployed at at edge server $k \in V_S$
THR_{CPU}	CPU utilization threshold value
THR_D	Disk storage utilization threshold value
THR_{RAM}	Random Access Memory (RAM) utilization threshold value

We assume the reference layered IoT architecture [52, 53]. The bottom layer accommodates physical IoT devices that belong to the real world, such as wearables, smartphones,

sensors, and actuators, etc.). Here, 5G infrastructure provides the options to connect these IoT devices.

The virtualization layer represents the top layer, wherein the digital replicas of physical objects, i.e., the SDTs, are deployed. Similar to DTs, SDTs augment the physical devices with computing capabilities and storage. Specifically, it allows caching and pre-filtering/aggregation of raw data transmitted by the associated IoT object before feeding IoT applications that process them. Alongside the semantic description of the associated physical object, the SDT stores information about all the social connections and "friends" created by the associated physical object [19]. More precisely, the SDT holds metadata describing the type of friend devices and the SIoT relationship type for each friend device.

An IoT device willing to query friend devices to discover services and/or push data to them read information about friendship stored in the SDT. Once such information has been collected, the SDT can interact with its peers on behalf of the physical IoT object. Then, SDTs of all (a subset of) friend devices can be contacted one by one, according to what we refer to in this thesis as *friend browsing*. It is assumed that SDTs are deployed as virtualized applications, for instance, as containers [78, 79], and instantiated at the network edge (i.e., in edge servers).

We assume that N IoT devices are located within the coverage area of M wireless access points (e.g., BSs, APs). At a given time instant, IoT device i is considered to be connected to a single BS/APs, namely, to the closest one [80].

The corresponding devices build relationships according to the SIoT paradigm. The resultant social network is represented by a social-based graph $G_P = (V_P, E_P)$. The set of vertices in the graph G_P, i.e., V_P, corresponds to the IoT objects connected by links in set E_P. The probability p_{ij}, reflects the existence of the social connections between IoT devices i and j.

We assume an edge infrastructure composed of M edge servers associated with each wireless BS/AP [80]. Since SDTs can store data and perform some processing, they have CPU and storage requirements that must be considered when deploying them on an edge server, which typically has a finite amount of resources. We define the parameter Γ, $\Gamma \geq 1$, limiting the number of SDTs hosted on an edge server.

The edge network is represented by graph $G_S = (V_S, E_S)$, where V_S is a finite set of edge servers, whereas E_S is a set of links between the edge servers. We consider that the number of IoT objects is higher than the number of edge servers, $|V_P| = N > |V_S| = M$, which does not limit the generality of the presentation.

The latency between each pair of edge servers k, l is given by L_{kl}. By analogy with [37,38], L_{kl} is estimated to be proportional to the distance between edge servers k and l. We denote as L_{ik} the latency between device i and its SDT located at edge server k. For L_{ik}, we disregard the delay over the radio interface and consider the latency between a BS that covers the IoT device and an edge server that hosts the corresponding SDT [49]. Therefore, L_{ik} and L_{kl} are estimated as in [81]:

$$L_{ik} = \epsilon d_{ik}, \tag{2.1}$$

$$L_{kl} = \epsilon d_{kl}, \tag{2.2}$$

where ϵ is the distance to latency mapping coefficient, d_{ik} is the physical distance between the BS that serves device i and edge server k, whereas d_{kl} is the distance between edge servers k and l.

2.1.3 Mathematical Characterization

We aim to place SDTs at edge servers so that a cost function, which is expressed as a combination of latency of connections between physical objects and their virtual counterparts and latency between SDTs of physical objects linked by social relationships, is minimized while meeting a set of constraints on communication latency and capabilities of edge servers (see Fig. 2.1, where the SDTs of two friend IoT devices, i and j, are placed at edge server k and l, respectively).

To model the problem of SDTs placement on the given set of edge servers, we introduce a decision variable, $x_{ik} \in \{0, 1\}$, which indicates whether the SDTs of device i is assigned to edge server k. The binary variable x_{ik} is equal to 1, if the SDTs of device i is deployed at edge server k, and $x_{ik} = 0$ otherwise. The variable p_{ij} is represented as:

$$p_{ij} = \begin{cases} 1, & \text{if device } i \text{ has a social relationship with device } j \text{ according to SIoT} \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

The problem can be formulated as follows:

$$\min_{x} \sum_{i \in V_P} \sum_{k \in V_S} x_{ik} L_{ik} + \sum_{i \in V_P} \sum_{k \in V_S} \sum_{j \in V_P} \sum_{l \in V_S} x_{ik} x_{jl} p_{ij} L_{kl}, \tag{2.4}$$

subject to

$$\sum_{k \in V_S} x_{ik} = 1, \ \forall \ i \in V_P, \tag{2.5}$$

$$\sum_{i \in V_P} x_{ik} \leq \Gamma, \ \forall \ k \in V_S, \tag{2.6}$$

$$L_{ik} \leq L_{\max}, \ \forall \ i \in V_P, \ \forall \ k \in V_S, \tag{2.7}$$

$$x_{ik}, x_{jl} \in \{0, 1\}, \ \forall \ i, j \in V_P, \ \forall \ k, l \in V_S. \tag{2.8}$$

Constraint (2.5) holds the condition that the SDT of device $i \in V_P$ can be assigned to one edge server only. Constraint (2.6) means that the maximum number of SDTs that can be deployed at an edge server $k \in V_S$ is limited by Γ. Constraint (2.7) includes a limitation on the latency between the physical device and the edge server hosting the corresponding SDT, which is upper bounded by $L_{\max}$. Constraint (2.8) reminds that we conveniently model the placement problem through binary variables.

Lemma 1. SDT placement problem is NP-hard.

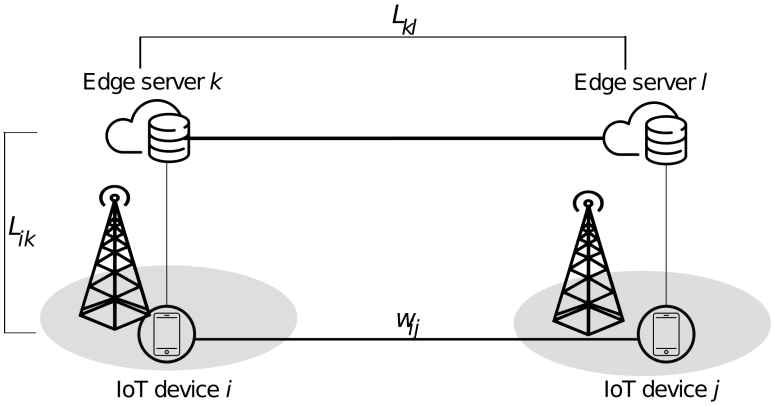

Fig. 2.1: Cost function [1].

Proof. We prove that the SDT placement problem is NP-hard by reducing it from the Quadratic Assignment Problem (QAP), well known to be NP-hard [82].

Quadratic Assignment Problem: QAP handles the problem of assigning a set of facilities to a set of locations, with the cost determined by the distance and flow between the facilities and the cost of installing a facility at a location. The optimization goal is to allocate each facility to a location while lowering the overall cost. Assuming that n is the number of facilities and locations, $N = \{1, 2, ..., n\}$, the facilities' placement is given by the bijection $N \to N$, i.e., a facility can be assigned to one location, and a location can accommodate one facility only.

Social-aware Closest Edge Placement problem: SoCEP is the relaxed form of QAP, in which the objectivity and surjectivity requirements for mapping the set of facilities (SDTs) to the set of placement sites (edge servers) are omitted. We consider the problem of allocating SDTs with the cost being a function of the latency L_{kl} between edge servers and the existence of social links p_{ij} between the IoT objects (see Fig. 2.1) and the placement cost associated with SDTs of IoT objects being deployed at edge servers. In our model, N SDTs are assigned to M edge servers such that SDTs corresponding to friend IoT devices are deployed closer with the maximum possible proximity between SDTs and their IoT object while allowing some flexibility for selecting/not selecting edge servers. More precisely, compared to QAP, p_{ij} can be correlated with the flow between a couple of facilities, L_{kl} can be associated with the distance between a couple of locations, and L_{ik} can be closely related to the cost of placing facilities at locations.

This is a formulation of the quadratic assignment problem, which is NP-hard.

Due to the quadratic form, one may infer that there is nonlinearity in the cost function in (2.4). To remove the nonlinearity, we perform the linearization of the objective function by introducing a new binary variable, y_{ikjl}, that equals 1 if SDTs of devices i and j are deployed at edge servers k and l, correspondingly [83, 84], i.e.,

$$y_{ikjl} = \begin{cases} 1, & \text{if } x_{ik} = 1 \text{ and } x_{jl} = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{2.9}$$

We also denote the cost contributions related to the latency between the SDTs of IoT objects i and j deployed at edge server k and l, correspondingly, as C_{ikjl}, by substituting $p_{ij}L_{kl}$.

Then, the cost function is given by:

$$\min_{x,y} \sum_{i\in V_P} \sum_{k\in V_S} x_{ik}L_{ik} + \sum_{i\in V_P} \sum_{k\in V_S} \sum_{j\in V_P} \sum_{l\in V_S} y_{ikjl}C_{ikjl}, \tag{2.10}$$

by satisfying the following constraints:

$$\sum_{k\in V_S} x_{ik} = 1, \ \forall\ i \in V_P, \tag{2.11}$$

$$\sum_{i\in V_P} x_{ik} \leq \Gamma, \ \forall\ k \in V_S, \tag{2.12}$$

$$\sum_{l\in V_S} y_{ikjl} = x_{ik}, \ \forall\ k \in V_S, \ \forall\ i,j \in V_P, \tag{2.13}$$

$$\sum_{k\in V_S} y_{ikjl} = x_{jl}, \ \forall\ l \in V_S, \ \forall\ i,j \in V_P, \tag{2.14}$$

$$L_{ik} \leq L_{\max}, \ \forall\ i \in V_P, \ \forall\ k \in V_S, \tag{2.15}$$

$$x_{ik}, x_{jl}, y_{ikjl} \in \{0,1\}, \ \forall\ i,j \in V_P, \ \forall\ k,l \in V_S. \tag{2.16}$$

Constraints (2.13) and (2.14) make it possible to account for mutual social relationships between two SDTs. SDTs of friend devices i and j may only be allocated at the respective edge servers k and l if the associated binary variables x_{ik} and x_{jl} are equal to 1. Constraint (2.15) restricts the latency between the physical device and the edge server hosting the matching SDT and is applied between the physical device and the edge server. Constraint (2.16) ensures that x_{ik}, x_{jl}, y_{ikjl} are binary variables.

2.1.4 Performance Evaluation

In this subsection, we evaluate the performance of the proposed framework. We start by describing the simulation environment, which utilizes the input parameters collected in Table 2.3. We then test our model through computer simulations by using the IBM ILOG CPLEX Optimization Studio 12.10.0 software suite [85, 86] on an Intel(R) Xeon(R) CPU E5 − 2620 v4 at 2.10 GHz with 19.7 GB RAM.

Simulation Settings

The simulation scenario, as shown in Fig. 2.2 (black squares represent the IoT objects in the considered area, whereas the links reflect social connections between a couple of IoT objects), corresponds to the city center of Santander, Spain. The settings we employed, in terms of area of interest and object's metadata, are detailed in [87]. We limit our analysis to a region

Table 2.3: System parameters related to Section 2.1.

Parameter	**Value**
Number of IoT devices, N	150
Number of edge servers, M	9
Capacity of an edge server, Γ	var
Maximum latency between a device and an edge server, $L_{\max}$	var
Distance to latency mapping coefficient, ϵ	3.33 ms/km [81]

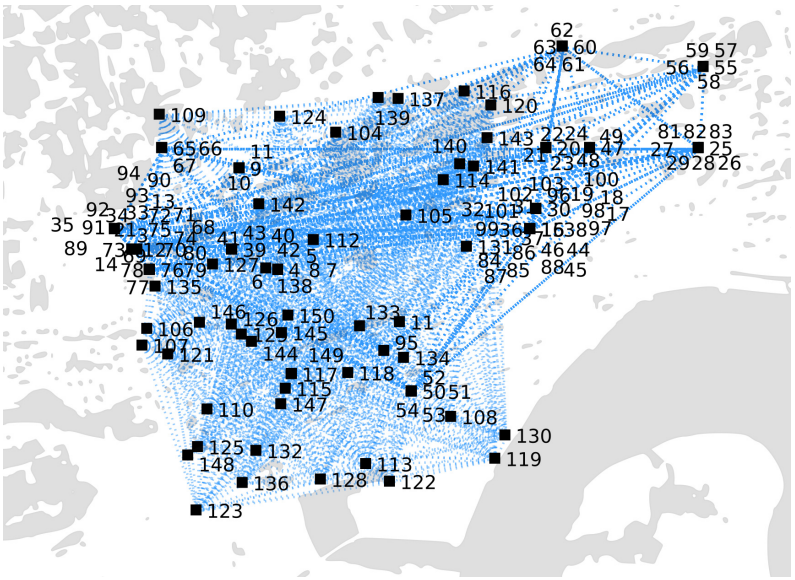

Fig. 2.2: Scenario of interest [1].

of 4 km $\times$ 4 km (including uninhabited zones). We assume the deployment of $M = 9$ base stations. Each BS is deployed at the center of the area and has a coverage of 1 km^2. An edge server is paired with each BS. There are $N = 150$ static IoT devices inside the area (which corresponds to public devices from the SIoT dataset (http://www.social-iot.org)).

From the dataset provided in [87], we extract information about physical IoT devices (IDs, types, coordinates, and adjacency matrix with the social relationships). In this case, value p_{ij} indicates the presence of a social relationship between a couple of devices (2.3). In this simulation, the set of social relationships is represented by $R \in \{\text{POR}\}$ [19], where Parental Object Relationship (POR) denotes the relationship among objects belonging to the same manufacturing batch [19]. The average number of social relationships per IoT device is 80.

Similar to [37, 38], the latency L_{ik} between device i and edge server k, as well as the latency L_{kl} between edge serves k and l are estimated to be proportional to the distance between them. More precisely, $L_{ik} = \epsilon d_{ik}$ and $L_{kl} = \epsilon d_{kl}$, where ϵ is the distance to latency mapping coefficient, d_{ik} and d_{kl} are physical distances between device i and edge server k and between edge servers k, l, respectively.

In addition, the following assumptions apply to our study: *(i)* when the edge server k that hosts the SDT is colocated to the same BS which its corresponding device, i, is connected to, then $d_{ik} = 0$, $L_{ik} = 0$; and *(ii)* when two SDTs are placed at the same edge server (i.e., $d_{kl} = 0, \forall k = l$), then $L_{kl} = 0$. The first assumption is commonly seen in the literature [49] and permits omitting the delay experienced over the radio interface. The second one disregards intra-edge server latency.

We evaluate the performance of the proposed model against a baseline approach, i.e., Closest Edge Placement (CEP), according to which digital counterparts of IoT objects are deployed at the closest edge server, ignoring social components. Therefore, the CEP optimization problem can be written as

$$\min_{x} \sum_{i \in V_P} \sum_{k \in V_S} x_{ik} L_{ik}, \tag{2.17}$$

while satisfying the same constraints as per the SoCEP optimization problem.

The following metrics are determined for assessment purposes:

- *Average latency among SDTs of friend IoT devices* is the average latency between each pair of edge servers hosting SDTs whose corresponding physical devices are friends.
- *Average latency for friends browsing* contains the latency between an IoT object and the edge server hosting the associated SDT and the latency for browsing, one-by-one, all friend devices' SDTs hosted in the edge infrastructure.

Performance Analysis

In this subsection, we assess the performance of the proposed SDT placement framework, SoCEP, against the benchmark approach, CEP, when varying the maximum latency, $L_{\max}$,

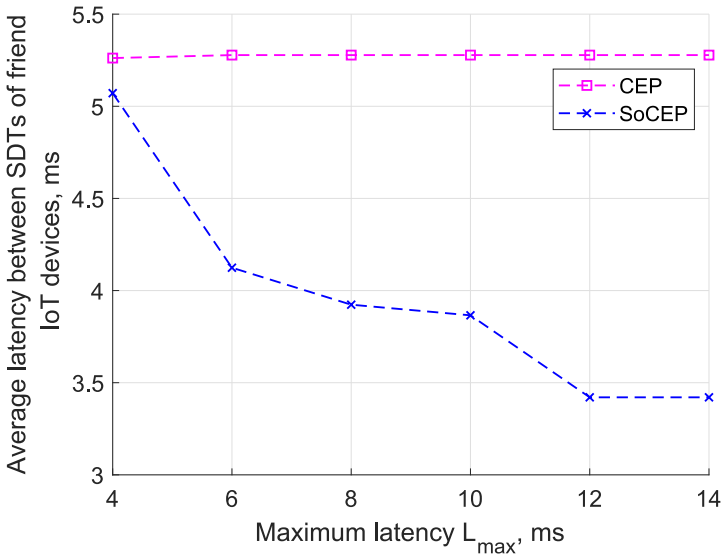

Fig. 2.3: Latency among SDTs of friend IoT devices, $\Gamma = 20$ [1].

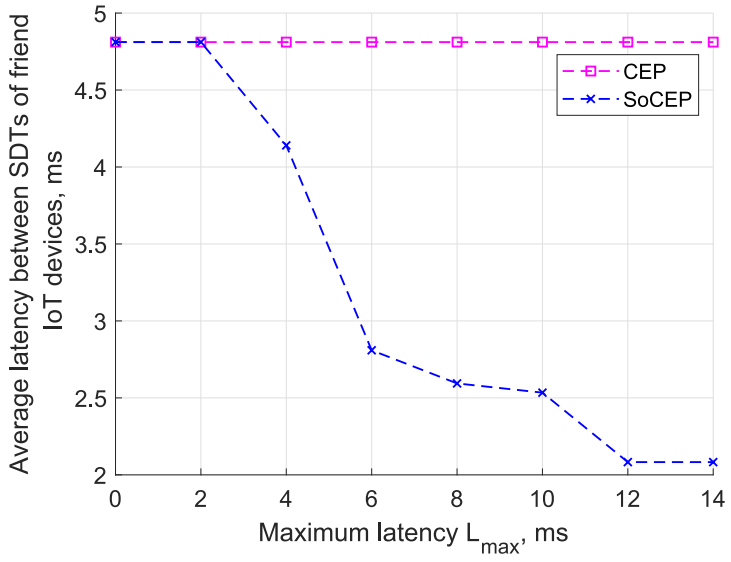

Fig. 2.4: Latency among SDTs of friend IoT devices, $\Gamma = 40$ [1].

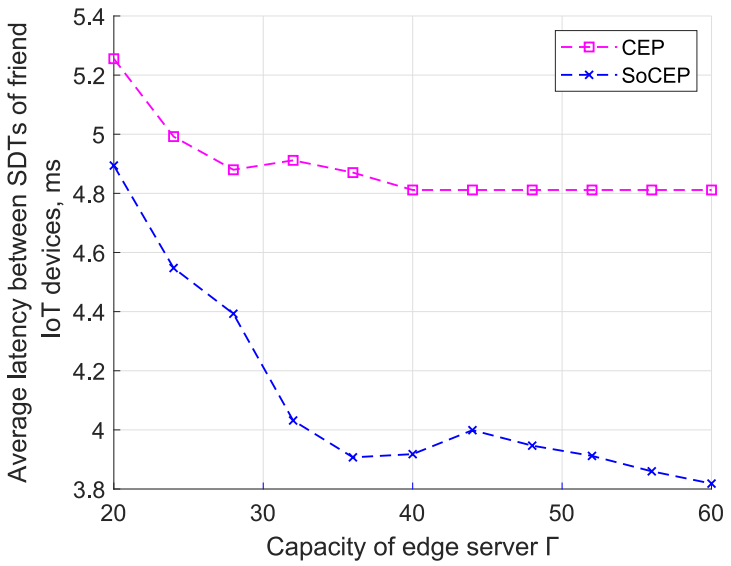

Fig. 2.5: Latency among SDTs of friend IoT devices, $L_{\max} = 5$ ms [1].

in the range $0 - 14$ ms and the edge server's capacity, Γ, in the range $20 - 60$. One can see that the proposed in this Chapter solution highly outperforms the benchmark solution with gains up to 35.1% and 56.7% for $L_{\max} = 14$ ms (as shown in Fig. 2.3 and Fig. 2.4, correspondingly) and up to 20.6% and 55.7% for $\Gamma = 60$ (as shown, respectively, in Fig. 2.5 and Fig. 2.6).

In Fig. 2.7 and Fig. 2.8, we examine the proposal's effectiveness in providing low latency in browsing the friends by the SDTs. The results indicate that the proposal is faster than

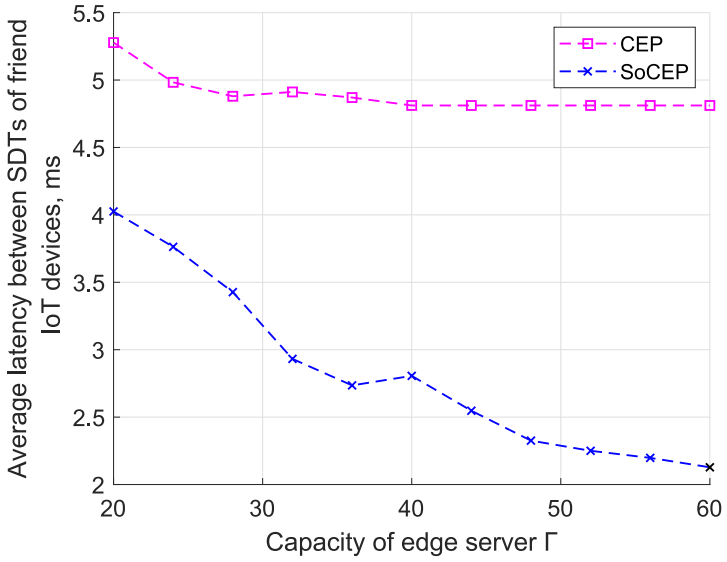

Fig. 2.6: Latency among SDTs of friend IoT devices, $L_{\max} = 7$ ms [1].

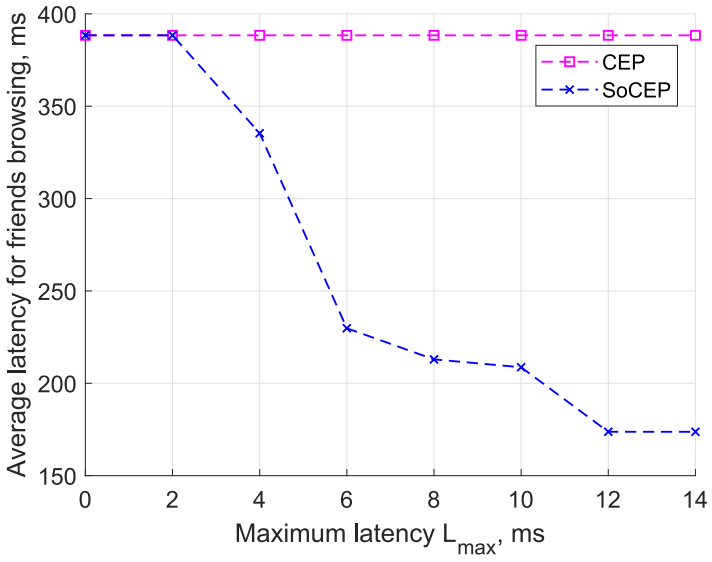

Fig. 2.7: Latency for friend browsing, $\Gamma = 40$ [1].

the CEP browsing approach. When the maximum latency $L_{\max}$ reaches 10 ms and $\Gamma = 40$, delay values are more than halved. This observation is essential since it demonstrates that accounting for social features in SDT placement enables one to enhance the browsing process considerably.

We solve the problem optimally by using a branch-and-bound algorithm, a variation of the exhaustive search approach that considers an acceptable solution set. In general, the computing complexity of such accurate methods is exponential. The problem difficulty

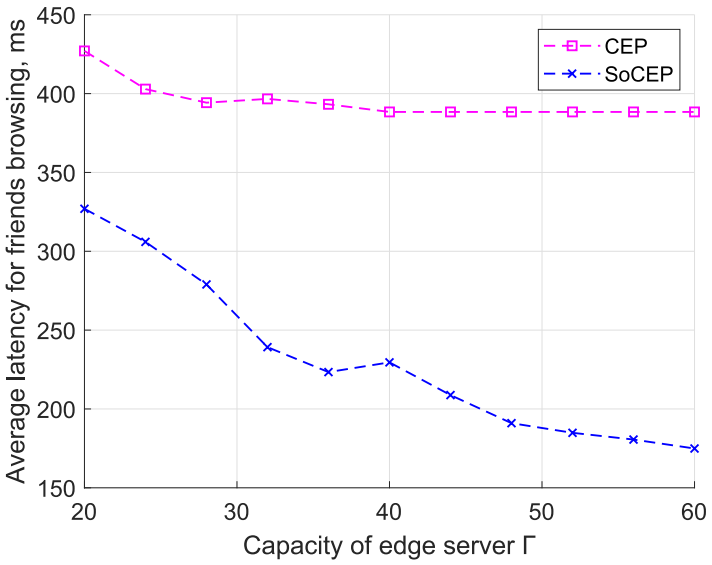

Fig. 2.8: Latency for friend browsing, $L_{\max} = 7$ ms [1].

increases in proportion to the problem size. Furthermore, we point out that the problem size and constraints impact the computational complexity. From Table 2.4, one may deduce that the relaxation on the maximum latency, $L_{\max}$, slows down the process of finding optimal solutions. This trend can be explained by the rise in the number of acceptable placement alternatives for hosting SDTs at the edge.

For less stringent latency limitations, the provided computation time values become acceptable. As a result, the next step (see Section 2.2) is to construct an efficient (in terms of computing time) heuristic algorithm that can address the problem in large networks while focusing on the SoCEP objectives with minimal compromise in solution optimality.

Table 2.4: Computation time.

$L_{\max}$	3 ms	3.5 ms	4 ms	4.5 ms	5 ms	5.5 ms	6 ms	6.5 ms	>7 ms
Time	1.6 s	3.7 s	5.8 s	7.1 s	46.8 s	1534.4 s	2601.5 s	2718.4 s	4380.9 s

2.1.5 Discussions

In this section, we presented a solution for DT placement at the edge, which accounts for proximity requirements and relationships among IoT devices established according to the SIoT paradigm. We have formulated the proposal as an optimization problem and obtained promising results. Indeed, the proposal can contribute to making the SIoT more viable by reducing browsing procedure times.

some devices may require the quicker engagement of certain SDTs compared to others. As a result, the type and number of social links in the social graph may be used to construct

the cost function by weighing the latency contributions between SDTs of friend devices. Furthermore, more intelligent placement solutions that take into account network deployment features, IoT device types, and mobility patterns are of particular importance to our research contributions. In the next section (Section 2.2), we address these issues by offering a placement framework and orchestration policies for the dynamic SDTs.

2.2 Social Digital Twin Orchestration Under Mobility

By 2050, there will be 24 billion linked devices [88], indicating that almost every thing around us, such as wearable gadgets, mobile phones, robots, electric meters, automobiles, and streetlights, will be connected to the Internet. Such devices will allow end-users to experience a broad range of novel applications, such as AR/XR and autonomous assisted navigation [89], which require high throughput, low latency, high reliability, and ubiquitous availability.

Despite the fact that 5G networks are being deployed in several countries, the requirements of most of the aforementioned applications are still poorly met, pushing the researchers to focus on B5G solutions by 2030. DTs appear to be the game-changer needed to enable the digital transformation and meet the constantly growing connectivity, computing, and storage demands of massively deployed heterogeneous IoT devices in 5G and beyond network use cases [90].

Connectivity between physical and virtual counterparts, i.e., DTs, is an open issue attracting considerable interest, especially in guaranteeing real-time data transfer [91]. To this aim, there is a broad consensus on placing DTs at the network edge to ensure low-latency interactions with their physical counterparts located in proximity [35,38]. However, the decision about the placement of DTs has to account for the limited and heterogeneous resources at edge servers. Such a decision becomes even more complicated when considering mobile devices in the physical realm that constantly trigger DT migrations among edge servers to ensure proximity to the physical devices. Recent works have addressed these issues; for instance, DRL and an algorithm based on iteratively solving a series of minimum graph cuts have been leveraged, respectively, in [75] and in [37], to approximate the optimal placement solution.

Another feature, which adds constraints to the placement policy, is the possibility for physical devices to establish mutual social relationships, e.g., according to the SIoT paradigm, which has gained tremendous popularity in recent years in the IoT research arena [10, 19]. A social network of devices is created by establishing and maintaining different types of Relationships, such as Co-Location, Ownership, and Parental, among others [19]. Indeed, IoT applications can be conceived to leverage data and services provided by "friend" devices. Their discovery may be facilitated with the presence of SDTs at the edge, used to expose the mentioned resource/services on behalf of the physical devices and keep track of the dynamic social relationships among their physical counterparts. Consequently, placement of SDTs might also be implemented to make social network browsing quicker and more effective.

In light of the aforementioned, it is evident that the dynamic placement of DTs with social features at the edge is a hard decision to be taken when both user-centric and operator-centric demands need to be simultaneously met. In Section 2.1, the issue has been addressed by formulating an initial optimization problem under basic conditions and without considering SDTs mobility, as the main objective was to provide a proof-of-concept of the introduced

paradigm. In this section, we offer extended research that provides the contributions summarized below:

- we offer an SIoT-edge framework, wherein the proposed placement of SDT strategy, named eSoCEP, is a functionality of the ETSI ME orchestrator;
- we formulate the optimal placement of SDTs as a QAP that extends the preliminary formulation in Section 2.1.1 by taking into account different types of IoT devices, their social capabilities, mobility patterns, and the limited computing resources of edge servers;
- we propose an approximation approach to find near-optimal solutions;
- we evaluate the performance of proposals against benchmark solutions;
- we analyze the SIoT-edge system under device mobility settings and define the selection of the time interval duration between consecutive runs of the SDT deployment policy.

The rest of this section is organized as follows. In Subsection 2.2.1, the SIoT-edge framework is introduced. In Subsection 2.2.2, the optimization problem is formulated, whereas in Subsection 2.2.3, an approximation algorithm and relaxation techniques for the SDT placement are described. Simulation results are reported in Subsection 2.2.4. The main findings and conclusions of the study are summarized in Subsection 2.2.5.

2.2.1 Orchestration

This subsection presents a general overview of the proposed SIoT-edge framework for the dynamic placement of the SDTs. The reference architecture consists of a *real-world layer* and a *virtualization layer* (see Fig. 2.9).

The real-world layer represents the physical world that accommodates IoT objects interconnected through facilities. Social relationships among objects are assumed and set up according to the SIoT paradigm [19]. The virtualization layer is responsible for hosting the SDTs, digital representations of real devices [92]. They provide the distinctive capabilities of a digital counterpart, such as caching and aggregation of the raw data supplied by IoT devices before processing by IoT applications. In addition, the SDT retains information defining the device type and the SIoT links that have been formed.

SDTs are installed as a virtualized ME application (through containers) and deployed in edge servers. According to the ETSI MEC architecture, the latter, defined as ME hosts, may be connected with BSs/APs [93]. To align our proposal with the ETSI MEC architecture [93], the SIoT-edge framework components, as shown in Fig. 2.10, are considered.

The ME orchestrator in the ETSI MEC architecture has visibility of the edge network's capabilities, which is composed of several ME hosts, and determines the most suitable ME hosts for instantiating the applications (i.e., ME apps) based on application requirements, available resources, and mobility conditions. The orchestrator initiates the migration operation if a virtualized application has to be migrated.

The ME orchestrator determines the ME hosts where each SDT should be placed in the envisioned architecture (see the corresponding functional module in Fig. 2.10). Furthermore,

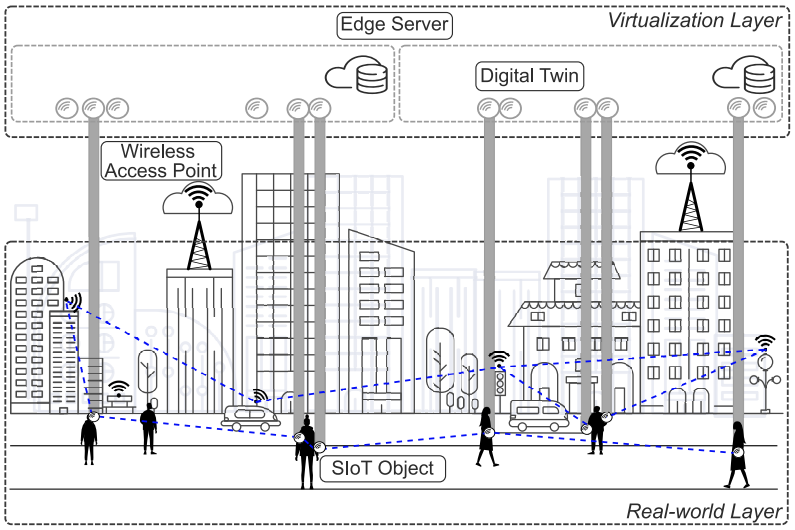

Fig. 2.9: Reference architecture [2].

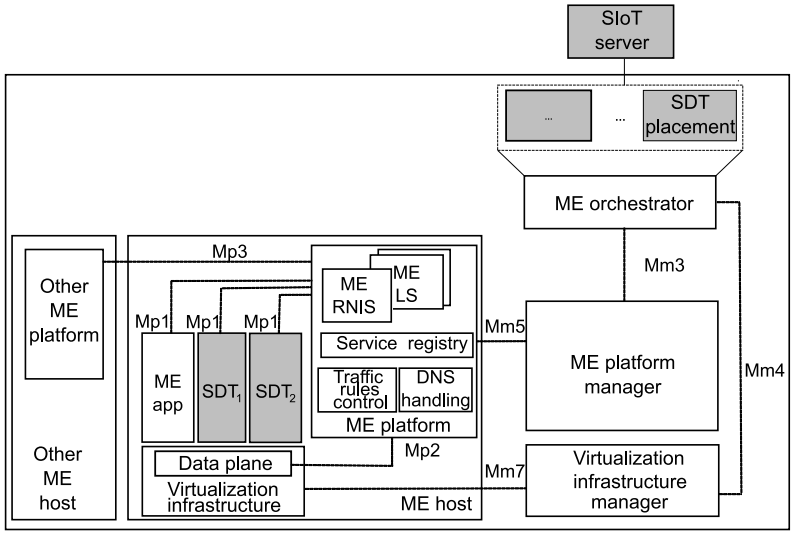

Fig. 2.10: The SIoT-edge framework [2].

the ME orchestrator may communicate with an external SIoT server to receive information about the existing social connections created by any given physical device to select the most suitable placement for its SDT. The SIoT server, in particular, may keep track of the profiles, relationships, and actions of SIoT devices. The device's position information may also be controlled and updated in the profile on the SIoT server.

2.2.2 Mathematical Characterization

In addition to the assumptions of the system model introduced in Section 2.1, the system is assumed to operate according to discrete timing based on a sequence of time slots $t \in \mathcal{T} = \{0, ..., T\}$ with the duration of τ (in minutes), introduced to capture the mobility features

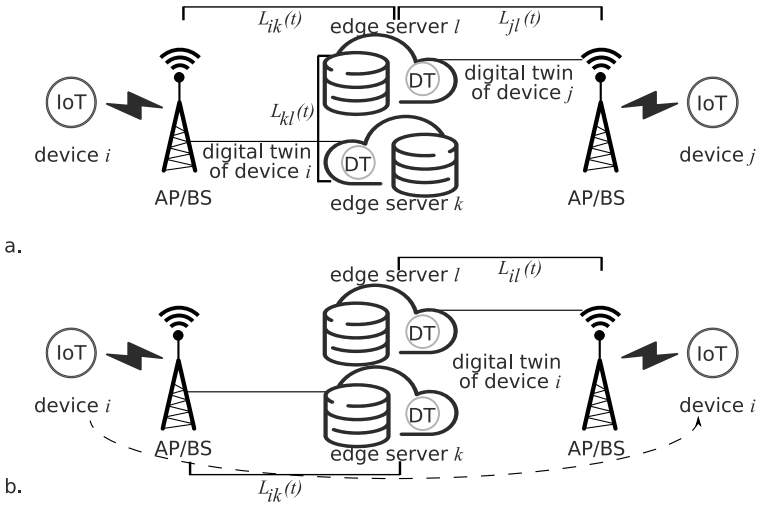

Fig. 2.11: SDT (a) placement and (b) migration [2].

and offer dynamic decisions. The assumption is quite common in the literature [38, 76, 94]. The Illustration of SDT placement and migration in case device i can move and change the connectivity point is presented in Fig. 2.11.

We assume heterogeneity among edge servers, i.e., an edge server k has a finite amount of CPU, disk, and RAM resources, denoted as $aCPU_k$, aD_k, and $aRAM_k$, respectively [95]. Such servers are in charge of hosting SDTs, which are, in turn, associated with IoT devices. SDTs can store data and perform processing, having specific CPU, disk, and RAM demands. The latter ones are indicated for SDT i as $CPU_i(t)$, $D_i(t)$, and $RAM_i(t)$.

The probability $p_{ij}(t)$, $0 \leq p_{ij}(t) \leq 1$, reflects the intensiveness of the data exchange between IoT devices i and j, and is associated with the SIoT links. It is straightforward to assume that the $p_{ij}(t)$ value is strongly correlated with the specific type of social relationship established between the two physical devices [96]. An IoT device linked by an Ownership Object Relationship (OOR) may need to frequently share data about the owner's smart home/car as well as her habits, preferences, and health status. This would not be the case for POR. In case more than one relationship is established between two devices, the maximum $p_{ij}(t)$ value is utilized.

We formulate the optimal SDT placement problem by targeting the following main objectives:

- to jointly minimize the latency between each IoT device and its relevant SDT placed at an edge server and the latency between friend SDTs, while accounting for the relationship existing between the corresponding physical devices, and, hence, for the intensiveness of the expected data exchange;
- to ensure that delay bounds on the interactions between IoT devices and their SDTs are met whenever requested;

- to guarantee effective utilization of the available resources for heterogeneous SDT demands.

We define the objective function as a cost to be minimized, given the sum of two latency contributions. The first component includes the latency experienced between a physical device and the corresponding SDT and is given by:

$$C_1(t) = \sum_{i \in V_P} \sum_{k \in V_S} x_{ik}(t) L_{ik}(t), \tag{2.18}$$

where $x_{ik}(t)$ is the SDT placement decision variable and is equal to 1 if the SDT of device i is placed at edge server k at time slot t, otherwise $x_{ik}(t) = 0$, i.e.,

$$x_{ik}(t) \in \{0, 1\}, \ \forall \ i \in V_P, \ \forall \ k \in V_S, \ \forall \ t \in \mathcal{T}. \tag{2.19}$$

The second component of the objective function includes the latency among SDTs of friend devices at time slot t, i.e., the time required to communicate and discover services querying the friends on the virtualization layer, and is as follows:

$$C_2(t) = \sum_{i \in V_P} \sum_{k \in V_S} \sum_{j \in V_P} \sum_{l \in V_S} x_{ik}(t) x_{jl}(t) p_{ij}(t) L_{kl}(t), \tag{2.20}$$

where $x_{jl}(t)$ is the SDT placement decision variable, $x_{jl}(t) \in \{0, 1\}$.

Hence, we define the total cost at time slot t as

$$C(t) = C_1(t) + C_2(t). \tag{2.21}$$

Allocation Constraint is responsible for the placement of SDTs without replication. Since each SDT is allocated only on one edge server, we have the following constraint for SDT placement decision $x_{ik}(t)$:

$$\sum_{k \in V_S} x_{ik}(t) = 1, \ \forall \ i \in V_P, \ \forall \ t \in \mathcal{T}. \tag{2.22}$$

Latency Constraint is compliant with the idea of meeting the proximity constraint for the SDT of a given physical device and preserves a limitation on the latency between IoT device i and its SDT deployed at edge server k, which is upper bounded by $L_{\max_i}$, i.e.,

$$L_{ik}(t) \leq L_{\max_i}, \ \forall \ i \in V_P, \ \forall \ k \in V_S, \ \forall \ t \in \mathcal{T}. \tag{2.23}$$

Resource Utilization Constraints guarantee efficient resource utilization while preventing the overload of a given edge server k. The constraints ensure SDT placement according to edge server resource availability and guarantee that the capacity constraint (i.e., CPU, THR_{CPU}, disk storage, THR_D, and RAM, THR_{RAM}, utilization) for each edge server at time slot t is not violated when multiple IoT devices simultaneously share the computing resources to host the corresponding SDT at edge servers, $\forall \ k \in V_S, \ \forall \ t \in \mathcal{T}$:

$$\sum_{i \in V_P} \frac{x_{ik}(t) CPU_i(t)}{aCPU_k} \leq THR_{CPU}, \tag{2.24}$$

$$\sum_{i \in V_P} \frac{x_{ik}(t) D_i(t)}{aD_k} \leq THR_D, \tag{2.25}$$

$$\sum_{i \in V_P} \frac{x_{ik}(t) RAM_i(t)}{aRAM_k} \leq THR_{RAM}. \tag{2.26}$$

The SDT placement problem can further be formulated as follows:

$$\min C(t) \tag{2.27}$$
$$\text{s.t. } (2.19), (2.22), (2.23), (2.24), (2.25), (2.26).$$

In each time slot t, an optimal placement can be obtained when solving (2.27) with the exhaustive search.

Lemma 2. Optimal SDT placement in the dynamic large-scale SIoT-edge environment problem is NP-hard.

Proof. We conducted the proof in Section 2.1.1 via a polynomial-time reduction from the quadratic assignment problem, which is known to be NP-hard [82].

We aim to remove the nonlinearity of function (2.27), specifically (2.20), and perform the linearization of the objective function. When elaborating $C_2(t)$, we first denote the cost contributions related to the latency between the SDTs of IoT devices i and j placed at edge servers k and l at time slot t, respectively, as $C_{ikjl}(t)$, by replacing $p_{ij}(t)L_{kl}(t)$. We reformulate (2.20) as follows [97]:

$$\begin{aligned}&\sum_{i\in V_P}\sum_{k\in V_S}\sum_{j\in V_P}\sum_{l\in V_S} x_{ik}(t)x_{jl}(t)C_{ikjl}(t)\\ &=\sum_{i\in V_P}\sum_{k\in V_S} x_{ik}(t)\sum_{j\in V_P}\sum_{l\in V_S} x_{jl}(t)C_{ikjl}(t).\end{aligned} \tag{2.28}$$

We then define $x_{ik}(t)\sum_{j\in V_P}\sum_{l\in V_S} x_{jl}(t)C_{ikjl}(t)$ by introducing $F_{ik}(t)$ and express the minimization of $C^L(t)$:

$$\min C^L(t) = C_1(t) + \sum_{i\in V_P}\sum_{k\in V_S} F_{ik}(t), \tag{2.29}$$

s.t.

$$f_{ik}(t)x_{ik}(t)+\sum_{j\in V_P}\sum_{l\in V_S} x_{jl}(t)C_{ikjl}(t)-F_{ik}(t)\leq f_{ik}(t), \tag{2.30}$$

$$F_{ik}(t)\geq 0,\ \forall\ i\in V_P,\ \forall\ k\in V_S,\ \forall\ t\in\mathcal{T}, \tag{2.31}$$

where $f_{ik}(t)$ is given by

$$f_{jl}(t) = \sum_{j\in V_P}\sum_{l\in V_S} C_{ikjl}(t). \tag{2.32}$$

2.2.3 Graph-Based Heuristic

In real situations, computing the optimal policy solution for SDT deployment is complex. Indeed, the problem is NP-hard. The exhaustive search technique may yield a design for a small network but has no practical use for large networks. In this subsection, we aim to develop a simpler-to-compute approximation solution for the SDT placement problem that achieves near-optimal performance. To simplify the original problem, the method conducts a relaxing transformation, from which a graph-based heuristic is constructed.

Algorithm 1: Graph-based Heuristic

1 **Input:** $G_P(t) = G_P(V_P(t), E_P(t))$; $G_S(t) = G_S(V_S(t), E_S(t))$;
2 **Output:** $V_P(t) \rightarrow V_S(t)$;
3 find sets of connected components $G_{P'}(t) = G_P(V_{P'}(t))$, $|G_{P'}(t)| = n$ such that
$V_{P'}(t) \subseteq V_P(t), E_{P'}(t) \subseteq E_P(t)$,
$\forall u, v \in V_{P'}(t)\ \exists\ (u, v)$,
$\forall u \in V_{P'}(t), w \notin V_{P'}(t)\ \nexists\ (u, w)$;
4 $MAX_W \leftarrow 0$;
5 **while** $G_{P'}(t) \neq \emptyset$ **do**
6 **for** $m = 1 : |G_{P'}(t)|$ **do**
7 find $W(t, m) = \sum_{i,j \in G_{P'}(t,m)} p_{ij}(t)$;
8 **if** $MAX_W < W(t, m)$ **then**
9 $MAX_W \leftarrow W(t, m)$;
10 $G_{P'_{\max}}(t) \leftarrow G_{P'}(t, m)$;
11 **end**
12 **end**
13 find spanning subgraph $T(G_{P'_{\max}}(t))$ such that
$V_P(T) = V_P(G_{P'_{\max}}(t)) \wedge E_P(T) \subseteq E_P(G_{P'_{\max}}(t)), |E_P(T)| = [V_P(G_{P'_{\max}}(t))| - 1$;
14 $G_{P'}(t) \leftarrow G_{P'}(t) \setminus G_{P'_{\max}}(t)$;
15 find optimal mapping $\Pi(t) = \{\pi(t) : V_P(T) \rightarrow V_S\}$;
16 **end**
17 **return** $V_P(t) \rightarrow V_S(t)$.

Due to its practical and theoretical relevance and complexity, the QAP has drawn the attention of researchers all over the globe. The QAP is one of the most challenging combinatorial optimization problems. However, to the best of our knowledge, there is no theoretical evidence for convergence of quality and computing time, particularly for large-scale dimension issues. In this work, we concentrate on reaching a short enough execution time while guaranteeing a decent approximation of the optimal solution.

We provide an approximation approach based on a graph-theoretic solution for the SDT placement problem. Algorithm 1 presents the pseudocode of the graph-based heuristic executed for each time slot $t \in \mathcal{T}$.

The formulation of the approximation solution in terms of graph theory is as follows. Let $G_P(t)$ be a weighted connected graph, $V_P(t)$ be the set of vertices of graph $G_P(t)$ corresponding to the SDTs, and $E_P(t)$ be the set of links of the graph $G_P(t)$ defining the connections between the SDTs allocated at edge servers. Let $V_S(t)$ be a finite set of positions intended for assigning vertices of the graph $G_S(t)$ corresponding to the set of edge servers.

Algorithm 1 begins with the definition of a connected component $G_{P'}(t)$ of graph $G_P(t)$, i.e., identification of individual connectivity components (line 3). It allows defining the num-

ber of strongly connected components in which a path from each vertex to another vertex exists. The algorithm considers each component separately (lines 5-16), starting from the strongest one between vertices of a component (lines 6-12).

Then, Algorithm 1 finds an approximating spanning subgraph or, in other words, a spanning tree (line 13). In the field of graph theory, a spanning tree T of an undirected graph $G_{P'}(t)$ is a subgraph that is a tree, which includes all of the vertices of $G_{P'}(t)$ with a minimum possible number of links. If all of the links of $G_{P'}(t)$ are links of a spanning tree T of $G_{P'}(t)$, then $G_{P'}(t)$ is a tree and is identical to T. The advantages of spanning tree usage and, therefore, problem simplification are the following. First, constructing a spanning tree takes a polynomial time when using well-known algorithms. Second, the problem of tree placement can be solved relatively quickly.

We then perform mapping $\Pi(t)$ (line 15) by placing the vertices of graph $G_P(t)$, assuming that vertex $i \in V_P(t)$ is allocated in the position $\pi(i) \in V_S(t)$, such that any vertex of $V_S(t)$ can either accommodate vertices of $V_P(t)$ or accommodate no vertices. The set of all mappings of set $V_P(t)$ into set $V_S(t)$ is given by

$$\Pi(t) = \{\pi(t) : V_P(t) \to V_S(t)\}. \tag{2.33}$$

We specify the following parameters. First, the distance $L_{ik}(t)$ between vertex $i \in V_P(t)$ and position $k \in V_S(t)$, defined in terms of the latency of the connectivity between physical device i and its SDT placed at edge server k. We refer to as distance $L_{ik}(t)$ the cost of placing vertex $i \in V_P(t)$ in position $k \in V_S(t)$. We then define the weight of the edge $p_{ij}(t)$ associated with the probability of data exchanging between IoT devices; the distance L_{kl} between positions $k, l \in V_S(t)$, defined in terms of the latency of the connectivity between edge server k and edge server l. The cost of communication between vertices $i, j \in V_P(t)$ placed in positions $k, l \in V_S(t)$ corresponds to $C_{ikjl}(t) = p_{ij}(t)L_{kl}(t)$.

As we aim to allocate the vertices of graph $G_P(t)$ in positions $V_S(t)$ by minimizing the total cost of placing the vertices $V_P(t)$ to positions $V_S(t)$, the problem is formulated in terms of mappings as follows:

$$\min_{\pi(t)\in\Pi(t)} \{ \sum_{i\in V_P(t)} L(i, \pi(i)) + \sum_{i\in V_P(t)} \sum_{j\in V_P(t)} C(i, j, \pi(i), \pi(j))\}. \tag{2.34}$$

Algorithm 1 defines the most suitable positions to host vertices of the spanning tree (T) to minimize the cost of the spanning tree's vertices assigned to the set of positions $G_S(t)$ (2.34). The Algorithm finishes when all vertices of $G_P(t)$ are mapped onto positions that belong to $G_S(t)$ by meeting constraints (2.19), (2.22)-(2.26).

Complexity Analysis: The computational complexity of Algorithm 1 is provided by

$$O(n) \cdot O(n) = O(n^2),$$

where n is the complexity due to the *while* cycle across all $|G_p(t)| = n$ vertices of graph $G_P(t)$ in the worst case when the number of connected components of graph $G_P(t)$ equals to the number of graph vertices (lines 5-16). For the second component, which is included

in the *while* cycle, n represents the complexity to search for the biggest connected component in terms of the number of communications between vertices (lines 6-12). Since initially $|G_{p'}(t)| = n$ and at each iteration one of the elements is deleted, this inner loop is executed n times first, then $n-1, n-2$, and so on until at the last iteration, the inner loop runs only once. The complexity of the sum $1+2+...+(n-1)+n$ is challenging to be precisely determined (lines 6-12). Instead, we determine an upper limit for it, which is $O(n)$. This means that every time inner loop runs exactly n times. The complexity of lines 13-16 is $O(1)$, but it executes within *while* cycle, hence $O(n \cdot 1) = O(n)$, which does not affect the complexity of the first component. Consequently, the maximum number of operations is in $O(n^2)$.

Differently, the optimal SDT placement problem in the dynamic large-scale SIoT-edge scenario has been stated to be an NP-hard QAP. The NP-hard problems are solvable, but not in polynomial time; that is, no solutions produce a result in $O(n^k)$ for any constant $k \geq 2$. In addition, QAP is one of the most difficult combinatorial optimization problems [98]. While theoretical, algorithmic, and technical advancements have resulted in considerable improvements in the sizes of solvable problems for many well-known NP-hard problems, QAP has remained a class that defies efforts to solve it except for extremely small sizes [98]. The QAP general form necessitates the inclusion of $O(n^4)$ cost terms for $C_2(t)$ [99]. The minimum number of operations for the dynamic, large-scale SIoT-edge environment issue is thus $O(n^4)$.

Therefore, the proposed heuristic guarantees a theoretical complexity substantially less than the formulation of the optimal problem.

2.2.4 Performance Evaluation

This subsection aims to assess the performance of the proposed SDT placement optimization strategy. First, we detail the simulation campaign, including the scenario, parameters, benchmark schemes, and measures of interest. Then, using a simulator tool based on the IBM ILOG CPLEX Optimization Studio 12.10.0 and Matlab R2021b software, we compare the results achieved through the optimal and approximation solutions, i.e., the graph-based and benchmark placement schemes. Simulation parameters are reported in the remainder of the subsection and are gathered in Table 2.5.

Simulation Settings

Similarly to [87], we consider the city center of Santander (Spain), which has an area of 4 km x 4 km. We assume the hexagonal grid cellular layout, in agreement with the Third Generation Partnership Project (3GPP) specifications [100], with $M = 8$ BSs. An edge server is associated with each BS, in agreement with the ETSI documents [80].

We evaluate the proposal based on a realistic object behavior taken from the large dataset generated in [87] that tracks device interactions based on real IoT objects and the Small World In Motion (SWIM) mobility model [101]. SWIM is based on a simple intuition about

Table 2.5: System parameters related to Section 2.2.

Area of interest	**Area:** Santander, Spain [87] **Size:** 4000 m x 4000 m [87]
Users	**Number:** 50 (100) **Mobility pattern:** SWIM [87, 101]
IoT Devices	**Smartphones:** 6% (12%), **Cars:** 15% (14%), **Tablets:** 12% (11%), **Smart Fitness:** 20% (24%), **Smartwatches:** 29% (25%), **PCs (static):** 1% (6%), **Printers (static):** 10% (2%), **Home Sensors (static):** 7% (6%) **Total number:** 113 (328) [87]
Social network	**Probability of data exchange:** 1 (OOR), 0.1 (C-LOR), 0.1 (SOR), 0.1 (POR) *: 1 (OOR), 1 (C-LOR), 0.1 (SOR), 0.1 (POR)
Base stations	**Cell layout:** 3GPP hexagonal grid [100] **Number of BSs:** 8 [100] **Cell area radius:** 450 m [100] **Intersite distance:** 1350 m [100]
Edge servers	**Deployment:** Co-location with BSs [80] **Number:** Equal to number of BSs [80] **Distance:** Geographical distance [37, 38] **Latency:** Proportional to the distance (with ϵ 3.33 ms/km [81]) **CPU capability:** 24000 MIPS [102] **Disk capability:** 2 TB [102] **RAM capability:** 24 GB [102]
Resource utilisation constraints	**CPU utilization threshold:** 0.6 [103] **Disk utilization threshold:** 0.9 [103] **RAM utilization threshold:** 0.9 [103]
Disk	**Disk demands:** Uniformly distributed in [10, 50] GB [95]
CPU demands / RAM demands	**High-CPU medium instance:** 2000 MIPS/0.85 GB [104] **Extra large instance:** 2500 MIPS/3.75 GB [104] **Small instance:** 1000 MIPS/1.7 GB [104] **Micro instance:** 500 MIPS/613 MB [104]
IoT devices-SDTs	**Distance:** Geographical distance [37, 38] **Latency:** Proportional to the distance with ϵ 3.33 ms/km [81]
Proximity constraint	**Physical device-SDT maximum latency:** Uniformly distributed in [1, 10] ms [105]

human mobility, i.e., people often go to places close to their homes and the most popular places. We consider two device density settings (i.e., portions of the dataset), namely, 50 and 100 users with $N = 113$ and $N = 328$ heterogeneous IoT devices, respectively, spanning from static devices to consumer devices carried by users moving in the selected area of interest (see Table 2.5 for the percentages of each type of IoT devices).

In edge facilities, an SDT is coupled with each IoT object and implemented as a container [78]. To account for the heterogeneity of IoT devices without sacrificing generality, we correlate SDTs with four types of containers (based on the respective device types) according to CPU requirements [102–104], as shown in Table 2.5. For instance, cars with autonomous navigation assistance and smartphones may demand a powerful CPU for their SDTs. In contrast, smartwatches, sensors, tablets, intelligent fitness devices, and printers may be linked with small and micro instances, respectively. The maximum latency between a physical de-

vice and its SDT is uniformly distributed within the interval $[1, 10]$ ms [105]. The latency is calculated according to the geographical distance between any two entities deployed in the reference region [37, 38, 81].

Social relationships are associated with each device in the dataset. For the extracted 113 devices, the percentage of established relationships corresponds to 50%, 21%, 15%, and 14%, for OOR, Co-Location Object Relationship (C-LOR), Social Object Relationship (SOR), and POR, respectively. For the 328 devices setup, the percentage of established relationships is as follows: 60% OOR, 8% C-LOR, 1% SOR, and 31% POR.

Unlike all the settings mentioned above, there are no specific clues about how to set the parameter $p_{ij}(t)$. Hence, without loss of generality, we consider a set of representative values throughout the simulation campaign to understand their impact on the SDT placement strategy compared to solutions oblivious to the social relationship information (see in the following). In particular, we fix $p_{ij}(t) = 0.1$ for SOR and POR. The logic for these numbers is that IoT devices with intermittent interaction (i.e., linked by a SOR) or belonging to the same brand or product batch (i.e., tied by a POR) are not expected to exchange large amounts of data frequently, but rather only in response to certain situations. For instance, POR-connected devices may seldom share software updates. Data transfers between objects of the same owner, i.e., those linked by an OOR, may be frequent and substantial, for example, to synchronize personal/health data or monitor the smart home. Hence, we set for OOR $p_{ij}(t) = 1$, and we vary it to be 0.1 and 1 for those devices tied by a C-LOR. However, the strategy is flexible enough to accommodate other settings.

The entire simulation covers a time-lapse of 5 hours. Within this period, we first simulate time slots, t, of duration equal to $\tau = 5$ minutes [76]. Then, we vary this setting up to 30 minutes in steps of 5 minutes. We determine the best time slot duration based on the analysis of simulation results. During each time slot t, we assume that the SDT placement does not change [76, 94].

We compare the optimal solution (labeled in the curves as eSoCEP with the following approximation solutions and placement strategies:

- Approximation techniques:
  - Proposed graph-based heuristic, as per Algorithm 1, labeled in the curves as eSoCEP Heuristic [2].
  - Local Branching (LB) [106], labeled in the curves as LB.
  - Relaxation Induced Neighborhood Search (RINS) [107], labeled in the curves as RINS.
- Placement strategies:
  - SoCEP, labeled in the curves as SoCEP, which takes into account proximity and social requirements. We also simulate the graph-based heuristic for SoCEP, labeled in the curves as SoCEP Heuristic [1].

- CEP, labeled in the curves as CEP, according to which SDTs paired with physical devices are always placed at the nearest edge server by neglecting social features [37, 38, 75, 108].
- Static Placement, labeled in the curves as No Migration, is a strategy according to which SDTs are initially placed at the nearest edge server and keep the same placement throughout the whole simulation duration without migration possibilities [47, 76].

For a fair comparison, a dynamic placement is also triggered at every time slot for both SoCEP and CEP.

We evaluate the performance of proposed and benchmark solutions by leveraging the following metrics:

- *Average latency between IoT devices and their SDTs.*
- *Average latency among friend SDTs.*
- *Total number of migrated SDTs.*

Performance Analysis

The first set of results aims to validate the effectiveness and efficiency of the proposed eSoCEP heuristic (for both $N = 113$ and $N = 328$: average number of friend IoT devices $f = 4$ for $N = 113$ and $f = 5$ for $N = 328$) when compared to the optimal solution, the considered relaxation techniques, and the benchmark solutions for time interval duration $\tau = 5$ min [76], probability of data exchange $p_{ij}(t) = 1$ for OOR and $p_{ij}(t) = 0.1$ for C-LOR, SOR, and POR. As shown in Fig. 2.12, eSoCEP and eSoCEP Heuristic preserve close values for all considered metrics and under all device density settings. This is especially true for the latency among friend SDTs, for which values are significantly lower compared to the CEP and No Migration benchmarks. Instead, the device-SDT latency, although higher for the eSoCEP Heuristic compared to the optimal solution, is in any case bounded by the proximity constraint.

Furthermore, we examine the average latency among friend SDTs per relationship category, as illustrated in Fig. 2.13 and Fig. 2.14 (time interval duration $\tau = 5$ min [76], probability of data exchange $p_{ij}(t) = 1$ for OOR and $p_{ij}(t) = 0.1$ for C-LOR, SOR, and POR). The optimal eSoCEP solution guarantees zero latency among OOR friends for $N = 113$ and the lowest values for $N = 328$. This means that SDTs of friend devices are co-located in the same edge server, in alignment with the targeted objectives, well captured by the parameter $p_{ij}(t) = 1$. Higher latency values are measured for the other types of relationships. In particular, the highest latency values are experienced among POR friends in the case of $N = 113$ IoT objects because devices establishing such a type of relationship are more likely spread throughout the topology.

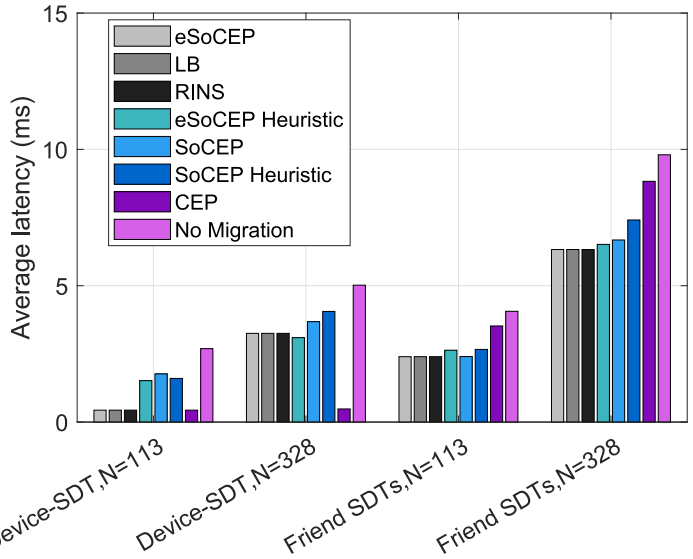

Fig. 2.12: Latency between an SDT and a device and among friend SDTs [2].

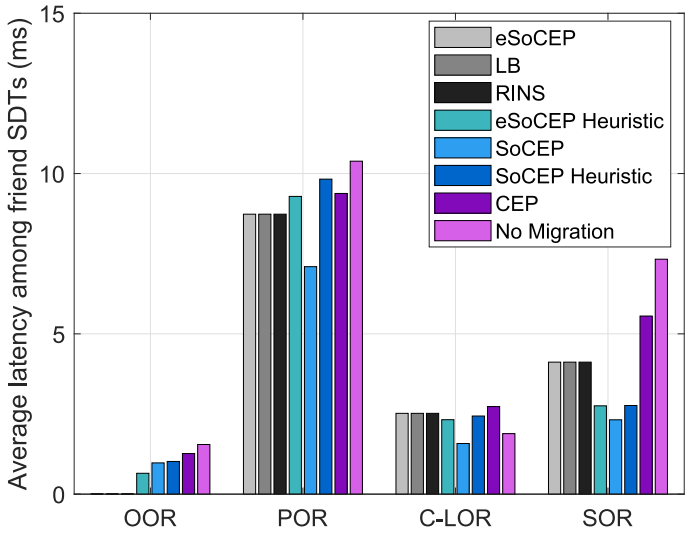

Fig. 2.13: Latency among friend SDTs per relationship type, $N = 113$ [2].

Next, we evaluate the computational complexity. To this end, Fig. 2.15 shows execution time as a function of the number of devices. We examine placement strategies with simulations on an Intel (R) Xeon(R) CPU E5 – 2620 v4 at 2.10 GHz with 19.7 GB RAM.

We start by comparing eSoCEP and SoCEP heuristics with the optimal solver. In eSoCEP, the introduction of stricter constraints on latency and resource usage as well as a linearization of the optimization function, which allows for faster optimal solution search, leads to a significant reduction in complexity. Contrary to eSoCEP, the SoCEP strategy fails

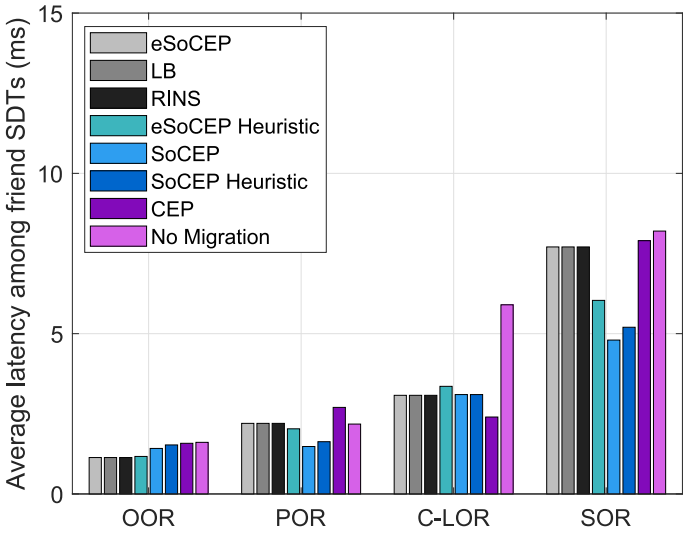

Fig. 2.14: Latency among friend SDTs per relationship type, $N = 328$ [2].

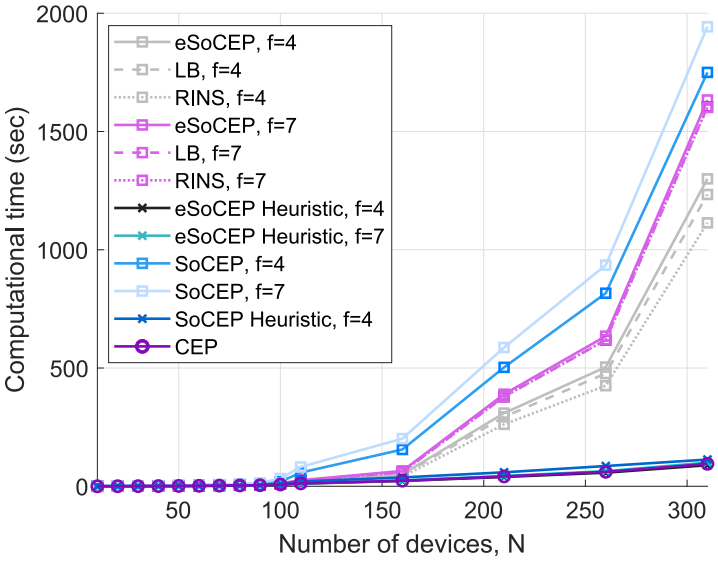

Fig. 2.15: Computational time [2].

to scale as the number of devices and the average number of friends increase. Moreover, we note that LB and RINS (dashed lines) applied to the eSoCEP solution respectively decrease the running time on average by 7.5% and 11.5% compared to eSoCEP for the average number of friend IoT devices, f, equal to 4, and by 8.8% and 11.3% for $f = 7$.

From the results in Fig. 2.15, it further emerges that the eSoCEP heuristic outperforms all the considered placement strategies and approximation solutions. It offers, on average, 43.3% and 46.9% reduction in the running time compared to the optimal solution for $f = 4$

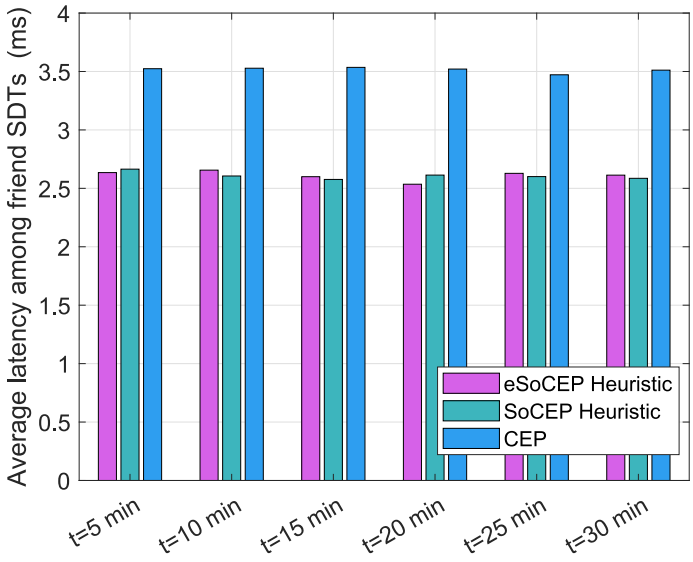

Fig. 2.16: Latency among friend SDTs [2].

and $f = 7$, respectively. For $N = 328$, such reduction is up to nearly 94%. Here, approximated spanning sub-graph usage is beneficial for the following reasons. First, generating a spanning subgraph in polynomial time is feasible when applying well-known methods. Second, the problem of tree location may be rapidly resolved. Regardless of f, the eSoCEP heuristic maintains a low enough computing time. Such an observation on the scalability of the suggested heuristic with the number of friends is especially relevant, given that IoT devices are expected to form numerous relationships. Interestingly, the eSoCEP heuristic, which represents a more advanced SDT placement technique, requires the same calculation time as the simplest CEP method, whose exact results are shown in the following.

We have analyzed the results of the dynamic placement strategies when fixing the time slot to 5 minutes, similarly to [76]. Fig. 2.16 reports the average latency contributions for different time interval duration, τ, for the average number of friend IoT devices f=4, probability of data exchange $p_{ij}(t) = 1$ for OOR and $p_{ij}(t) = 0.1$ for C-LOR, SOR, and POR, $N = 113$.

SoCEP and eSoCEP outperform other schemes in terms of latency among friend SDTs (Fig. 2.16), indicating their capacity to account for social connection needs, as specified by the formulated problem. Although close to 4 ms delay values are measured for the No Migration policy, they are omitted from Fig. 2.17 to reduce redundancy.

From Fig. 2.17 (average number of friend IoT devices f=4, τ=20 minutes, probability of data exchange $p_{ij}(t) = 1$ for OOR and $p_{ij}(t) = 0.1$ for C-LOR, SOR, and POR, $N = 113$), interestingly, the proposed eSoCEP solution is more efficient than SoCEP. It can be observed that it always triggers fewer migration events compared to SoCEP. As a consequence, the overhead incurred by migration procedures is lower, i.e., a lower amount of data is exchanged

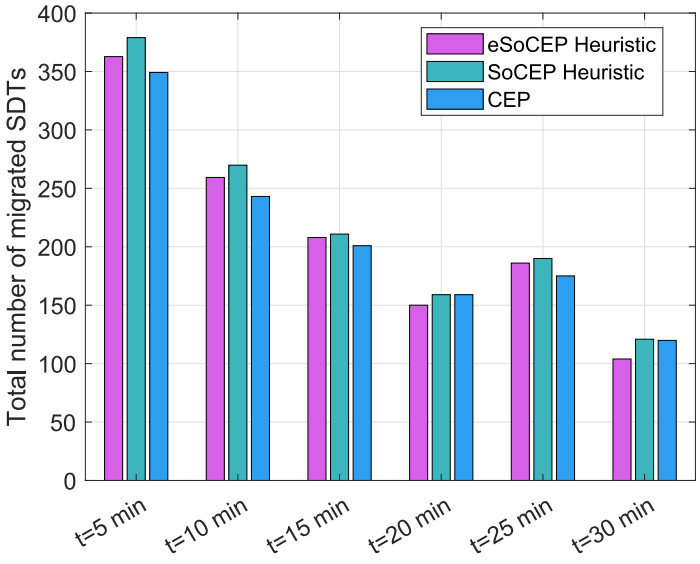

Fig. 2.17: The number of migration events [2].

over backhaul links interconnecting the edge servers acting as the source and destination of migrating SDTs. It can be further noted that, unlike latency metrics, the total number of migration events is vulnerable to the time slot settings, t. Such a finding suggests further investigating the results for a specific τ value, i.e., 20 minutes, which allows achieving a trade-off between the number of migration events and the latency metrics.

2.2.5 Discussions

In this section, we developed a framework for the dynamic placement of DTs associated with physical IoT devices establishing social relationships.

The proposed placement strategy considers the social characteristics of IoT devices, their motion patterns, and the limited computational capabilities of edge servers. A QAP has been developed to optimize SDT placement. We designed a heuristic to handle the complexity of the optimization problem. The numerical results revealed that the proposed graph-based heuristic preserves polynomial time complexity while maintaining results close to the optimal solution.

The proposed placement, eSoCEP, achieves the goal of assuring the lowest latency among SDTs of friend devices that are more likely to share data, such as those linked by OOR, while ensuring adequate proximity between physical devices and their counterparts. Lower latency among SDTs has two benefits: *(i)* it reduces network strain when SDTs exchange data because packets transit fewer connections, and *(ii)* it ensures rapid interactions among them, which is critical for service discovery methods that include traversing the social graph.

The eSoCEP heuristic is demonstrated to be efficient in terms of computation time when compared to the benchmark solutions, with its execution even quicker than the most com-

mon placement technique, CEP, which is myopic in terms of the need to assure closeness among SDTs of friend devices. Furthermore, the eSoCEP heuristic is faster and more efficient in terms of placement decisions than the heuristic for our previous approach, SoCEP. In comparison with SoCEP, eSoCEP has fewer migrations. Hence, a lower communication footprint is incurred since data must be exchanged from the source to the target edge server whenever a migration is triggered to match a new placement decision.

However, when the latency limitations for communication between physical IoT devices and their digital counterparts are not met, the placement selection can be dynamically adjusted. In this case, the difficulty of executing real-time DT replacement is connected not only to the requirement to rerun the optimization model but also to the movement of the SDT from one edge server to another (see, e.g., [109]). How to enable smooth migration among edge servers is still an open problem in the literature. These concerns are addressed in the next section (Section 2.3).

2.3 Social-Aware Digital Twin Orchestration Under Heterogeneous Mobility

The deficiencies of the 5G mobile system as a platform for SIoT applications are currently driving research efforts toward B5G wireless networks [110, 111]. Such systems are envisioned to revolutionize next-generation applications and services by ensuring intelligent and autonomous operations. Moreover, recent progress in wireless communications, edge computing, and intelligent technologies are likely to fuel the growth of SIoT with pervasive sensing and computing capabilities [112]. This will ensure seamless connections and autonomous management among SIoT objects without human interaction, potentially changing industries and providing major societal advantages [113].

The constantly growing heterogeneity, fuelled by an avalanche of more intelligent and capable SIoT devices, necessitates an order-of-magnitude improvement in energy efficiency without sacrificing communication quality. The level of support provided by the operator infrastructure affects resource utilization efficiency. To this end, user-centric and network-centric techniques have been proposed to efficiently control user connectivity and improve performance. The former relies on end-user decisions by improving connectivity patterns. The latter entails a central coordinating unit to make decisions based on system-wide data collected in a timely manner across the network. However, one of the crucial considerations in both approaches is the re-optimization frequency [114]. It is related to the periodicity with which the associated computing protocols might be run on the network infrastructure side, as well as the actual device latency and user experience [115].

As for user-centric approaches, the higher the frequency of re-optimization, the closer the system's time-averaged performance will remain to the optimal state values [116–118]. However, due to prohibitive signaling and computation overheads, the re-optimization frequency cannot be arbitrarily high in network-centric schemes. Moreover, in practice, the re-optimization periods are fixed so that the resulting decisions are left intact for the re-optimization interval. The constant re-optimization interval might lead to the following situations: *(i)* the system's time-averaged behavior may deviate from the optimum due to unexpected SIoT device mobility depending on the device type, time, space, and situation; *(ii)* the system's time-averaged behavior may trigger wasteful network resource usage, e.g., when the devices are static, i.e., during nighttime.

In this section, differently from recent academic and industry efforts, we study time-, device-, space-, and scenario-depending network re-optimization intervals and their enablers for optimal deployment of SDTs at the network edge proposed in Section 2.2 (as illustrated in Fig. 2.18).

Specifically, we present the following contributions:

- we offer a review of device motion patterns that might depend on the time, the device type, space, and the scenario;

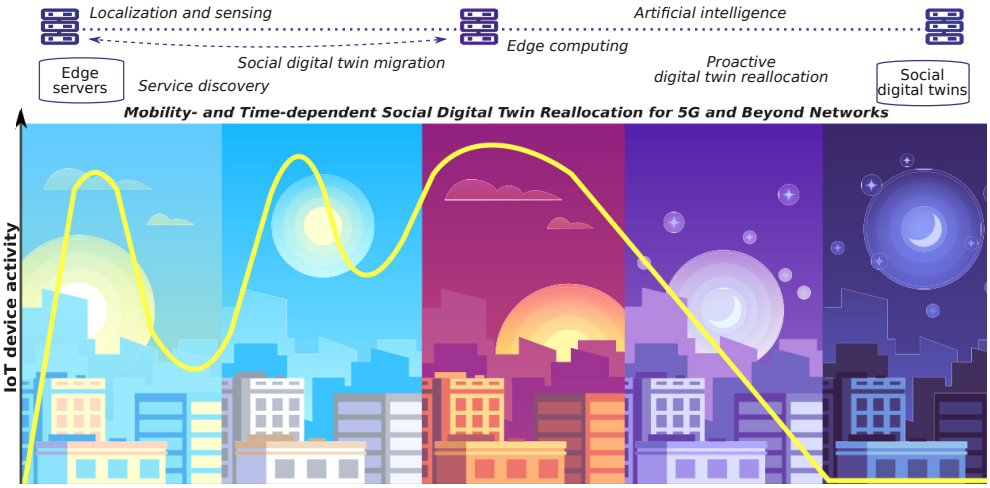

Fig. 2.18: Convergence of IoT device motion pattern and SDT reallocation on the edge [3].

- we propose a design of social-aware orchestration comprising NFV, SDN, edge/fog and cloud computing, DL-based user activity prediction, and sensing and tracking technologies;
- we evaluate the performance of the proposed orchestration;
- we analyze the re-optimization time interval concerning the impact on the service discovery latency for traditional system design and co-design of localization, sensing, and AI-driven communication and computation.

The rest of this section is structured as follows. In Subsection 2.3.1, a review of device motion patterns that might depend on the time (e.g., morning, day, evening, or night hours), the device type (e.g., static sensor, cars, and wearable devices), space (i.e., environment), and the scenario (e.g., critical situation or everyday routine actions) is offered. In Subsection 2.3.2, a social-aware orchestration comprising NFV, SDN, edge/fog and cloud computing, DL-based user activity prediction, and sensing and tracking technologies. In Subsection 2.3.3, a simulation campaign on the re-optimization time interval concerning the impact on the service discovery latency for traditional system design and co-design of localization, sensing, and AI-driven communication and computation is performed. Discussions are drawn in Subsection 2.3.4.

2.3.1 Mobility Behaviour in Beyond-5G SIoT Environment

SIoT involves interconnected heterogeneous mobile devices with movements depending on time, device type, space, and situation.

Time-Dependent Mobility

Time is one phenomenon that affects SIoT device motion. Recent literature has been rich in daily pattern examination. In [119], the days and hours with the highest level of motion activity have been analyzed. More specifically, there exist peak hours (i.e., 7 am–9 am, 12 pm–1 pm, and 4 pm–6 pm) and non-peak hours (i.e., 10 am–11 am). In [120], the morning and evening peaks begin from 7 am and 5 pm and end at 8 am and 6 pm.

Moreover, the motion patterns vary within the days of the week [121, 122]. The busiest days are usually weekdays. In contrast to weekday trends, weekend pedestrian traffic displays a consistent and progressive increase throughout the day, peaking in the early afternoon and gradually decreasing in the evening. In addition to a deviation between weekdays and weekends, each day has its peculiarity. For example, Friday evening hours are usually more active than other weekdays [119]. Furthermore, the night activity is more frequent during the night on weekends and Fridays than during Monday-Thursday time intervals [119, 120].

Furthermore, the user motion patterns correlate with population growth. According to the estimates in [119], the average annual growth rate of detections has been around 3.69% with over 500 million devices identified by 2030. Research in this field indicated that other factors, such as engagement with a specific device or application [5], environment (e.g., a city with a specific regime) [119, 120], or even a situation, may also influence user motion.

Device-Dependent Mobility

The next phenomenon that triggers deviations in mobility patterns is the device type, i.e., bicycles, trains, cars, sensors, or people with wearable devices, AR, VR, XR, holographic glasses, and smartphones. This area is rich in research experiments and studies.

We first introduce the investigations for bicycle patterns [121] and compare them to daily pedestrian motion. The first peak hours for bicycle activities correspond to typical workplace start times and are often between 8 am and 10 am, whereas pedestrian activity initiates one hour earlier. Around 2 pm, a second peak initiates (during people's lunch breaks) and terminates at 4 pm in comparison with 12 pm–1 pm, and 4 pm–6 pm time intervals of the boosted activity of pedestrians. Finally, at around 7 pm, the evening peak hours begin (typical end time of most workdays).

The weekend pattern differs from weekdays because it does not include the early morning peaks. Instead, 8 am is the least active hour. The activity continuously increases until 2 pm, right before lunchtime, when it declines. After that, the activity rises again, following a pattern similar to that seen during working days with peaks at 4 pm and 8 pm, but the difference is less obvious than on working days [121]. Even though the bicycle activity during the weekdays is somewhat close to pedestrian behavior, the weekend motion differs significantly. Moreover, the speeds and paths taken to reach certain destinations vary, and are specific to each device.

Another set of investigations has been concentrated on analyzing XR glasses or Head-Mounted Displays (HMDs) compared to mobile phone motion patterns. The findings are consistent across research communities, demonstrating that XR usage results in diverse movement patterns compared to mobile phone usage due to unique content presentation and navigation experience [123]. More specifically, wearing an HMD causes movements with shorter stride lengths, longer stance time, and higher speed variability. Furthermore, walking of users engaged in message writing and audio recording, for example, is distinct in mobile phone applications. Typing a message on a tiny phone screen necessitates much concentration due to mobility restrictions, but voicemails can be received and transmitted without limitation. Moreover, when walking with a cell phone while dual-tasking, walking pace changes are substantially lower than when walking alone. Variations in single- and dual-task phases, on the other hand, modestly lower head-up walking, validating XR stability and multitasking sustainability [5].

Space-Dependent Mobility

Depending on the location, mobility patterns differ and have diverse degrees of unpredictability [121, 122]. Although human movements and activities fluctuate over time and across sites, there is a pattern of geographical dependence in the observed activity and information flow [124].

As an example of motion that is dependent on location, the study of bicycle traffic [121] found that the average patterns of activity at individual bicycle stations (local activity cycles) differ from the global patterns presented in Section 2.3.1. The peaks of activities near a university are between 8 am and 1 pm, which is typical for an institution that offers morning classes or workplaces. However, the second peak in the afternoon begins at 3 pm and ends at 4 pm. This could be due to people leaving the institution for lunch or a shift change between morning and afternoon classes. Finally, after 8 pm, there is a spike in the activity, most likely due to the popular nearby area of bars and restaurants. Moreover, the weekend activity patterns begin later than during the weekdays.

Another site that has been investigated in [121] is next to a hospital and office buildings. The increase in activity in such locations is at around 8 am, which is more likely to be caused by a steady work schedule in businesses or hospitals than fluctuating start times of university courses. Typical residential districts, where people leave in the morning and return later in the afternoon or evening, exhibit the opposite behavior compared to the above-mentioned sites. Furthermore, places next to malls on weekends show a unique bimodal distribution, possibly due to the attraction of afternoon visitors [121]. These studies only cover a small portion of the research on mobility that is dependent on location. Still, one may see the complexity of the motion behavior considering heterogeneous SIoT devices in various environments.

Situation-Dependent Mobility

Finally, we analyze situations different from everyday life, i.e., emergency scenarios. There might appear unexpected but situation-triggered peaks in activity [125]. On the other hand, it may result in complete cessation of movement (such as during a lockdown). Despite the apparent necessity for such research, few works can be found in the literature due to a lack of data on spatiotemporal movement patterns during catastrophes, disasters, or other rare events [124].

In summary, device mobility is an aspect of SIoT systems that depends on various factors such as time, device type, space, and situation, not to mention other features, i.e., age of users and family status [126], which are out of the scope of our work. However, these behavior patterns are expected to affect service discovery procedures, depending on the SDT placement at the network edge, and introduce new challenges, especially in systems that rely heavily on large amounts of data, i.e., SIoT systems.

2.3.2 Social-Aware Orchestration

We overview social-aware service discovery and the framework for SDT placement at the network edge. We propose an advanced orchestration in B5G-IoT ecosystems that can address the above-mentioned mobility issues.

Service discovery involves discovering the objects that offer services relevant to the users. When objects submit a request to find nearby objects/services, the request is matched with the available objects/services that exist in their exact or nearby location and match their preferences and interaction history. The requester receives a list of all applicable objects/services that fulfill their quality of service level [127, 128].

The following classification of discovery systems is based on the physical features of the SIoT objects and their surrounding environment:

- *Location-based service discovery* establishes a spatial social structure among objects.
- *Time-based service discovery* identifies services by constructing a temporal social structure among objects.
- *Spatiotemporal-based service discovery* finds services by constructing a spatiotemporal social structure among objects.
- *Event-based service discovery* uncovers services by building social relationships between objects based on real-world occurrences.

Therefore, applications and services might leverage data and resources provided by the social network of IoT objects when services use data from particular categories of friend objects. To this end, the SIoT builds social structures between IoT objects and people that intend to deliver services as follows. Parental object relationships are utilized when a software patch might need to be delivered to devices or sensors of the same brand or model. In the case of person-level service, certain data might reach all other devices belonging to the same

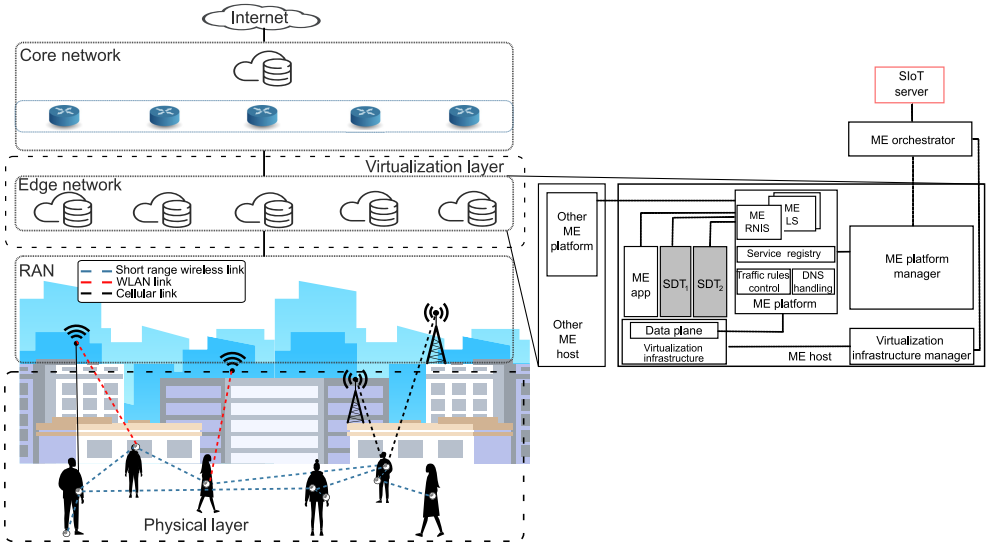

Fig. 2.19: Framework for SIoT communications [3].

user linked by an OOR. Devices that are currently in the same area frequently and visit the same location may be marketed with business services linked by C-LOR and Co-Work Object Relationship (C-WOR), respectively. SOR covers composite services that can be created using the resources given by the resource chain [129].

Existing Orchestration

We offer a short overview of the SIoT communication framework [2] (partially presented in Section 2.2) that ensures effective service discovery and comprises the real-world and the virtualization layers (see Fig. 2.19). The real-world layer accommodates SIoT devices that can be either connected via D2D and/or with other remote entities, either through a gateway node (e.g., a smartphone) or directly through the 5G/B5G Radio Access Network (RAN) facilities. The virtualization layer is in charge of hosting the SDTs, which provide storage and computing capabilities to physical devices.

The SDT stores information on all the social links created by the corresponding physical device according to the SIoT paradigm [19]. Specifically, the SDT holds information specifying the friend device type and SIoT relationship. They are installed as virtualized applications, i.e., containers, on edge servers (ME hosts), which are associated with BSs/APs. The RAN may encompass both 3GPP and non-3GPP access.

The ME orchestrator is aware of the resources and capabilities of the edge network, which comprises multiple ME hosts, and determines the most appropriate ME hosts for instantiating applications based on requirements (e.g., latency, processing requirements), available resources, and mobility conditions. In addition, an SDN controller may organize the backhaul links interconnecting BSs/APs and, therefore, ME hosts. In the architecture, it is responsible for determining the ME hosts where each SDT should be installed. Furthermore,

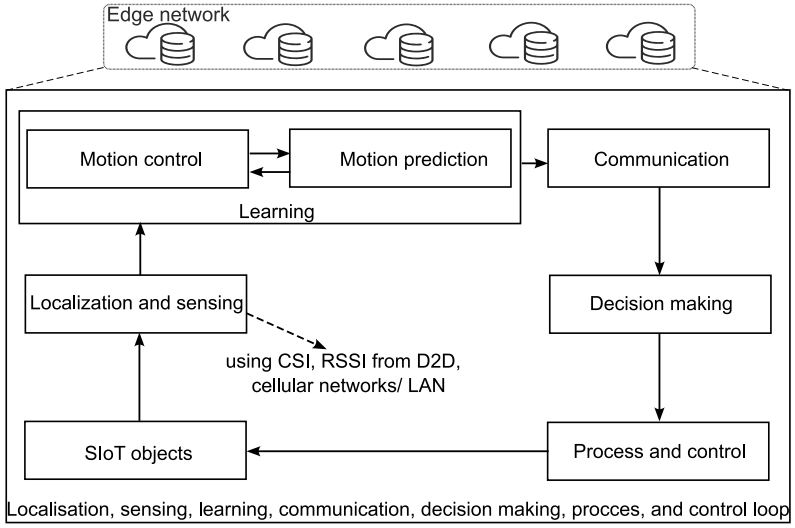

Fig. 2.20: Localization, sensing, learning, communication, decision making, process, and control loop [3].

the ME orchestrator may interface with the SIoT server to receive information about the social links between devices to determine the optimal placement of SDTs.

Proposed Orchestration

Because user mobility is one of the inherent features of massive wireless networks, the small range of microwave, Millimeter Wave (mmWave), or even Terahertz (THz) networks/cells may cause handovers and switching among edge servers to ensure fast and reliable operation. However, relying solely on frequent switching of edge servers and handovers is not efficient from a network standpoint, as it results in the wastage of many resources without any improvement in service delivery. Therefore, seamless and well-orchestrated integration should be incorporated in B5G systems, especially when considering high dynamic and dense deployment scenarios, power- and battery-constrained devices, and communication requirements, among others, to maintain optimum system state (i.e., optimal deployment of SDTs at the network edge).

To this end, the orchestration must incorporate Machine Learning (ML) and AI, such as DL-based user activity prediction, along with sensing and tracking technologies that will be employed to learn about the static and dynamic system components as presented in Fig. 2.20. High-level localization and sensing can be obtained from low-level raw measurements, e.g., Channel State Information (CSI), Received Signal Strength Indicator (RSSI), between peers via Device-to-Device (D2D) communications [130], and from local area networks, such as Wi-Fi. Moreover, D2D links can facilitate localization and sensing through cooperative positioning [131]. First, peers can exchange necessary data, such as common physical layer estimates and position information, to increase positioning accuracy. Alternatively, with the

implementation of D2D communications, peers inherently receive signals from one another that provide additional signal observations and may be used for location estimation.

The information about *SIoT object* location obtained through the *localization and sensing* techniques is further sent to the AI-based *learning component*. At this stage, *motion information predictions* for all devices are obtained for a given time interval. The outputs of the learning model might be time-, device-, space-, and scenario-dependent physical motion predictions, which are sent to ME hosts. Then, at the *communication component*, contexts shape the actions and form conditions for objects to comprehend their actions to allow data transmission prediction, which also depends on the object latency requirements. The *decision* on the SDT re-deployment is then taken based on the data transmission predictions for all SIoT devices over the time period for which motion has been predicted. This ensures an intelligent SDT re-deployment since the data is analyzed about all devices in the system and over a particular time, avoiding triggering redundant reallocations of SDTs. Intelligent storage and distributed processing capabilities enable the fusion of sensor data for detecting trends and deviations, helping make decisions on edge server switching and handover. At the final stage, *re-deployment is triggered and controlled*.

As a component of the social-aware orchestration (see Section 2.3.2), the localization, sensing, learning, communication, decision making, process, and control loop is anticipated to guarantee an optimum system state with reconfigurable re-optimization frequency and intelligent network resource consumption.

2.3.3 Performance Evaluation

Our analysis of IoT device/user motion reveals that mobility patterns differ among heterogeneous devices, locations, situations, and over time, holding diverse degrees of unpredictability. This can result in either an inability to meet user-specific requirements or inefficient use of network resources. To characterize the impact of IoT device motion on the user- and network-centric parameters and validate the novel orchestration for optimal SDT placement and efficient reallocation, we evaluate the performance in terms of service discovery delay and re-optimization frequency. The simulation campaign is based on Section 2.2. However, to make this section self-contained, we outline the scenario of interest and simulation parameters gathered in Table 2.6. We then offer selected numerical results in the following.

Simulation Settings

We study Santander (4 km x 4 km) city center and assume a 3GPP-compliant hexagonal grid with 8 BSs, each co-located with an edge server. We assess the concept using realistic object behavior [87] that records device interactions based on actual IoT objects and the SWIM mobility model. We assess settings with 100 users and 328 heterogeneous static and dynamic IoT devices. Each IoT device has an SDT implemented as a container at the edge. We associate SDTs with four kinds of containers according to CPU demands. Cars with autonomous

navigation and cellphones may need high-CPU SDTs, but smartwatches, sensors, tablets, smart fitness gadgets, and printers may demand small or micro instances. The maximum delay between a physical device and its SDT is within $1 - 10$ ms. The simulation covers a 5-hour time-lapse. Within this period, according to the proposed framework, re-optimization for the placement of SDTs triggers on-demand, whereas, for benchmark schemes, we set re-optimizations every 1, 5, 10, 15, 20, 25, and 30 minutes. We assess the service discovery latency for the device setting (see "all" devices) as well as specific categories of devices (see results for "smartphones", "cars", and "smartwatches").

Table 2.6: System parameters related to Section 2.3.

Area of interest	Santander, Spain (4 km x 4 km)
Number of Users	100
Number IoT Devices	328
Cell layout	3GPP hexagonal grid, 8 BSs with cell area radius: of 450 m and intersite distance of 1350 m
Edge servers	Co-located with BSs
CPU, DISK, and RAM capability	24000 MIPS, 2 TB, and 24 GB
Disk demands	Uniformly distributed in $[10, 50]$ GB
CPU and RAM demands	2000 MIPS, 0.85 GB; 2500 MIPS, 3.75 GB; 1000 MIPS, 1.7 GB; 500 MIPS, 613 MB
Latency requirement	Uniformly distributed in $[1, 10]$ ms

Performance Analysis

In Fig. 2.21, one may observe that the proposed architecture improves average service discovery time by up to 1 ms, which is critical for delay-sensitive applications. We note that average service discovery latency includes the delay associated with the connection between an IoT device and the edge server hosting the corresponding SDT and the delay needed to reach the SDTs of friend devices one by one. Given re-optimization statistics, the estimated overall latency reduction when employing the proposed method compared to average re-optimization values is 9%. Moreover, the difference between the median, which separates the higher half from the lower half of the data, and the proposed novel orchestration is up to 10.5%. Furthermore, the proposed solution demonstrated an improvement of 13% in latency compared to results obtained using various re-optimization intervals.

The overall re-optimization has been triggered 58 times utilizing the proposal (in comparison with 300, 60, 30, 20, 15, 12, and 10 times). The average number of migrated SDTs is 6.2, whereas every 1, 5, 10, 15, 20, 25, and 30 minute re-optimizations trigger on average 3.1, 5.4, 8.1, 9.8, 10.4, 13.5, and 11.8 migration events, respectively. The total number of migrated SDTs within 5-hour time-lapse corresponds to 360 when utilizing the novel orchestration and 930, 324, 243, 196, 156, 162, 118 for fixed optimization intervals.

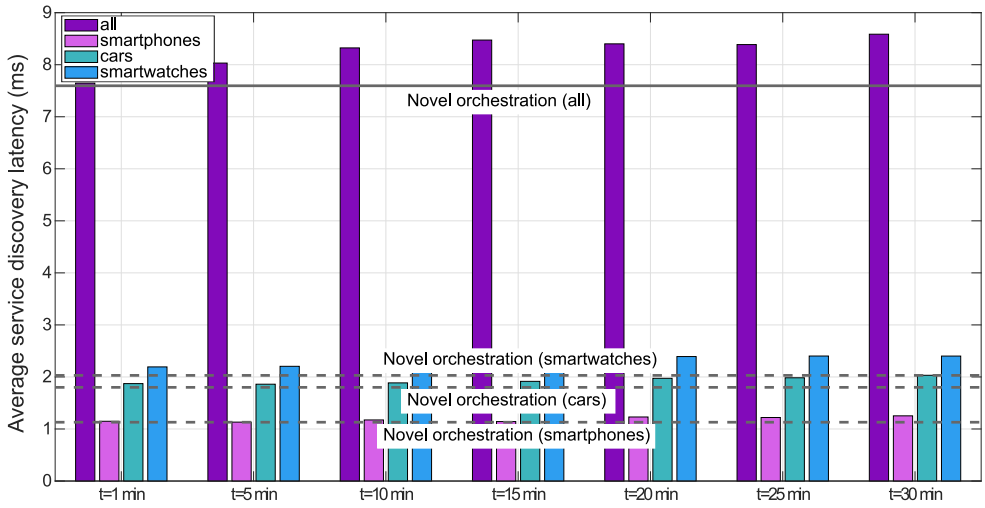

Fig. 2.21: Service discovery latency assessment [3].

As per our additional results, we extend the simulation to the 24-hour time-lapse and assess the re-optimization frequency for the proposed orchestration and benchmark solutions. For 113 IoT devices, ranging from stationary equipment to mobile consumer devices carried by mobile users in the region of interest (i.e., smartphones 6%, cars 15%, tablets 12%, smart fitness devices 20%, smartwatches 29%, personal computers 1%, printers 10%, and home sensors 7%), the reallocation of SDTs according to the novel orchestration has been triggered 159 times. This frequency is lower than the re-optimizations that occur every 1 and 5 minutes.

2.3.4 Discussions

The proliferation of advanced IoT devices and sensors is expected to bring unpredictability to device mobility, which might significantly affect system performance, demanding frequent network re-optimizations. On the contrary, repeated re-deployment of SDTs might be ineffective when the situation is almost static, such as during nighttime hours, resulting in wasted network resources. We investigated the required network re-optimization intervals to bridge this gap for effective SDT deployment at the network edge. We provided an overview of device motion patterns that may vary based on time, device type, location, and scenario. We then analyzed service discovery in SIoT systems and provided novel orchestration that consists of NFV, SDN, edge/fog and cloud computing, learning-based user activity prediction, and sensing and tracking technologies. Finally, we launched a simulation campaign on the re-optimization time interval comparing conventional system design with the proposed co-design of location, sensing, and AI-driven communication and computation. The results

confirmed the effectiveness of the proposed novel orchestration from both user and network perspectives.

We emphasize that the orchestration is intended to provide connectivity with extremely low end-to-end latency, high reliability, and adaptability to current or future networks. This includes immersive services such as XR, which is rapidly evolving towards likely mass adoption, bringing new challenges and technical problems. In the literature, XR systems are analyzed in steady-state operating conditions [132, 133]. However, due to recent findings in XR interaction freedom, state-of-the-art solutions may have limited applicability to practical implementations. The challenge is to address dynamicity and non-stationarity inherent to immersive reality behavior. In the next Chapter (Chapter 3), we address this gap and provide system designers with a tool for evaluating the XR system components considering both communication and computing perspectives.

3

Dynamic Behavior-Aware Interactions in Edge-Aided IoT

This chapter focuses on dynamic behavior-aware interactions in edge-aided IoT.

In Section 3.1, we focus on a performance evaluation methodology for XR services in edge-assisted networks. We offer a fluid approximation to characterize the XR content evolution in dynamic networks. The proposed approach captures the time and space dynamics of the content distribution process in its transient phase considering the computing plane.

In Section 3.2, we focus on a joint communication and computing performance evaluation methodology for XR services in edge-assisted wireless highly directional networks. The methodology is based on a fluid approximation and is particularly effective for the analysis of non-stationary processes with periodic arrival rates. We note that, in these two sections, we focus on the processes that tend to be in that transient phase and occur in the transition between various steady-state conditions.

In Section 3.3, we focus on assessment methodology for XR systems. We investigate the impacts of XR user mobility from the perspective of communication and computing. We provide a review of mobility patterns in XR and a comprehensive simulation study of the effect of interaction-dependent gait patterns on latency and resource utilization. The sources of evidence on the impact of user movements on network operations are then provided. We propose a case study for mobile XR that characterizes the system performance in terms of user mobility, communication, and processing.

3.1 Computing Performance Evaluation Methodology

XR, including AR/VR/MR, is unquestionably a widely discussed topic in the field of advanced audio-visual experiences [134, 135]. XR is a rendered representation of a supplied audio and visual scene designed to simulate real-world sensory inputs in a lifelike manner for the observer while a user moves within the constraints set by the program and the equipment [136].

XR experiences may be categorized into Three Degrees of Freedom (3DoF) XR services and Six Degrees of Freedom (6DoF) XR services based on the perceived experience. 3DoF supports rotational user movement along the x, y, and z- axes, where $(0, 0, 0)$ represents the center of the user's head, enabling the user to look around from a single fixed viewing point. 6DoF allows for movement and rotation within a 3D environment, enabling the user to freely navigate an XR scene [137].

According to [138, 139], there are the following XR device technological specification-based phases: fair-experience and comfortable-experience phases. Combined with versatile video coding and light field rendering, the development of content and terminals yields an ideal XR experience. The full-view transmission solution requires high network capacity to support XR video services. Each phase involves advancements in XR device technology, primarily in terms of hardware development. We can determine the service requirements for each phase based on these specifications (see Table 3.1).

As the upper bound for actual XR requirements, i.e., data rate and communication latency, we can relate to the limitations of human vision. Human eyes are capable of seeing dots as small as 0.3 arc-minutes per degree, which can be translated to around 200 unique dots per degree. Human eyes can mechanically shift across 150 degrees horizontally and 90 degrees vertically, requiring a region of 540 million pixels for full view. Adding the ability to turn and rotate the body, the visual field can be expanded to 360 degrees horizontally and approximately 270 degrees vertically. It would require a region of 3.888 billion pixels for full view.

In this case, a static image requires up to 540 million/3.888 billion pixels. Multiple static pictures are flashed in series for motion video. The human eye is capable of sensing motion at a much faster rate, with some estimates reaching up to 200 frames per second. To avoid motion blur and confusion, high-speed immersive experiences require at least 60 frames per second and, in some cases, up to 120 frames per second. Moreover, other characteristics of the human eye exceed current display technologies. The human eye is capable of perceiving a contrast ratio of up to nearly 1 million brightness levels, requiring up to 8 bytes per pixel to fully encode the perceptible color gamut for each screen.

Therefore, the upper limit corresponds to 15.2 Terabytes of data per second with 540 million pixels at 8 bytes per pixel at 120 frames per second. However, no digital system or network in the foreseeable future can handle that kind of throughput. Fortunately, there is significant redundancy in visual data that allows a great deal of compression depending on

Table 3.1: XR network requirements.

Requirement	**Fair Experience**	**Comfortable Experience**	**Ideal Experience**	**Human Visual Perception**
Commercial application time	2018	2019-2020	2023-2025	-
Content full-view resolution	4K	8K	12K 24K	-
Video full-view resolution	(1080x1200x2)	(1920x1920x2)	(3840x3840x2) (7680x7680x2)	(150x200x90x200) (360x200x270x200)
Field-of-View	90°-110°	120°	120°-140°	-
Color depth	8	8	10-12	8
Coding Standard	H.264/H.265	H.265	H.265/H.266	-
Frame rate	50-90	90	120-200	120
Bit Rate	5.6 Gbps	15.93 Gbps	212.34 Gbps (12K) 849.35 Gbps (24K)	15.2 Tbps 11.1 Tbps
Bit Rate (20:1)	0.28 Gbps	0.8 Gbps	10.62 Gbps (12K) 42.47 Gbps (24K)	77.76 Gbps 559.9 Gbps
Bit Rate (300:1)	0.02 Gbps	0.05 Gbps	0.71 Gbps (12K) 2.83 Gbps (24K)	5.18 Gbps 37.32 Gbps
Network RTT	20 ms	15 ms	8 ms	8 ms
Packet loss rate	10^{-5}	10^{-5}	10^{-6}	10^{-6}

*Bit rate = 3 x Color depth x Full-view x Frame Rate / Compression ratio

the complexity of the images. Even with a high compression ratio of 300 : 1, which would require powerful computers to encode and decode the compressed video, the data rate would still be 5.18 GB per second. Current commercial 3D displays are not capable of providing such high resolution.

However, XR is rapidly evolving toward likely mass adoption. To address *dynamicity* and *non-stationarity* inherent to immersive reality behavior and assist network planning engineers with a means to evaluate XR system performance, we offer the following contributions:

- we offer a practical methodology to characterize XR content evolution in dynamic networks as a continuous fluid considering the computing plane;
- we evaluate the performance of the proposed methodology based on a fluid approximation;
- we assess the XR system characteristics under different network settings;
- we offer practical conclusions for designing XR networks considering computing plane.

The rest of this section is structured as follows. The motivation behind the investigated topic is presented in Subsection 3.1.1. The system model is outlined in Subsection 3.1.2, whereas the proposed methodology is characterized in Subsection 3.1.3. In Subsection 3.1.4, an evaluation campaign is offered. Discussions are drawn in Subsection 3.1.5.

3.1.1 Motivation

In the literature, XR systems are usually analyzed in steady-state operating conditions [132, 133]. However, due to the recent findings in the context of XR interaction [5], state-of-the-

art modeling solutions may have limited applicability to practical implementations. The challenge is, thus, to address the modeling of *dynamicity* and *non-stationarity* [140] aspects inherent to operational behavior of modern immersive reality systems under conditions of periodic arrival processes [141] to provide network planning and optimization engineers with an effective means to evaluate XR system performance.

To examine time and space dynamics and capture the system effects in its transient phase, the actual discrete number of users interested in the contents may be substituted with the equivalent continuous fluid. Specifically, in [142], the fluid approximation is used to model the evolution of calling/noncalling vehicles along a highway, whereas in [143] and [144], the number of users retrieving the content and those who already have received the content are considered as fluids. In [145], the volume of traffic in wireless sensor networks is modeled using a macroscopic fluid dynamic model.

However, naturally, XR has a unique set of characteristics, i.e., novel form factors that pose stringent requirements on the power consumption and heat dissipation of user devices (since devices are worn on the body). Therefore, computations cannot be executed on the XR HMDs or glasses, thus requiring task offloading to an edge server [146, 147]. To this end, the *computing* plane must be considered when analyzing XR content evolution. In this section, we close this gap and address *dynamicity* and *non-stationarity* of XR systems.

We offer a practical methodology to characterize XR content evolution in dynamic networks as a continuous fluid. Specifically, we model the non-static XR behavior with a departure rate that strictly depends on the computing characteristics of edge nodes. We then apply our methodology to assess the system characteristics under different network settings.

3.1.2 System Model

This section outlines the reference scenario and the system model assumptions on user dynamics and content distribution.

We consider an outdoor environment where multiple users engage in an XR interactive experience through HMDs. As XR devices have limited computing performance due to the constraints on size, power consumption, and heat dissipation, we assume that an HMD acts as a thin client and receives the personalized video stream from a proximate edge computing server.

Assumption 1. User State: A user is associated with one of the two states: *(i) active in the processing (computing) phase* if awaiting and downloading the generated XR 360 video from the server, or *(ii) idle*, otherwise.

Assumption 2. User Motion: We consider a one-way motion around the semi-infinite pedestrian zone. For example, users may move from one site to another or follow a straight route along the street. We note that the studied one-way formulation can easily be generalized to the case of two-way traffic. The pedestrian zone is divided into K sub-zones (as

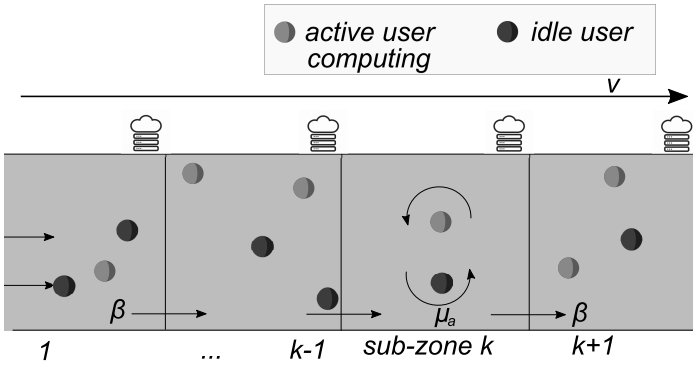

Fig. 3.1: System illustration.

illustrated in Fig. 3.1), i.e., $k \in \mathcal{K} = \{1, ..., K\}$. Both active and idle users move forward within pedestrian zone k with a constant speed v.

Assumption 3. System Dynamics: We consider the following source of arrivals of active and idle users into the system:

(i) zone 1: external *active communicating/idle* users with specified deterministic intensity λ;
(ii) zone k: *active computing / idle* users from zone $k-1$ arriving with the constant rate β;
(iii) zone k: transition from *idle* to *active computing* due to the initiated transmission with the constant rate μ_a;
(iv) zone k: transition from *active computing* to *idle* due to the completed content processing at the edge node.

Active users become idle only after processing the required content at the edge, i.e., after passing through the processing (computing) state. In turn, idle users switch to the active communicating state based on the content demand, and thus, the rate μ_a is constant.

Assumption 4. Processing (Computing) State: The BS/AP that serves zone k is associated with an edge node, where multiple Virtual Machines (VMs) are utilized to implement parallel processing of the content. To consider the input/output interference among VMs of the same node, we introduce the degradation factor d [148], so that the individual share of available resources in the case of $N_k^{a_c}(t)$ users in the zone k is estimated as

$$\left[N_k^{a_c}(t)(1+d)^{(N_k^{a_c}(t)-1)}\right]^{-1}. \tag{3.1}$$

If B_c is the constant size of the downloaded content and R_0^c is the total capacity of elaboration that the edge server has, then the actual transition rate from their state of *active user in the processing phase* (i.e., the processing of their content is ongoing) to that of *idle* (i.e., the processing of their content has been completed), can be determined as

$$C^c(t) = \frac{1}{N_k^{a_c}(t)(1+d)^{(N_k^{a_c}(t)-1)}} \cdot C_0^c, \tag{3.2}$$

where $C_0^c = R_0^c/B_c$.

For the sake of analytical tractability, we approximate (3.2) as

$$C^c(t) = \frac{1}{N_k^{a_c}(t)(N_k^{a_c}(t)+\alpha)} \cdot C_0^c, \tag{3.3}$$

where α is a fitting parameter.

3.1.3 Mathematical Characterization

In this subsection, we outline the proposed methodology to assess the performance of a highly dynamic XR system. We employ the method of fluid approximation, which allows replacing integer-valued processes with deterministic real-valued ones and is particularly suited for the analysis of non-stationary processes.

We denote the total number of idle and active users in the processing states in the zone k at time t as $N_k^d(t)$ and $N_k^{a_c}(t)$, respectively, and consider them as non-negative real numbers.

Furthermore, let $C_k^{a_c+}(t)/C_k^{a_c-}(t)$ denote the number of active computing users that arrive to/depart from zone k during time interval $(0,t]$. Similarly, $C_k^{d+}(t)/C_k^{d-}(t)$ represent the respective number of idle users arriving to/departing from zone k during time $(0,t]$. We note that for the sake of clarity, we further omit index (t).

The evolution of active and idle users in zone k is governed by a system of coupled Ordinary Differential Equations (ODEs) for $x_k \geq 0$, $0 < t < \infty, k \in \mathcal{K}$, as follows:

$$\begin{cases} \frac{dN_k^{a_c}}{dt} \equiv C_k^{a_c+} - C_k^{a_c-}, \\ \frac{dN_k^{d}}{dt} \equiv C_k^{d+} - C_k^{d-}, \end{cases} \tag{3.4}$$

where

$$C_1^{a_c+}(t) = \lambda + \mu_a N_1^d(t) \text{ and } C_1^{a_c-}(t) = \frac{C_0^c}{(N_1^{a_c}(t)+\alpha)} + \beta N_1^{a_c}(t),$$
$$C_1^{d+}(t) = \lambda + \frac{C_0^c}{(N_1^{a_c}(t)+\alpha)} \text{ and } C_1^{d-}(t) = \mu_a N_1^d(t) + \beta N_1^d(t),$$
$$C_k^{a_c+}(t) = \mu_a N_k^d(t) + \beta N_{k-1}^{a_c}(t) \text{ and } C_k^{a_c-}(t) = \frac{C_0^c}{(N_k^{a_c}(t)+\alpha)} + \beta N_k^{a_c}(t),$$
$$C_k^{d+}(t) = \frac{C_0^c}{(N_k^{a_c}(t)+\alpha)} + \beta N_{k-1}^{d}(t) \text{ and } C_k^{d-}(t) = \mu_a N_k^d(t) + \beta N_k^d(t).$$

We assume that the content requests start at time instant $t = 0$. The number of active users during the computing phase and idle users, $N_k^{a_c}$ and N_k^d, can be obtained by solving the Cauchy problem (3.5):

$$\begin{cases} dN_1^{a_c}/dt = \lambda + \mu_a N_1^d - \frac{C_0^c}{(N_1^{a_c}+\alpha)} - \beta N_1^{a_c}, \\ dN_1^{d}/dt = \lambda + \frac{C_0^c}{(N_1^{a_c}+\alpha)} - \mu_a N_1^d - \beta N_1^d, \\ dN_k^{a_c}/dt = \mu_a N_k^d + \beta N_{k-1}^{a_c} - \frac{C_0^c}{(N_k^{a_c}+\alpha)} - \beta N_k^{a_c}, \\ dN_k^{d}/dt = \frac{C_0^c}{(N_k^{a_c}+\alpha)} + \beta N_{k-1}^{d} - \mu_a N_k^d - \beta N_k^d, \\ \text{under initial conditions: } N_1^{a_c}\big|_{t=0} = 0,\ N_1^d\big|_{t=0} = M_1,\ N_k^{a_c}\big|_{t=0} = 0,\ N_k^d\big|_{t=0} = M_k. \end{cases} \tag{3.5}$$

We address the system (3.5) by substituting $N_k = N_k^{a_t} + N_k^{a_c} + N_k^d$ and obtain the following:

$$\begin{cases} dN_1/dt = 2\lambda - \beta N_1, \\ dN_k/dt = \beta N_{k-1} - \beta N_k. \end{cases} \tag{3.6}$$

By solving the first differential equation of (3.6), we obtain the number of users in zone 1 as

$$N_1 = C_1 e^{-\beta t} + \tfrac{2\lambda}{\beta}, \tag{3.7}$$

where $C_1 = \frac{2\lambda}{\beta}$.

By induction, we further obtain N_k:

$$N_k = \sum_{n=1}^{k} \frac{(\beta t)^{k-n} e^{-\beta t}}{k-n!} C_n + \frac{2\lambda}{\beta}. \tag{3.8}$$

$N_k^{a_c}(t)$ and $N_k^d(t)$ can be found by using the following expressions:

$$N_k^{a_c}(t) = C_{ac_k} e^{-\beta t} + \frac{2\lambda}{\beta}, \tag{3.9}$$

$$N_k^d(t) = C_k e^{-\beta t} - C_{ac_k} e^{-\beta t}, \tag{3.10}$$

where C_{ac_k} can be obtained by solving (3.5) for any zone k by substituting (3.9)-(3.10). This technical task, however, is out of the scope of this work.

In summary, in this section, we offered a practical methodology based on a fluid approximation to characterize the non-static process of the XR content evolution. We derived (3.8) to express the number of users $N_k(t)$, $\forall k \in (0, \infty)$, $t \in (0, \infty)$, whereas the number of active users during the processing phase and the number of idle ones can be obtained from (3.9)-(3.10).

3.1.4 Performance Evaluation

This section compiles selected numerical results of the performance evaluation of a dynamic XR system. The simulation scenario is modeled after a large social/public XR event, e.g., a concert or an outdoor exhibition, where user devices offload extensive computations to the edge. We study the time and space dynamics of the XR content distribution process captured by the deterministic fluid model (A) and Monte Carlo simulations (S). The simulation data agree with analytical results for all considered metrics of interest.

Simulation Settings

In simulations, the transition rate is defined as per (3.2), whereas when applying the analytical model, we employ the approximation of (3.2) provided in (3.3) with $\alpha = 0.12$. We examine the XR system performance in terms of the number of users for two scenarios respectively characterized by *(i)* $B_t = 100\,\text{Mb}$, $B_c = 150\,\text{Mb}$, $R_0^t = 196\,\text{Mbps}$, $R_0^c = 150\,\text{Mbps}$ and *(ii)* $B_t = 160\,\text{Mb}$, $B_c = 350\,\text{Mb}$, $R_0^t = 150\,\text{Mbps}$, $R_0^c = 150\,\text{Mbps}$ [149, 150]. Table 3.2 summarizes the main parameters [149, 151–154].

Table 3.2: Simulation parameters related to Section 3.1.

Parameter	**Value**
Area	200 m x 50 m [151]
Carrier frequency, f_c	28 GHz [152]
Number of BSs	1 BS per zone [153]
Number of edge servers	1 server per zone [153]
User velocity	1.55 m/s [154]
Downloaded content size, B_c	150, 350 Mb [149]
Edge node total capacity of elaboration, R_0^c	150, 150 Mbps [149]
Fitting parameter, α	0.12
Transformation rate, μ_a	0.6^{-1} 1/s
Arrival rate, λ	0.5^{-1} 1/s

Performance Analysis

In Fig. 3.2 and Fig. 3.3, we observe that the number of active users involved in the processing phase increases. As a result, the processing speed decreases, which causes rising delays. One may also observe that at time instant $t = 221$s, users start to arrive at sub-zone 2, leading to a gradual system unloading and a decrease in the number of users involved in the processing phase in sub-zone 1.

We then consider the case of relatively heavy edge processing loads (please refer to Fig. 3.4 and Fig. 3.5). Differently from the previous setting, the number of active users drastically increases. As a result, the bottleneck occurs during the processing phase.

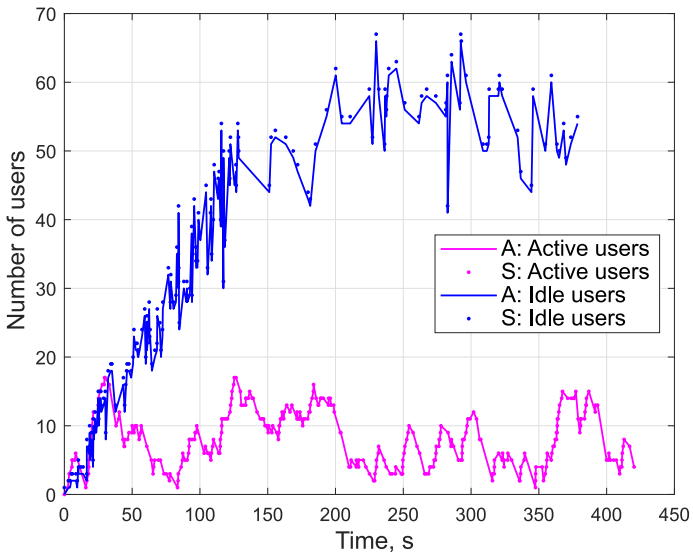

Fig. 3.2: Performance assessment (number of users, zone 1): $B_c = 150$ Mb, $R_0^c = 150$ Mbps.

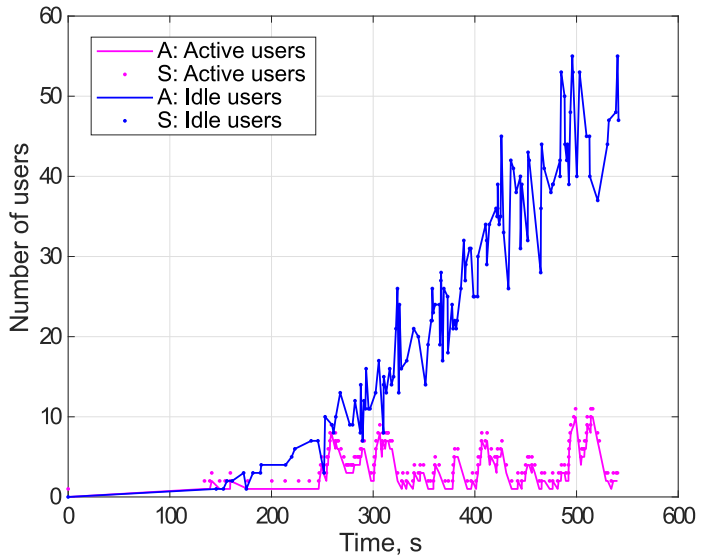

Fig. 3.3: Performance assessment (number of users, zone 2): $B_c = 150\,\text{Mb}$, $R_0^c = 150\,\text{Mbps}$.

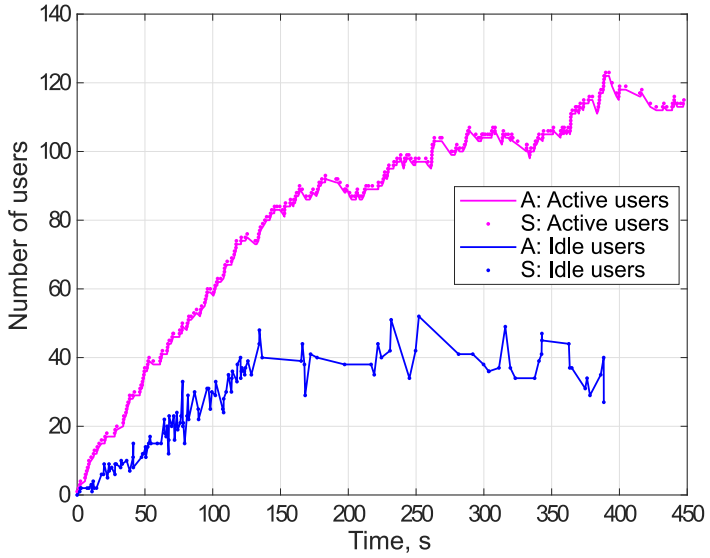

Fig. 3.4: Performance assessment (number of users, zone 1): $B_c = 350\,\text{Mb}$, $R_0^c = 150\,\text{Mbps}$.

The main finding is that *the computation procedure for processing XR video appears to be very demanding in terms of computational resources. System designers may recourse to both parallel and distributed computing to reduce the computation time for video processing at the edge server.*

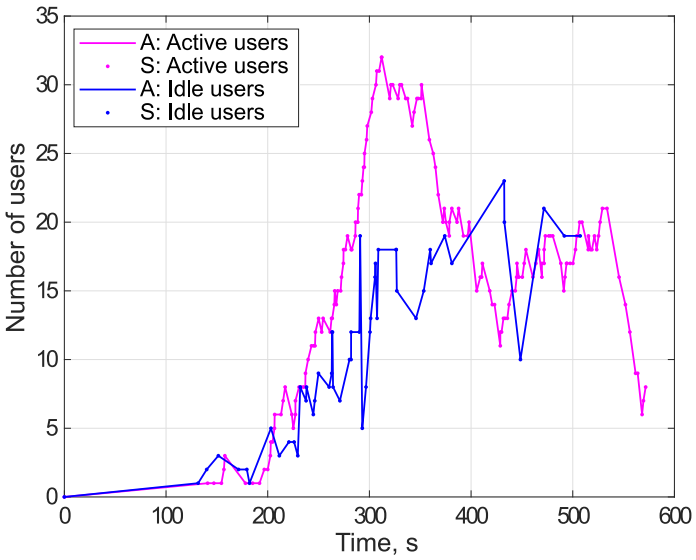

Fig. 3.5: Performance assessment (number of users, zone 2): $B_c = 350\,\text{Mb}$, $R_0^c = 150\,\text{Mbps}$.

In this section, we assessed the effectiveness of the fluid approximation that captures the time and space dynamics of the XR content distribution process in its transient phase considering the computing plane. We evaluated the XR system performance for different system settings. We are confident that our proposed methodology and practical conclusions have the potential to drive system designers to design computing immersive experience systems effectively.

3.1.5 Discussions

In this section, we provided a methodology for characterizing the evolution of immersive reality content in dynamic networks. The methodology considers the time and spatial dynamics of the content distribution process during the computing plane. *However, for designing XR communication networks, it is vital to consider the communications component together with the computing one. To this end, in the next section (Section 3.2), we focus on the content distribution for XR services in edge-assisted wireless highly directional networks.*

3.2 Joint Communication and Computing Performance Evaluation Methodology

B5G wireless networks are expected to substantially differ from the current systems in terms of applications and services, interactions, and even devices, especially when comparing immersion in different realities. As the consumer interest in various immersive applications grows explosively, the XR technology rapidly evolves, offering more affordable, compact, and powerful hardware coupled with rapid developments in software and connectivity. Recently, XR freedom of mobility and interaction has been unlocked by introducing a 60GHz HTC VIVE Wireless Adapter, which allows removing cables in a one-room environment, and later, a 5GHz Oculus Air Link that utilizes available Wi-Fi connectivity [155], thereby shifting to a new era of XR experience with the complete freedom of interaction.

As we mentioned in the previous section, by design, XR has a unique set of features, i.e., new form factors leading to strict requirements on user equipment power consumption and heat dissipation. Therefore, computations (i.e., content elaborations) cannot be executed on the XR HMDs or glasses, thus requiring task offloading to an edge server over a wireless network. To this end, both *communication* and *computing* planes must be considered when analyzing XR content evolution. In this section, we close this gap and address *dynamicity* and *non-stationarity* of XR systems as follows:

- we offer a practical methodology based on a fluid approximation to characterize the XR content evolution in dynamic wireless networks by capturing the time and space dynamics of the content distribution process considering both communication and computing planes;
- we evaluate the performance of the proposed methodology based on a fluid approximation;
- we assess the XR system characteristics under different network settings;
- we provide practical conclusions for designing XR wireless communication networks considering both communication and computing planes.

The rest of this section is organized as follows. In Subsection 3.2.1, we offer a practical methodology to characterize XR content evolution in dynamic wireless networks as a continuous fluid. Specifically, we model the non-static XR behavior with a periodic arrival process and assume a departure rate that strictly depends on the communication characteristics of the 5G New Radio (NR) access technology. Simulation results are reported in Subsection 3.2.2. The main conclusions of the study are summarized in Subsection 3.2.3.

3.2.1 Mathematical Characterization

In addition to the assumptions on the system model introduced in Subsection 3.1.2, in this section, we consider both communication and computing planes.

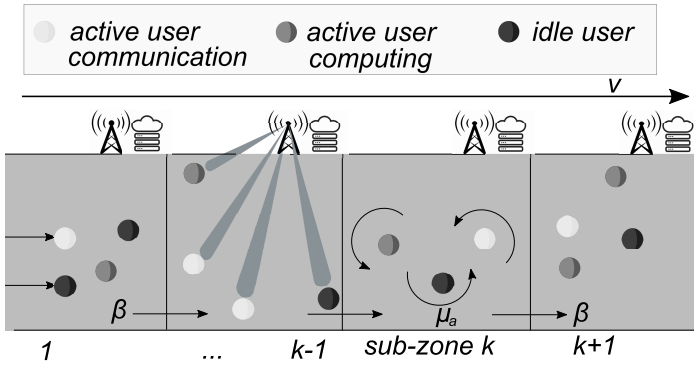

Fig. 3.6: Joint communication and computing system [4].

As XR devices have limited computing performance due to the constraints on size, power consumption, and heat dissipation, we assume that an HMD acts as a thin client and receives the personalized video stream from a proximate edge computing server over wireless (e.g., mmWave) connection.

A user is associated with one of the three states: (i) *active in the communication phase* while sending the location and motion information to the server, (ii) *active in the processing* phase if awaiting and downloading the generated XR 360 video from the server, or (iii) *idle*, otherwise (see Fig. 3.6).

We consider the following source of arrivals of active and idle users into the system:

(i) zone 1: external *active communicating*/*idle* users;
(ii) zone k: *active communicating* / *active computing* / *idle* users from zone $k-1$ arriving with the constant rate β;
(iii) zone k: transition from *idle* to *active communication* due to the initiated transmission with the constant rate μ_a;
(iv) zone k: transition from *active communication* to *active computing* due to completed transmission;
(v) zone k: transition from *active computing* to *idle* due to the completed content processing at the edge node.

We assume a periodic external arrival rate with phase shift $w_0^{a/d}$ and angular frequency $w^{a/d}$ for active/idle users. We limit the maximum and minimum values of the external arrival rate by $\lambda_1^{a/d}$ and $\lambda_2^{a/d}$, respectively, and, hence, the external arrival rate for actively communicating and idle users at time t may be expressed as:

$$\lambda^a(t) = \sin(w^a t + w_o^a)(\lambda_1^a - \lambda_2^a)/2 + (\lambda_1^a + \lambda_2^a)/2, \tag{3.11}$$

$$\lambda^d(t) = \sin(w^d t + w_o^d)(\lambda_1^d - \lambda_2^d)/2 + (\lambda_1^d + \lambda_2^d)/2. \tag{3.12}$$

Active users become idle only after completing the uplink transmission and processing the required content at the edge, i.e., after passing through both communication (transmission)

and processing (computing) states. In turn, idle users switch to the active communicating state based on the content demand, and thus, the rate μ_a is constant.

Each zone k is served by a BS or AP that provides wireless connectivity to the users within coverage. We assume resources at BS/AP, which are equally shared in time and/or frequency among $N_k^{a_t}(t)$ active users.

Furthermore, we assume that B_t is the constant uplink packet size and R_0^t is the total uplink capacity. In the case of $N_k^{a_t}(t)$ participants, we may estimate the transition rate of users from their status of *active in the communication phase* (i.e., transmitting) to that of *active in the processing phase* (i.e., the processing of their content is ongoing), as

$$C^t(t) = \frac{1}{N_k^{a_t}(t)} \cdot C_0^t, \tag{3.13}$$

where $C_0^t = R_0^t/B_t$, whereas B_t is the constant size of the uplink packet.

We employ the method of fluid approximation, which allows replacing integer-valued processes with deterministic real-valued ones and is particularly suited for the analysis of non-stationary processes.

We denote the total number of idle and active users in the communication and processing states in the zone k at time t as $N_k^d(t)$, $N_k^{a_t}(t)$, and $N_k^{a_c}(t)$, respectively, and consider them as non-negative real numbers.

Furthermore, let ${C_k^{a_t}}^+(t)/{C_k^{a_t}}^-(t)$ and ${C_k^{a_c}}^+(t)/{C_k^{a_c}}^-(t)$ denote the number of active communicating and computing users that arrive to/depart from zone k during time interval $(0, t]$. Similarly, $C_k^{d+}(t)/C_k^{d-}(t)$ represent the respective number of idle users arriving to/departing from zone k during time $(0, t]$. We note that for the sake of clarity, we further omit index (t).

Table 3.3: Expressions for arriving/departing flows in (3.14).

$C_1^{a_t+} = \lambda^a + \mu_a N_1^d$	$C_1^{a_t-} = C_0^t + \beta N_1^{a_t}$
$C_1^{a_c+} = C_0^t$	$C_1^{a_c-} = \frac{C_0^c}{(N_1^{a_c}+\alpha)} + \beta N_1^{a_c}$
$C_1^{d+} = \lambda^d + \frac{C_0^c}{(N_1^{a_c}+\alpha)}$	$C_1^{d-} = \mu_a N_1^d + \beta N_1^d$
$C_k^{a_t+} = \mu_a N_k^d + \beta N_{k-1}^{a_t}$	$C_k^{a_t-} = C_0^t + \beta N_k^{a_t}$
$C_k^{a_c+} = C_0^t + \beta N_{k-1}^{a_c}$	$C_k^{a_c-} = \frac{C_0^c}{(N_k^{a_c}+\alpha)} - \beta N_k^{a_c}$
$C_k^{d+} = \frac{C_0^c}{(N_k^{a_c}+\alpha)} + \beta N_{k-1}^d$	$C_k^{d-} = \mu_a N_k^d + \beta N_1^d$
$\lambda^a = \sin(w^a t + w_o^a)\frac{\lambda_1^a - \lambda_2^a}{2} + \frac{\lambda_1^a + \lambda_2^a}{2}$	
$\lambda^d = \sin(w^d t + w_o^d)\frac{\lambda_1^d - \lambda_2^d}{2} + \frac{\lambda_1^d + \lambda_2^d}{2}$	

The evolution of active and idle users in zone k is governed by a system of coupled ODEs for $x_k \geq 0$, $0 < t < \infty, k \in \mathcal{K}$, as follows:

$$\begin{cases} \frac{dN_k^{a_t}}{dt} \equiv C_k^{a_t+} - C_k^{a_t-}, \\ \frac{dN_k^{a_c}}{dt} \equiv C_k^{a_c+} - C_k^{a_c-}, \\ \frac{dN_k^{d}}{dt} \equiv C_k^{d+} - C_k^{d-}, \end{cases} \tag{3.14}$$

where variables $C_k^{d/a\,+/-}$ are gathered in Table 3.3.

We assume that the content requests start at time instant $t = 0$. The number of active users during communication and computing phases and idle users, $N_k^{a_t}$, $N_k^{a_c}$, and N_k^d can be obtained by solving the Cauchy problem (3.15):

$$\begin{cases} dN_1^{a_t}/dt = (\sin(w^a t + w_o^a)\frac{\lambda_1^a - \lambda_2^a}{2} + \frac{\lambda_1^a + \lambda_2^a}{2}) + \mu_a N_1^d - C_0^t - \beta N_1^{a_t}, \\ dN_1^{a_c}/dt = C_0^t - \frac{C_0^c}{(N_1^{a_c}+\alpha)} - \beta N_1^{a_c}, \\ dN_1^{d}/dt = (\sin(w^d t + w_o^d)\frac{\lambda_1^d - \lambda_2^d}{2} + \frac{\lambda_1^d + \lambda_2^d}{2}) + \frac{C_0^c}{(N_1^{a_c}+\alpha)} - \mu_a N_1^d - \beta N_1^d, \\ dN_k^{a_t}/dt = \mu_a N_k^d + \beta N_{k-1}^{a_t} - C_0^t - \beta N_k^{a_t}, \\ dN_k^{a_c}/dt = C_0^t + \beta N_{k-1}^{a_c} - \frac{C_0^c}{(N_k^{a_c}+\alpha)} - \beta N_k^{a_c}, \\ dN_k^{d}/dt = \frac{C_0^c}{(N_k^{a_c}+\alpha)} + \beta N_{k-1}^{d} - \mu_a N_k^d - \beta N_k^d, \\ \text{under initial conditions: } N_1^{a_t}|_{t=0} = 0,\ N_1^{a_c}|_{t=0} = 0,\ N_1^d\big|_{t=0} = M_1, \\ N_k^{a_t}|_{t=0} = 0,\ N_k^{a_c}|_{t=0} = 0,\ N_k^d\big|_{t=0} = M_k. \end{cases} \tag{3.15}$$

We address the system (3.15) by substituting $N_k = N_k^{a_t} + N_k^{a_c} + N_k^d$ and obtain the following:

$$\begin{cases} dN_1/dt = \sin(w^a t + w_o^a)\frac{\lambda_1^a - \lambda_2^a}{2} + \frac{\lambda_1^a + \lambda_2^a}{2} \\ \qquad\quad + \sin(w^d t + w_o^d)\frac{\lambda_1^d - \lambda_2^d}{2} + \frac{\lambda_1^d + \lambda_2^d}{2} - \beta N_1, \\ dN_k/dt = \beta N_{k-1} - \beta N_k. \end{cases} \tag{3.16}$$

By solving the first differential equation of (3.16), we obtain the number of users in zone 1 as

$$N_1 = C_1 e^{-\beta t} + H, \tag{3.17}$$

where constant C_1 is determined by initial conditions; H and C_1 are given by expressions (3.18) and (3.19), correspondingly.

$$H = \frac{\lambda_1^a + \lambda_2^a + \lambda_1^d + \lambda_2^d}{2\beta} + \frac{(\lambda_1^a - \lambda_2^a)(\beta \sin(w^a t + w_0^a) - w^a \cos(w^a t + w_0^a))}{2(\beta^2 + w^{a^2})} + \frac{(\lambda_1^d - \lambda_2^d)(\beta \sin(w^d t + w_0^d) - w^d \cos(w^d t + w_0^d))}{2(\beta^2 + w^{d^2})}, \tag{3.18}$$

$$C_1 = M_1 - \frac{\lambda_1^a + \lambda_2^a + \lambda_1^d + \lambda_2^d}{2\beta} - \frac{(\lambda_1^a - \lambda_2^a)(\beta \sin(w_0^a) - w^a \cos(w_0^a))}{2(\beta^2 + w^{a^2})} - \frac{(\lambda_1^d - \lambda_2^d)(\beta \sin(w_0^d) - w^d \cos(w_0^d))}{2(\beta^2 + w^{d^2})}. \tag{3.19}$$

Furthermore, using (3.17), we may obtain the total number N_k for $k = i + 4m$, where $i = 1, 2, 3, 4$ and $m \geq 0$ as follows:

$$\begin{cases} N_{k=1+4m}= Z_0 + \beta Z_1^a - w^a Z_2^a + \beta Z_1^d - w^a Z_2^d, \\ N_{k=2+4m}= Z_0 - \beta Z_2^a - w^a Z_1^a - \beta Z_2^d - w^d Z_1^d, \\ N_{k=3+4m}= Z_0 - \beta Z_1^a + w^a Z_2^a - \beta Z_1^d + w^d Z_2^d, \\ N_{k=4+4m}= Z_0 + \beta Z_2^a + w^a Z_1^a + \beta Z_2^d + w^d Z_1^d, \end{cases} \tag{3.20}$$

where parameters $Z_0, Z_{1/2}^a, Z_{1/2}^d$ are given by

$$\begin{cases} Z_0 = \sum_{n=1}^{k} \frac{(\beta t)^{k-n} e^{-\beta t}}{(k-n)!} C_n + \frac{\lambda_1^a + \lambda_2^a + \lambda_1^d + \lambda_2^d}{2\beta}, \\ Z_1^a = \frac{(\lambda_1^a - \lambda_2^a)\sin(w^a t + w_0^a)}{2w^{a(k-1)}(\beta^2 + w^{a2})}, Z_2^a = \frac{(\lambda_1^a - \lambda_2^a)\cos(w^a t + w_0^a)}{2w^{a(k-1)}(\beta^2 + w^{a2})}, \\ Z_1^d = \frac{(\lambda_1^d - \lambda_2^d)\sin(w^d t + w_0^d)}{2w^{d(k-1)}(\beta^2 + w^{d2})}, Z_2^d = \frac{(\lambda_1^d - \lambda_2^d)\cos(w^d t + w_0^d)}{2w^{d(k-1)}(\beta^2 + w^{d2})}. \end{cases} \tag{3.21}$$

Knowing the expressions for the total number of users, N_k, we may obtain $N_k^{a_t}(t), N_k^{a_c}(t)$, and $N_k^d(t)$ separately as

$$N_k^{a_t} = C_{at_k} e^{-\beta t} + H, \tag{3.22}$$

$$N_k^{a_c} = C_{ac_k} e^{-\beta t} + H, \tag{3.23}$$

$$N_k^d = C_k e^{-\beta t} - C_{at_k} e^{-\beta t} - C_{ac_k} e^{-\beta t} - H, \tag{3.24}$$

where C_{at_k} and C_{ac_k} can be obtained for any zone k by substituting (3.22)-(3.24) into the initial conditions of (3.15). This technical task, however, is out of the scope of this work.

In summary, in this section, we offered a fluid approximation method to characterize the non-stationary process of the XR content evolution with periodic user arrivals. We derived (3.20) to express the total number of users, whereas the number of idle and active users in the communication and processing states can be obtained by (3.22)-(3.24).

3.2.2 Performance Evaluation

This subsection compiles the results of a performance evaluation campaign of the XR system. The simulation scenario is modeled after a large social/public event, such as a concert hall or an outdoor set of exhibitions, when large crowds move from one XR immersive site to another. We provide numerical results for the time and space dynamics captured by the deterministic fluid model (A) and obtained through Monte Carlo (S) simulations. The simulation data agree with analytical results for all considered metrics of interest.

Simulation Settings

During simulations, the transition rate is defined as per (3.2), whereas when applying the analytical model, we employ the approximation of (3.2) provided in (3.3) with $\alpha = 0.12$. We examine the XR system performance in terms of the number of users, transmission rate[1]/processing speed[2], and latency for two scenarios respectively characterized by *(i)* $B_t = 100\,\text{Mb}$, $B_c = 150\,\text{Mb}$, $R_0^t = 196\,\text{Mbps}$, $R_0^c = 150\,\text{Mbps}$ and *(ii)* $B_t = 160\,\text{Mb}$,

[1] Measured as $R_0^t/(N_k^{a_t}(t))$.
[2] Measured as $R_0^c/(N_k^{a_c}(t)(N_k^{a_c}(t) + \alpha))$.

$B_c = 350\,\text{Mb}$, $R_0^t = 150\,\text{Mbps}$, $R_0^c = 150\,\text{Mbps}$ [149, 150]. Table 3.4 summarizes the major parameters [149, 151–154].

Table 3.4: Simulation parameters related to Section 3.2.

Parameter	**Value**
Area	200 m x 50 m [151]
Carrier frequency, f_c	28 GHz [152]
Number of BSs	1 BS per zone [153]
Number of edge servers	1 server per zone [153]
User velocity	1.55 m/s [154]
Uplink packet size, B_t	100, 160 Mb [149]
Downloaded content size, B_c	150, 350 Mb [149]
Total uplink capacity, R_0^t	196, 150 Mbps [150]
Edge node total capacity of elaboration, R_0^c	150, 150 Mbps [149]
Fitting parameter, α	0.12
Transformation rate, μ_a	0.6^{-1} 1/s
Phase constants, w_o^a, w_o^d	$\pi, \pi/4$
Angular frequencies, w^a, w^d	$\pi/32, \pi/16$
Maximum altitudes, λ_1^a, λ_1^d	2, 1
Min altitudes, λ_2^a, λ_2^d	0, 0

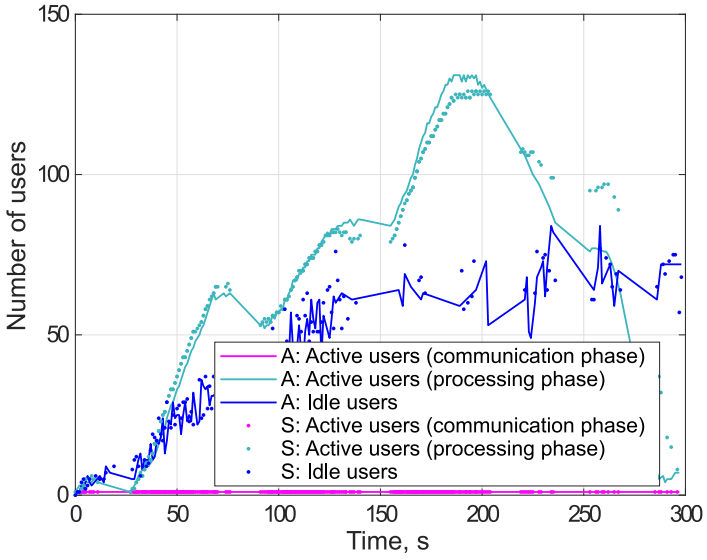

Fig. 3.7: Performance assessment (number of users): $B_t = 100\,\text{Mb}$, $B_c = 150\,\text{Mb}$, $R_0^t = 196\,\text{Mbps}$, $R_0^c = 150\,\text{Mbps}$ [4].

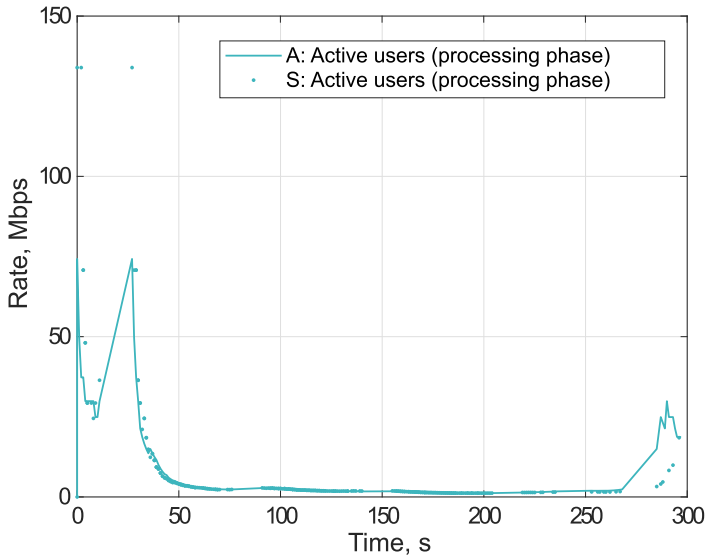

Fig. 3.8: Performance assessment (actual transmission rate/processing speed): $B_t = 100\,\text{Mb}$, $B_c = 150\,\text{Mb}$, $R_0^t = 196\,\text{Mbps}$, $R_0^c = 150\,\text{Mbps}$ [4].

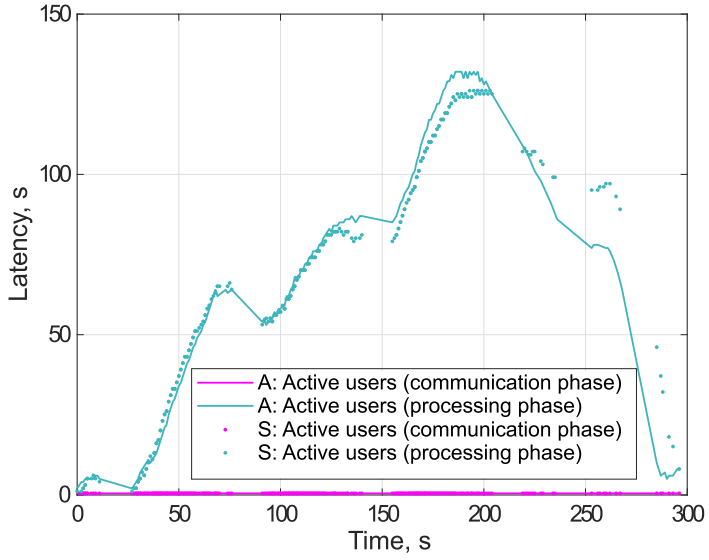

Fig. 3.9: Performance assessment (latency): $B_t = 100\,\text{Mb}$, $B_c = 150\,\text{Mb}$, $R_0^t = 196\,\text{Mbps}$, $R_0^c = 150\,\text{Mbps}$ [4].

Performance Analysis

In Fig. 3.7, Fig. 3.8, and Fig. 3.9, we observe that the number of active users involved in the communication and processing phases increases with a repetitive trend triggered by the periodic arrival rate function. As a result, the uplink transmission rate and processing speed decrease, which causes rising delays. The primary traffic bottleneck appears to be the

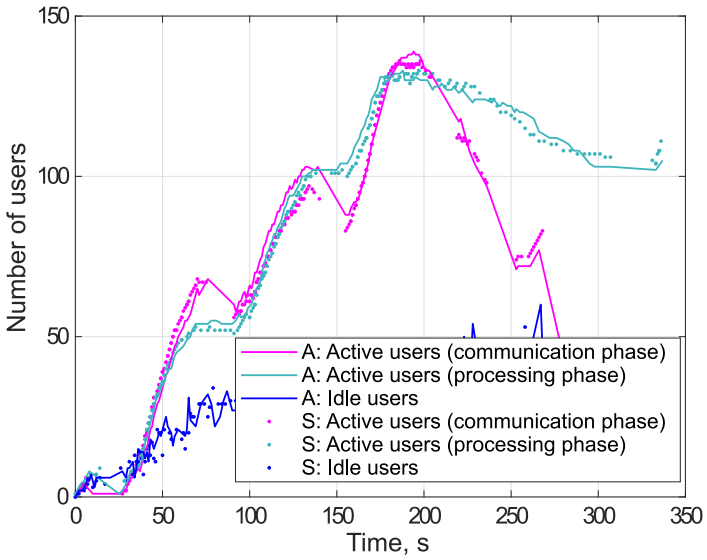

Fig. 3.10: Performance assessment (number of users): $B_t = 160\,\text{Mb}$, $B_c = 350\,\text{Mb}$, $R_0^t = 150\,\text{Mbps}$, $R_0^c = 150\,\text{Mbps}$ [4].

processing phase at the edge servers. At time instant $t = 221\text{s}$, no more content requests arrive at the system, leading to a gradual system unloading and a decrease in the number of users involved in the processing phase.

The key discovery is that the computation process for processing XR video is highly demanding in terms of computational resources. To address this issue, system designers can resort to using both parallel and distributed computing approaches to reduce the computation time for video processing at the edge server.

In Fig. 3.10, Fig. 3.11, and Fig. 3.12, differently from the previous setting, the uplink data rate drastically decreases. As a result, the number of active users involved in the communication and processing phases is comparable, and the bottleneck occurs during both the communication and processing phases.

The main finding is that both communication and computing planes are the key barriers to meeting the requirement for real-time transmission of spatial information from XR and video content processing. A B5G cellular network with multi-RAT multi-connectivity functionalities might be a promising candidate for supporting such XR-aided system operations.

In this section, we assessed the effectiveness of the fluid approximation that captures the time and space dynamics of the XR content distribution process in its transient phase considering the communication and computing planes. We evaluated the XR system performance for different system settings and identified the service bottleneck. We believe that our proposed method and practical conclusions have the potential to inspire system designers to create mobile communication and computing systems that deliver an immersive experience effectively.

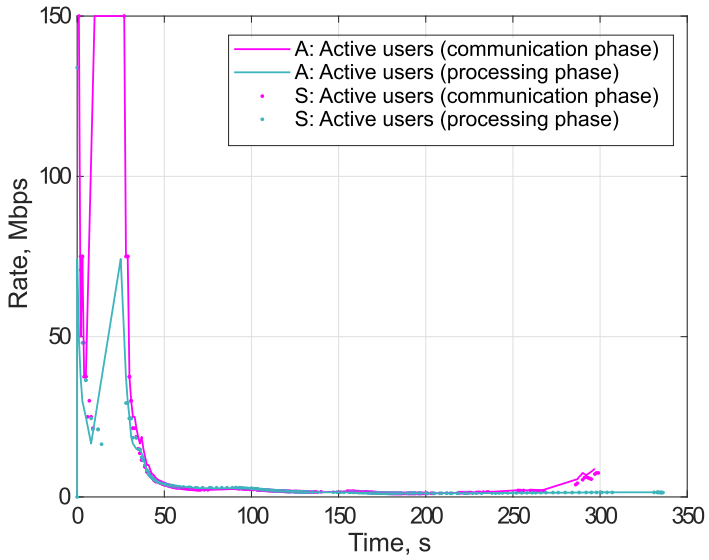

Fig. 3.11: Performance assessment (actual transmission rate/processing speed): $B_t = 160\,\text{Mb}$, $B_c = 350\,\text{Mb}$, $R_0^t = 150\,\text{Mbps}$, $R_0^c = 150\,\text{Mbps}$ [4].

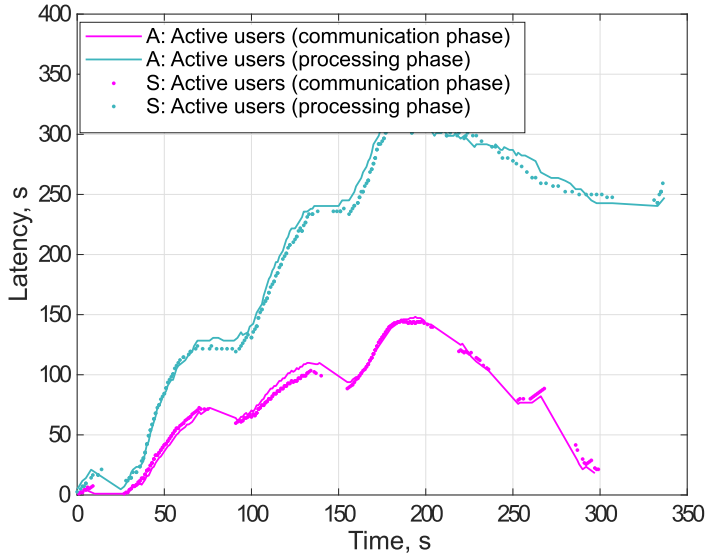

Fig. 3.12: Performance assessment (latency): $B_t = 160\,\text{Mb}$, $B_c = 350\,\text{Mb}$, $R_0^t = 150\,\text{Mbps}$, $R_0^c = 150\,\text{Mbps}$ [4].

3.2.3 Discussions

In this section, we provided the methodology to characterize the immersive reality content evolution in dynamic wireless networks with non-stationary arrival processes, i.e., a periodic arrival rate function. The approach captures the temporal and spatial dynamics of the content distribution process while taking into account both communication and computing planes.

We evaluated our analytical tool by testing it in different network configurations. As a result, we presented system designers with a way to evaluate the XR system components as well as provided immersive reality system characterizations that have significant implications for their businesses. The proposed tool may also serve as a valuable instrument for developing practical operational schemes and promoting ongoing standardization efforts.

However, as immersive interactions distract users from the real world and modify their behavior and motion, which, in turn, may affect the network operations, communication patterns need a profound transformation, which is investigated in the next section (Section 3.3).

3.3 Joint Behavior, Communication, and Computing Assessment Methodology

Mobile XR offers unique "anywhere anytime" interactive experiences such as real-time collaboration, training, and gaming, enabled by the ability to navigate virtual space through physical movement [156]. Unlike traditional interfaces such as mobile phones or tablets, XR submerges the users into a virtual world, enabling immediacy to complete immersion in different realities and distracting them from the surrounding environment [157]. While the XR users can freely navigate around the area (e.g., a room or a pedestrian way) and circumvent obstacles, such mobility might be affected by the patterns of use [158] and unique immersive interactions [159–161].

Compared to traditional user behavior, using an HMD leads to shorter stride lengths, longer stance times, and increased speed variability [159, 162, 163]. In addition, due to the unique characteristics of XR content presentation and navigation, HMD wearers' motion patterns may vary significantly from those of mobile phone users. In typical mobile phone applications, for instance, the walking patterns of people engaged in message writing and audio recording are distinct [123]. Typing a message on a small phone screen requires intense concentration and severely limits mobility, but voicemails may be received and transmitted with fewer restrictions. Due to the improvements in user perception, this significant difference in motion pattern may no longer define XR encounters.

The provided examples demonstrate that the XR uniqueness is derived not only from the stringent application requirements, such as high peak data rate and low latency for a fully immersive experience with a sense of reality but also from the interaction models and motion patterns, which may impact network performance. It is anticipated that XR applications would be adaptable and dynamic, necessitating a real-time reaction dependent on communication and computation capabilities. In addition, since the mobility of an XR user is likewise highly dependent on service provisioning, the output of such a system is redirected to interaction and movement dependence, so establishing a feedback loop. The convergence of communication, processing, and use/motion patterns (see Fig. 3.13) is the next stage in developing advanced XR services and future communications in general.

By bridging the current research gap of investigating communication networks from the perspective of user interaction patterns, this section uncovers the basic characteristics of XR user behavior and mobility and identifies the associated communication and computation challenges. First, we provide an assessment of user behavior patterns that verifies use case-dependent changes in gait characteristics, including direction, velocity, stride length, step width, and stance duration. In addition, we present the proof sources on the influence of user movements on network operation. Finally, we propose a case study for mobile XR that characterizes system performance in terms of user mobility, communication, and computation. We quantify the resulting interaction using system-level simulations and compare XR

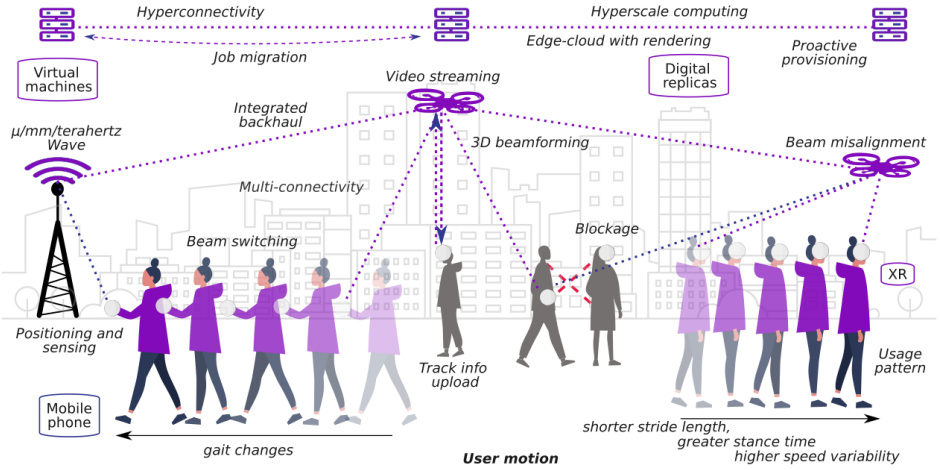

Fig. 3.13: Motion, communication, computing, and usage pattern loop [5].

with conventional mobile broadband services to determine the impact of use patterns on wireless communications performance.

3.3.1 Application-Dependent Mobility

Mobility patterns are designed to accurately imitate user motion in various scenarios by providing a realistic representation of real-world situations. Community-adopted mobility models allow for assessing the impact of multiple parameters, such as speed, direction, distance, cadence (defined as the number of steps per minute), stride length, and stance time, on the operation of communication systems. This subsection discusses XR-driven user behavior patterns collected by the research communities from various fields outside communications engineering, such as psychology, neuroscience, and intelligent transportation.

Multisensory immersion is one of the most influential elements of user behavior. Recent literature is filled with reports of walking experiments in XR. *Variations in gait characteristics, which represent motion instability*, have been intensively studied ever since the groundbreaking work in [158], which has been supported by recent studies in [162]. In the virtual world, *individuals walk with wider step widths, shorter step lengths, and more velocity variability*. These differences have also been evaluated in immersive XR vs. walking in the real world: the average walking speed reduces by 46 %, cadence and stride length decrease by 14 % and 33 %, respectively. Simultaneously, the stride width and stance duration rise by 18 % and 7 %, respectively [159].

Gait instability and variability are but one example of a phenomenon caused by submerging into immersive XR. Another one is a different adaptation in physical and virtual environments and related changes in circumvention – the process of avoiding obstacles and barriers by steering the body in another direction. Such behavior, being a fundamental part of daily routine, causes different adaptations in physical and virtual environments. *When employing immersive applications, avoiding collisions and seeking appropriate clearance for*

all parts of the human body need more cautious trajectory and speed adjustments, resulting in a more "conservative" circumvention strategy [158, 162]. Experiments on locomotor behavior in the virtual environment have shown slight but discernible variations in paths, more significant maximum deviation, higher obstacle clearance, and slower speeds compared to walking through obstacles in the real world.

Circumvention strategies have also been examined in real and virtual environments, with other pedestrians being obstacles. In this case, *circumvention patterns become role-dependent by dividing opposing users into passing and bypassing. Subjects who engage in immersive experiences have been found to have lower walking speeds and increased distances from potential interferences* [162]. However, as the number of repeated tests in the same environment grows, this disparity narrows, but users still prefer more "conservative" circumvention strategies for avoiding both stationary (items) and moving (other pedestrians) obstacles. The collision avoidance tendencies in actual and virtual worlds are nearly equivalent.

Another series of experiments have been focused on multitasking, which is described in neuropsychology as executing many tasks concurrently (typically two, which is referred to as dual-task activity), and comparing multi-task and single-task performance. *Speaking, texting, and calculating impact overall walking performance by affecting pace, cadence, and gait pattern* [164]. The gait, in turn, affects multitasking efficiency. Paced motion reduces user performance while doing dual-task. It makes walking more complicated, particularly when the displayed material limits the visibility of surrounding items essential for active dynamic locomotion [165].

Much research has focused on comparing the influence of mobile phones and XR wearables (such as HMDs) on gait variability. The findings are similar across several study groups and demonstrate that *head-up tasks (those utilizing HMDs) decrease walking performance less than head-down activities (those involving mobile phones)* [123]. The walking pace drops considerably for dual-task vs. single-task activity while using a mobile phone. The difference in head-up walking between single- and dual-task operations is negligible, supporting XR stability and multitasking resilience.

Equipment variety is another critical component that characterizes user motion. The gait pattern, in particular, is susceptible to both hardware and software [162]. This fact was discovered by testing XR wearables from various manufacturers, generations, and models. The findings revealed *a relation between user motion and the equipment model, influencing walking metrics such as speeds (minimum, maximum, and average) and trajectories.* Different degrees of equipment usability (measured by temperature, texture, tightness, weight, size, and shape) and perceived comfort in the virtual world (depends on image resolution, colors, time perception, and degree of realism) cause the variety in user behavior patterns.

In summary, application type significantly influences user motion. Compared to traditional mobile phone usage, engagement in XR applications results in broader and shorter steps and increased pedestrian pace fluctuation. Variations in gait characteristics arise from the physical limits of the user equipment, the difference between the virtual and real worlds,

and the unique immersive experience. Moreover, dual- and multi-task activities had a lesser impact on the mobility of XR users than mobile phone usage, demonstrating the uniqueness of immersive services. These unconventional behaviors of XR users may substantially affect radio communications.

3.3.2 Mobility-Dependent Communication

On-the-go connectivity presents significant issues, particularly for data-intensive systems such as immersive XR, which process large amounts of data. In this regard, we examine enabling management and optimization strategies for mobility support in the context of XR usage in the approaching B5G era. The future B5G technologies are expected to support seamless operation of extreme performance through advanced wireless access solutions, allowing "anywhere anytime" communications [166].

Today, many network operators already utilize microwave (μWave) radio systems operating in the 4.1–7.125 GHz band for basic coverage, which is expected to carry most of the traditional cellular traffic. However, unlike the Extremely High Frequency (EHF) bands, such as mmWave and THz, μWave systems are unable to meet the demand for multi-gigabit-per-second throughput and low latency communication in dynamic networks. To support high data rate services, network operators are expected to deploy both lower frequency and mmWave or even THz technologies [167, 168].

Integration of μWave/mmWave combines extreme transmission data rates with the reliability of legacy μWave channels. Specifically, μWave band can be utilized for lighter XR traffic, i.e., location information, whereas higher frequency bands are dedicated to heavier video streaming. However, user mobility challenges resource allocation in such multi-Radio Access Technology (RAT) scenarios since the backhaul network requires more frequent traffic re-routing [169]. In addition, high-frequency channels are prone to severe fluctuations and involve more complex beam management and Medium Access Control (MAC) protocols in general.

Multi-cell connectivity allows devices to ensure reliable data transmission by maintaining several signal paths from/to different BSs [170, 171]. The BS and users continuously monitor potential wireless links via dynamic beam tracking and beam refinement, which results in significant overhead in the case of the EHF band [172]. Moreover, high user mobility causes rapid load changes at each BS due to the smaller cell size in highly directional networks [4]. The network may benefit from accurate positioning and sensing information; however, this information is collected using the same radio resources, which may significantly raise the network overhead.

In addition, the specific position of a mobile XR device on the body necessitates further research evaluating motion and rotation patterns. For instance, the relative mobility of body components, such as the head or hands, might sometimes result in signal losses due to beam misalignment and/or obstruction of mmWave networks [173, 174]. Additionally, the height

at which the user holds the device affects blockage conditions, particularly in settings with a high user density [175, 176]. Since the XR wearable is connected to the user's head, it is anticipated that it will be less susceptible to blockage and, thus, less influenced by channel quality deterioration than smartphones held at chest level. This effect may lead to the need for new service provisioning models mindful of the diversity of use cases and corresponding behavior patterns.

In summary, efficient B5G wireless access for XR relies on different technologies and techniques, such as multi-RAT and multi-cell connectivity, integrated backhaul, localization and sensing, advanced beam-tracking and beam-training procedures, which are significantly affected by the user motion patterns. Reliable XR connectivity requires novel scheduling, resource, and mobility management protocols along with coordination and control algorithms that jointly account for and are flexible to be adjusted to the specific use cases. The discussion above concludes that traditional communication techniques might not be sufficient for immersive XR applications, demanding, inter alia, the development of novel tailor-made AI and ML tools that efficiently adapt to diverse and dynamic conditions.

3.3.3 Communication-Dependent Computing

XR may demand resource-constrained edge nodes to do intense computation (e.g., Three Dimensional (3D) rendering and analyzing user movements or camera feed). From this viewpoint, we address advanced computing approaches in the context of applications that initiate specific motion patterns. In addition, several XR services frequently request information, such as background sceneries, which require vast amounts of storage space and therefore pose a problem to typical edge cache management. This might impair data replication [177], necessitating extra processing and storage expenses for constant synchronization between digital copies, allowing real-time interaction and reliable communication between the digital domain and physical systems.

Furthermore, supporting seamless low-latency connectivity (i.e., 5 ms) with high data rates poses challenges in both the communication and computation domains, mainly due to user mobility. On the communication level, user mobility causes handovers. Consequently, virtualized representations of users and their data follow the user's route from one edge node to another [178]. As a result, proactive provisioning is critical for effective resource management under low latency constraints [179]. Compared to reactive techniques or data replication, sophisticated proactive solutions have various benefits, including precise synchronization with back-end storage and rapid access to an on-demand state, which aids in maintaining application performance.

Pre-loading computational jobs or data onto the target edge server is only one component of efficient proactiveness, which is highly dependent on user mobility and needs precise motion predictions [180]. Significant inaccuracy in this context may result in content regeneration and, as a result, an increase in latency, which XR applications cannot tolerate. The

association between XR users and edge servers is another essential component that benefits from predicting user location and mobility. Since the EHF band signal is susceptible to obstruction, multi-connectivity may improve data rates and mobility resilience. In this case, effective prediction of user orientation and mobility patterns in the immersive environment is crucial for the proactive association of users with BSs and edge servers.

Load balancing is further complicated by frequent radio handovers and migration of computing jobs/results when moving out of the edge server coverage. The network optimizes migration techniques using relevant information on nodes' capacity and current load. For instance, computations may be performed on the previously served edge node such that the results are sent to the moving user via a new server in close proximity. The computations may also migrate to an adjacent server or be sent to a network server with more processing capability [181].

In summary, user mobility, application type, and a vast volume of produced data demand increased network architecture flexibility, new application-specific configuration choices that provide dynamic adaptation, and consistent cross-application performance. Therefore, the design of novel advanced algorithms for configurable and reliable coordination of computations and communications becomes of paramount importance. To offer a truly immersive XR experience, the network requires further enhancements in virtualization, digitalization, cloudification, and device/network programmability. Notably, advanced AI solutions that predict motion and anomalies in network operation may help significantly improve the convergence of communication and compute functionalities.

3.3.4 Performance Evaluation

Our analysis of trends reveals that XR applications provide distinct usage patterns that influence user movements and, therefore, communication and computing capabilities. Regardless of usage trends, the research community relies on current pedestrian mobility models despite extensive work on B5G systems. We evaluate the communication and computation performance of mobile XR in terms of *(i)* total delay, the sum of communication and computing delays, and *(ii)* resource utilization, the ratio of utilized resource blocks to available resource blocks. Below we summarize the considered scenario of interest, simulation settings, and selected simulation results. The key system parameters are listed in Table 3.5.

Simulation Settings

We assume a user terminal with 4K resolution and a content provider that renders 8K video [182] and focus on two types of services, termed *weak interaction* and *strong interaction*. Weak-interaction applications include various audiovisual services, such as video and live broadcasts. In such cases, users have limited or no interactions with the surroundings, i.e., they may not trigger physical exchanges, but they may choose their viewing point and location. Since users do not move their heads often when information is displayed, freedom

is unavoidably constrained. Weak-interaction services tolerate the end-to-end/motion-to-photon latency of around 30 ms and content quality of 30 fps [182].

Table 3.5: System parameters related to Section 3.3.

Area of interest	**Area:** Street Canyon, 50 m x 200 m
Pedestrians	**Number:** 20 – 60 **Mobility:** Social force model [183] **Speed:** 3 km/h (baseline) **Height:** Normal distribution ($\mu = 1.65$, $\sigma = 0.08$ m)
Behavior models	Mobile phone / XR wearable usage **1.** Single-task mode **2.** Dual-task mode
Devices	**Category:** Mobile phone / HMD **Number:** Number of users
Traffic	**Uplink motion information data rate:** 150 kbps **Downlink frame size:** 0.425 Gb (150 : 1 rate)
Weak-interaction/strong-interaction	**Quality of experience:** 8K with 30/90 fps **Period between the requests** 33/11 ms **Typical Round-Trip Time (RTT) requirement:** 30/10 ms
Edge servers	**Deployment:** Servers are co-located with BSs
Edge processing	**Frame rendering time:** 16.9 ms **Degradation factor Input/Output (I/O) interference:** $d = 0.02$ **Number of VMs on an edge server:** 50
mmWave radio	**Frequency:** 28 GHz **Bandwidth:** 400 MHz **Signal degradation with human blockage:** 15 dB **Resource block size:** 1.44 MHz
μWave radio	**Frequency:** 3.5 GHz **Bandwidth:** 100 MHz **Signal degradation with human blockage:** 4 dB **Resource block size:** 0.72 MHz
Propagation	**Model:** 3GPP Urban Microcell (UMi) Street Canyon **Building blockage:** Line-of-Sight (LoS), Non-Line-of-Sight (NLoS) **Blockage:** Blocked, nBlocked
BSs	**Deployment:** Strauss process ($c = 0.9$, $\delta = 200$ m) **Transmit power:** 33 dBm **Height:** 10 m **Multi-connectivity degree:** $2, 4, 6, 8, 10$ **Handover delay:** $[2 - 10]$ ms
Devices	**Transmit power:** 10 dBm **Mobile phone / HMD height:** Normal distribution ($\mu = 1.50$ / 1.65, $\sigma = 0.08$ m)

In strong-interaction immersive scenarios such as virtual gaming arcades or XR social media, users interact with the virtual space and respond in real-time. The resolution is greatly enhanced, increasing the required bandwidth, while the end-to-end latency requirement ap-

proaches 10 ms. To provide a truly immersive experience, such services demand higher frame rates (90 fps) compared to the weak-interaction scenarios [182].

In the given settings, user devices communicate with multiple BSs, each co-located with an edge computing server, via a dual mmWave/μWave radio interface. We assume the 3GPP channel model in an UMi environment [184] for both the mmWave band at 28 GHz and the μWave band at 3.5 GHz. The BSs are located across the tracking area according to the Strauss process with the inhibition coefficient of 0.9 and the inhibition distance of 200 m [185]. Devices can transition to the BS providing the best Signal-to-Interference-plus-Noise Ratio (SINR) ratio with a handover delay of $2 - 10$ ms [186, 187].

We examine a system in which users send tracking data, such as user location, to the selected BS in the uplink channel and then to a back-end server utilized for precise synchronization and instant access to an on-demand state. The edge node renders video frames, which are then delivered back to the user through the serving BS. Different uplink and downlink communication bands in XR may be used for more efficient transmissions [188].

The End-to-End (E2E) delay (excluding encoding/decoding) includes the uplink transmission over μWave links, processing, migration, and downlink mmWave transmission latencies. The processing delay is calculated based on the measurements from Huawei 5G network XR test with edge/cloud services [189], while communication latencies depend on channel conditions. We also assume the implementation of virtual machines for parallel computing of multiple tasks at the same edge server with the degradation factor of 0.02, which defines computation-service rate reduction when multiplexed with other VMs due to I/O interference.

The period between requests is 33 ms and 11 ms for weak- and strong-interaction scenarios, respectively [182]. The uncompressed video frame size that has to be downloaded is 63.7 Gb (i.e., 8K resolution, 8-bit color depth). We utilize 150 : 1 compression rate that reduces the bandwidth and bit rate requirements, thereby decreasing the interaction latency. The uplink channel supports data rates of 150 kbps for transmitting the motion information [188].

We employ pedestrian flow modeling to simulate real-world user behavior in various application settings. The simulation is based on a model of human behavior based on social forces [183]. The model replicates realistic crowd dynamics as seen in applications such as collective XR and virtual gaming. We run simulations under various density situations (i.e., $20 - 60$ users in the area).

We experiment with pedestrian movement by considering the diversity in speed, stance duration, step length, head direction, and the presence of obstacles that define human behavior in XR and mobile phone applications. As a baseline model, we simulate user movement in single- and dual-task modes by modeling the motion of a pedestrian walking at 3 km/h [123].

In a *single-task* setting, XR user motion changes in speed (-46%), step length (-33%), stance time ($+7\%$), and distance from an interferer or obstacle ($+3\%$) compared to baseline setup [159, 162]. Regarding users with mobile phones, the difference in speed, step length,

stance time, and distance from an interferer is -25%, -20%, 0%, $+2\%$, respectively. In a *dual-task* mode, XR leads to the difference of -70% in speed, -65% in step length, $+20\%$ in stance time, $+7\%$ in the distance from an interferer, compared to the baseline model. For mobile phone services, the corresponding variations in parameters are given by -80%, -69%, $+27\%$, and $+5\%$. We also model the variability in the mobility of the user equipment (located on the head or hand) by reducing the range and the frequency of motion for dual-task activities with respect to the single-task regime [164].

To illustrate the above claims, we provide the results in terms of speed, distance, and acceleration, see Fig. 3.14, Fig.3.15, and Fig. 3.16 for four motion/use models, i.e., the human motion with mobile phone and single-task mode, XR wearable and single-task mode, mobile phone and dual-task mode, XR wearable and dual-task mode.

Performance Analysis

The application type, particularly XR, has been linked to user response in terms of gait patterns. In reality, interaction patterns have a considerable influence on user mobility, which impacts – on the communication level – multi-connectivity and handover and – on the computing level – job migration. This subsection evaluates the convergence of usage, motion, communication, and computing patterns by quantifying the differences between XR and conventional mobile broadband applications in terms of the E2E delay (as shown in Fig. 3.17 and Fig. 3.18) and resource utilization (Fig. 3.19) subject to user density, service quality, and BS deployment settings.

We begin our analysis by examining the E2E latency for four mobility models linked with mobile phone and XR use under single- and dual-task conditions. First, Fig. 3.17 reports the

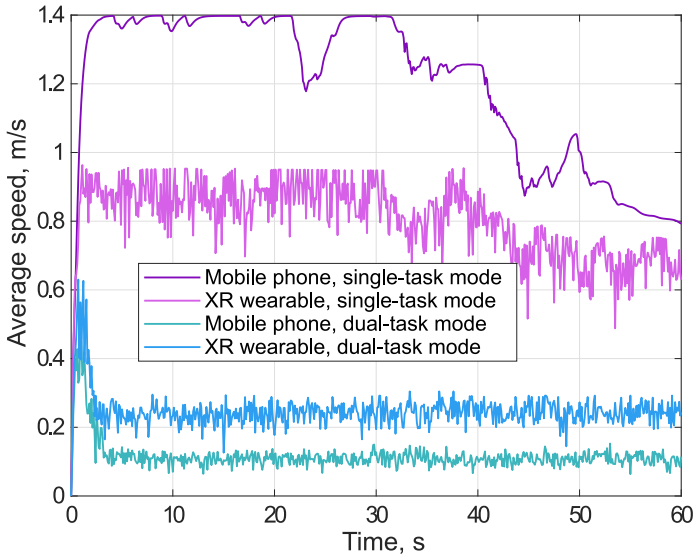

Fig. 3.14: Speed for four user behavior models.

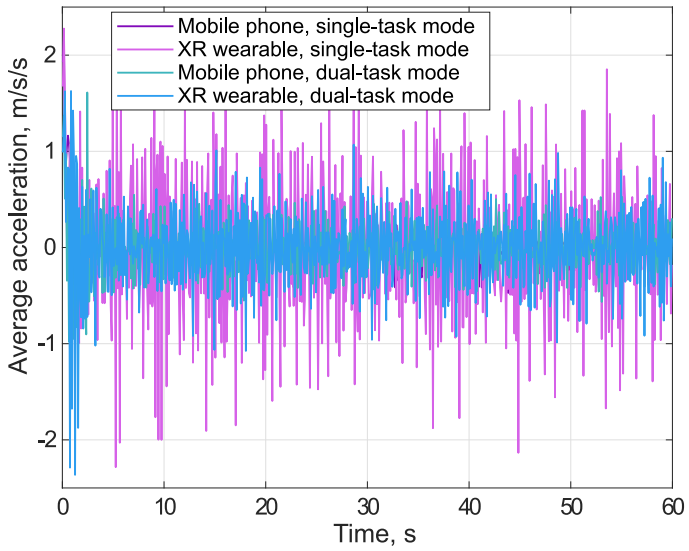

Fig. 3.15: Acceleration for four user behavior models.

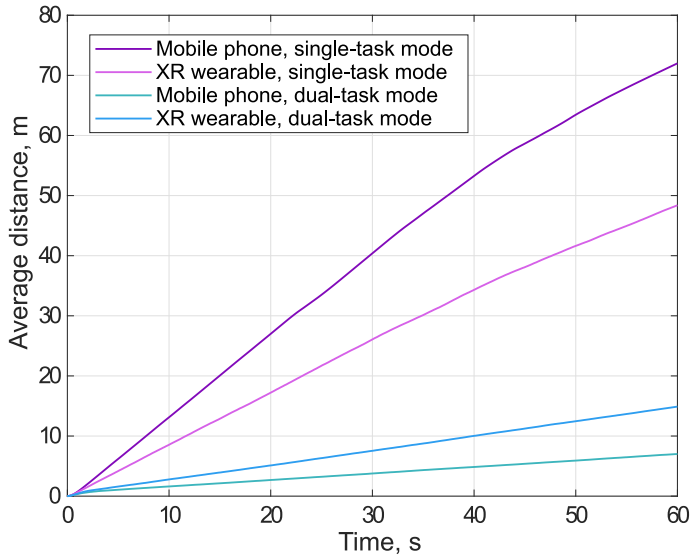

Fig. 3.16: Distance for four user behavior models.

latency for weak-interaction (30 fps) and strong-interaction (90 fps) services under low (20 users) and high (60 users) density. Under single-task and dual-task situations, the system-level performance of XR and mobile phone applications changes significantly. The observed gap is caused by distinct motion patterns that affect connection and different channel conditions that depend on factors such as equipment height. The difference becomes more visible when service quality improves, i.e., in the case of strong-interaction services, which cause heavier system demand, and in the case of high user density, due to both load and blockage.

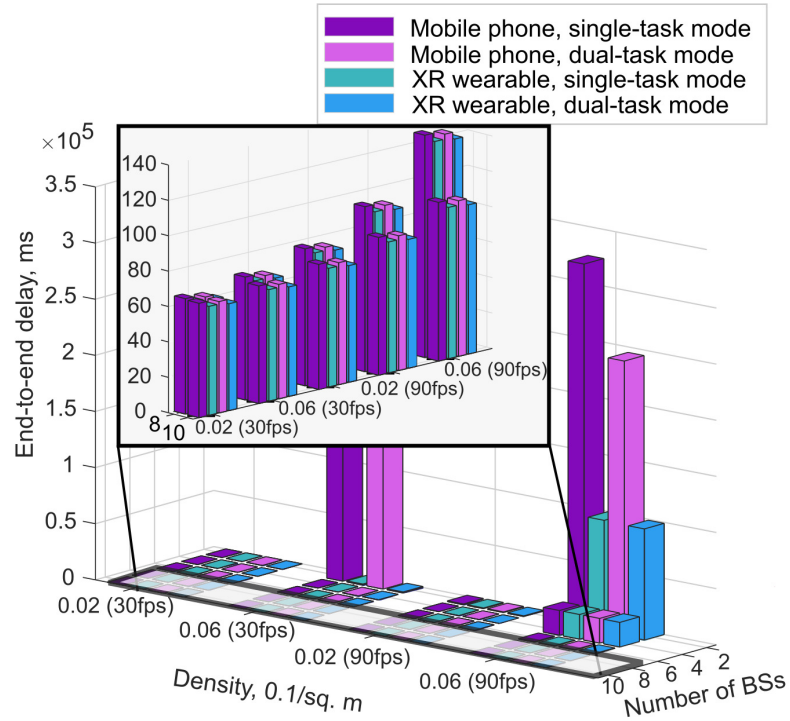

Fig. 3.17: E2E delay assessment [5].

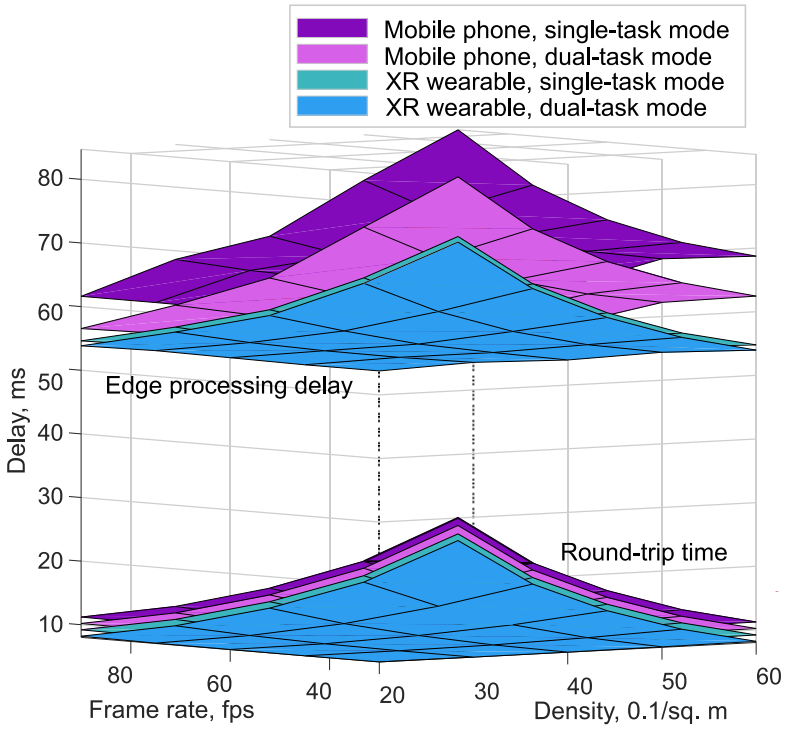

Fig. 3.18: RTT and edge processing delay assessment, 10 BSs [5].

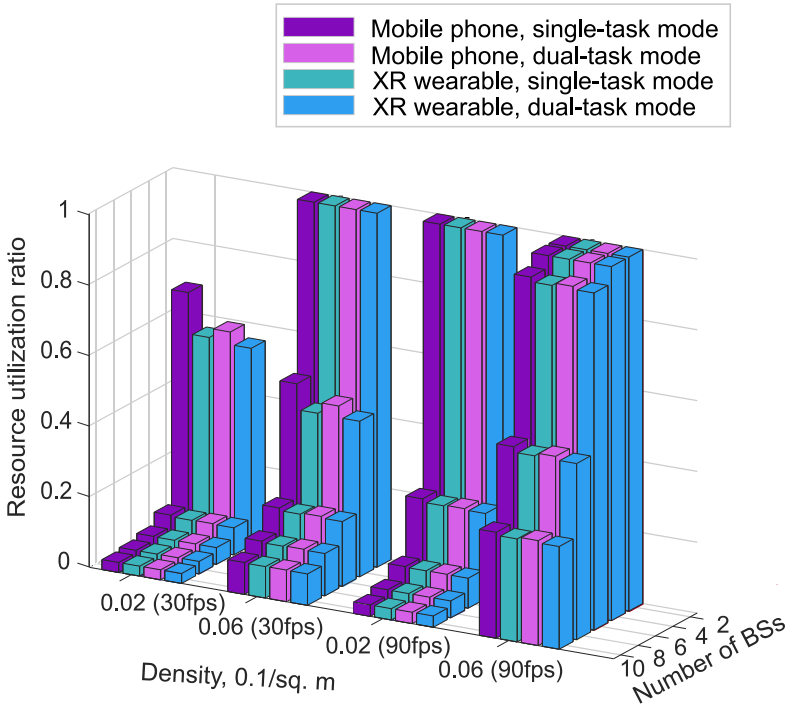

Fig. 3.19: Resource utilization assessment [5].

Furthermore, in Fig. 3.17, we observe the impact of the multi-connectivity degree (the number of available BSs) on the system latency. The gap between XR and phone applications is substantially lower for higher degrees of multi-connectivity than for 2-4. The reason for these findings is that multi-connectivity guarantees a better processing environment as the number of available BSs, servers, and VM rises regardless of application type. In particular, significant discrepancies between XR and mobile phone use cases occur in systems with the multi-connectivity degree of 2; if 10 BSs are deployed in the area, these gaps decrease to 11 % and 9 % for single- and dual-task modes, respectively.

We evaluate communication- and computing-related delay separately to better understand the influence of user movements on communication and processing patterns. The difference between the mobile phone and XR models in Fig. 3.18 preserves the same trend as in Fig. 3.17. In the case of increased user density and strong-interaction mode, RTT variations between the mobile phone and XR approach 2 ms and 3 ms, respectively, for single-task and dual-task modes. For lower density, however, mobile phone and XR-driven communication delays exhibit similar trends. In contrast, the differences in edge processing latency between mobile phone and XR models are observable under any density and service quality conditions, reaching up to 10 ms (single-task) and 17 ms (dual-task). Different body blockage patterns cause this phenomenon, which leads to more frequent beam switching for mobile phones as user density increases. In this case, blockage impacts the frequency of handovers and job migrations.

As per our additional results, we study the impact of multi-connectivity on communication and computation performance. The BS density affects communication and computation functionalities differently in terms of delay. As the number of BS alternatives available to customers increases, the average SINR rises, and the transmission latency decreases, approaching the same values regardless of application type. However, handovers create more frequent job migrations, increasing the overall edge processing time as the multi-connectivity degree increases. Furthermore, we performed additional simulation experiments to evaluate the effect of motion parameters on the performance. The key variables influencing system-level outcomes are the distance from interferers, XR/phone height, head direction/hand position, and changes in terminal location induced by head or arm movements.

Furthermore, Fig. 3.19 depicts the effect of the application on radio resource utilization for downlink transmission. To that end, we assume the 28 GHz mmWave carrier frequency and the corresponding NR numerology $mu = 3$ with a physical resource block size of 1.44 MHz. Since downlink transmission delays dominate RTT in systems such as XR, the increase in resource deviations reflects the variations in RTT delay, which also increase with BS densification. The distances between the BS and the users decrease as the number of BS options increases. The SINR increases in this situation, boosting the resource usage ratio. However, the downlink transmission of the processed video frames significantly contributes to system load, acting as a second bottleneck (after computing resources) in XR systems. As one may notice, the communication resources of 2 BSs are utilized at total capacity for strong-interaction services at any user density. For 30 fps and multi-connectivity of degree 2, the difference between XR and mobile phone use cases is 21% and 7% in single- and dual-task modes, respectively; for 90 fps and 10 BS setup, it is less noticeable but remains at the level of 1% and 0.7%. The observations in Fig. 3.19 related to the need for high degrees of multi-connectivity are confirmed in Fig. 3.17. We may infer that BS densification is essential to fulfill the demand for high-quality XR services and assure low-latency communication in future wireless networks, which are projected to offer high-quality XR services under high user density.

As per our additional results, these conclusions also maintain for different mobility models, e.g., the Lévy walk process. Since we consider such applications as collective XR and virtual games, we focused on crowd dynamics. Recent research studies in psychology, neuroscience, and intelligent transportation have demonstrated that individuals anticipate the movements of their neighbors to find their routes in dynamic pedestrian flow [190–192]. This path-seeking behavior causes pedestrians to deviate from their intended course, that is, the direct path to their destination. The results in [192] have confirmed that vertical pedestrian motions are governed by a superdiffusive dynamic (Lévy walk) process. It has also been demonstrated that the path-seeking behavior is performed when using a scale-free movement strategy known as a Lévy walk, which may assist the transition to group-level behavior.

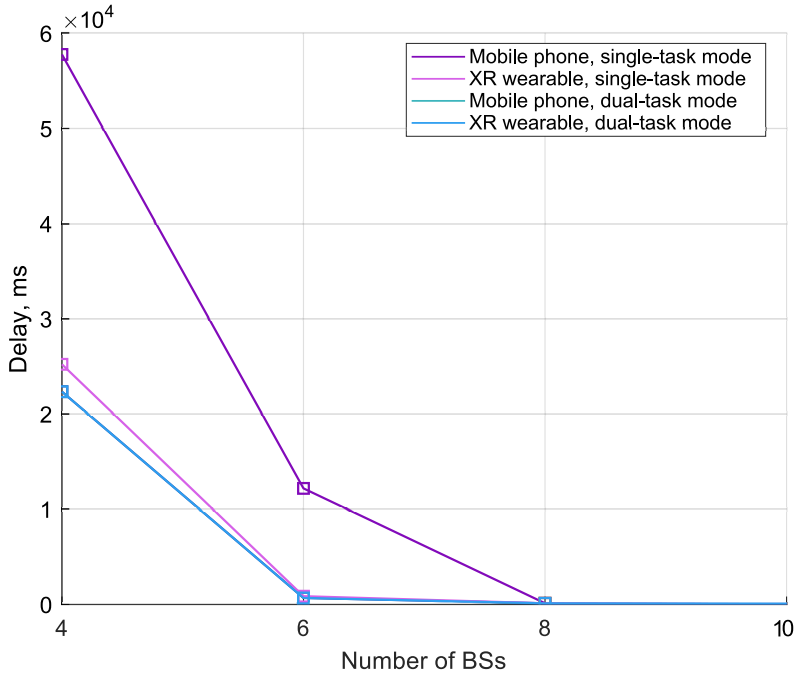

Fig. 3.20: E2E delay assessment, 60 users, 90 fps.

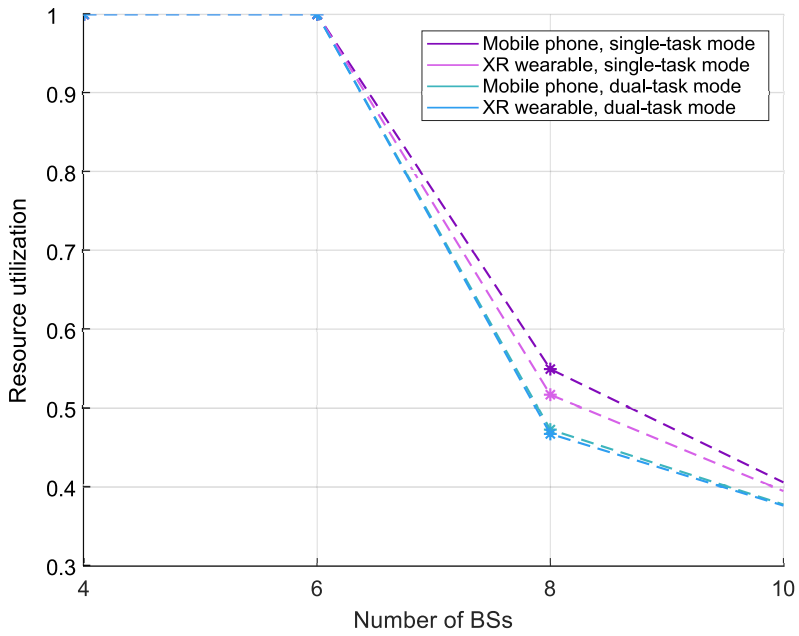

Fig. 3.21: Resource utilization assessment, 60 users, 90 fps.

We have conducted additional simulations with the Lévy walk process, a specific type of random walk in which the walker makes a few long steps and a large number of short steps, resulting in a power-law distribution of step lengths [190, 193–196]. In Fig. 3.20, Fig. 3.21, Fig. 3.22, and Fig. 3.23, we provide an E2E delay and resource utilization assessment for weak- (30 fps) and strong- (90 fps) interaction services. Specifically, delay deviation between

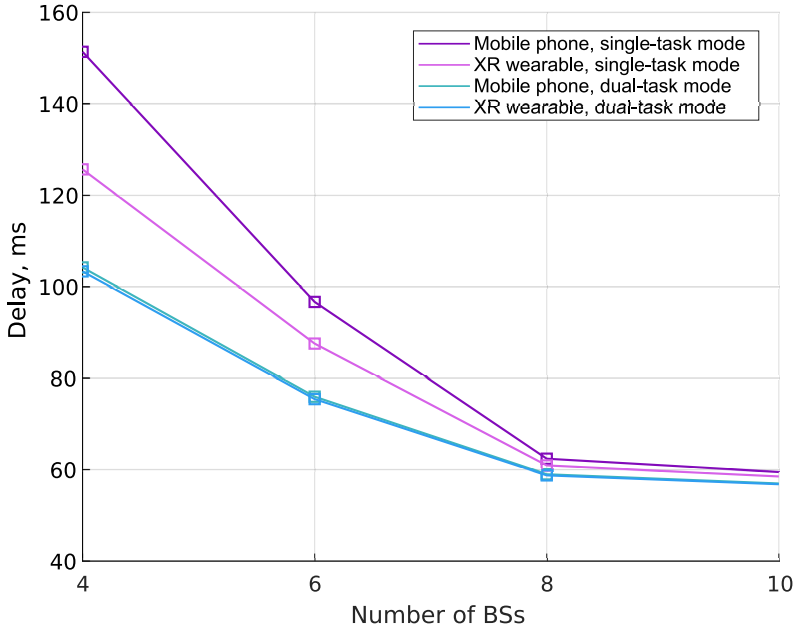

Fig. 3.22: E2E delay assessment, 60 users, 30 fps.

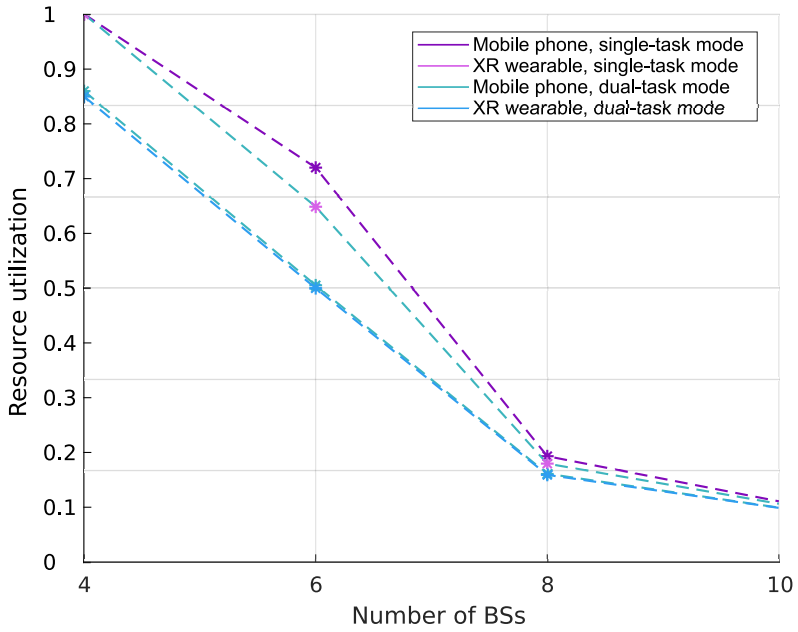

Fig. 3.23: Resource utilization assessment, 60 users, 30 fps.

mobile phone and XR usage reaches 32% and 17% for weak-interaction services in single- and dual-task modes, respectively, and 62% and 12% for strong-interaction services. The deviations in resource utilization between models reach up to 30% and 23% for weak-interaction services in single- and dual-task modes, respectively, and 15% and 12% for strong-interaction services.

In conclusion, our research proved that application usage patterns influence user behavior models and communication and computing performance. Mobile XR is distinctive not just in terms of system requirements but also in terms of interaction, mobility, computing, and communication patterns. The development of efficient architectures and algorithms for the network edge, as well as the design of innovative protocols and coordination and control methods for the radio network, are essential for more flexible, configurable, reliable, and robust computations and communications.

3.3.5 Discussions

With a trend toward immersive interactive and contextualized experiences, new use cases may be identified that require not only low-latency and high-bandwidth connectivity but also influence user mobility and, therefore, system performance. This interaction also operates in reverse. The engagement and user gait patterns may differ depending on the quality of service provisioning. Consequently, there is a feedback loop that includes patterns of use, mobility, communication, and computing. However, the research community has not paid enough attention to this significant influence.

To address this gap, we examined XR-driven motion patterns, reviewed relevant information, and performed comprehensive simulation research on the influence of XR usage on communication and computing performance. Our system-level analyses indicated that the usage of untethered XR had a specific influence on motion models and, more broadly, service provisioning. We envision collecting real-world mobility pattern datasets and applying advanced methods of data analysis to refine the employed dependencies and identify other factors that affect system performance. This study intends to motivate the research community to reconsider the standard mobility patterns and service models currently in use and move toward algorithms, architectures, and service provisioning methods that accurately capture user motion based on patterns of use.

4

Conclusions

We conclude this thesis with a summary of the main research contributions and conclusions. We then present future research avenues.

4.1 Summary

In this thesis, to support dynamic social-aware networking in 5G and B5G, we provided *(i)* the design of a reference architecture for the orchestration of SDTs at the network edge to accelerate the service discovery procedure across the SIoT; *(ii)* a set of methodologies to evaluate the highly dynamic system performance considering jointly communication and computing resources; *(iii)* a set of practical observations and outcomes helpful in designing future digital twin-enabled B5G networks.

The study of this thesis has led to the following conclusions:

- Accounting for social features in SDT placement offers considerable improvements to the SIoT browsing procedure.
- Lower latency among SDTs has two benefits: *(i)* it reduces network load when SDTs exchange data because packets traverse fewer links, and *(ii)* it ensures rapid interactions among SDTs, which is critical for service discovery procedures that include traversing the social graph.
- Recent advancements in wireless communications, edge computing, and smart device technologies are expected to promote the growth of SIoT with pervasive sensing and computing capabilities. This will ensure seamless connections and autonomous management among SIoT objects without human interaction, potentially changing industries and providing significant societal benefits.
- Both communication and computing planes are the key barriers to meeting the requirement for real-time transmission of spatial information from XR and video content processing.
- New use cases affect user mobility and, therefore, system performance due to the trend toward immersive, interactive, and contextualized experiences. This interaction also works

in reverse. Interaction and user gait patterns may vary depending on the quality of service provisioning.

- There is a need to revisit the typical mobility patterns and service models being used by the research community and move toward algorithms, architectures, and service provisioning techniques that accurately capture user movements based on usage patterns.

4.2 Future Research and Challenges

In this section, we discuss future research avenues and critical issues in digital twinning and XR systems modeling.

4.2.1 Digital Twins from Networking and Modeling Perspectives

Despite the enormous promise of DTs, their development and implementation in real-world applications remain challenging. Specifically, although the concept is currently being utilized in various fields, numerous obstacles should be overcome before DT can be identical to their physical counterparts. This subsection briefly discusses the key networking-related challenges in DT research and their potential application.

DT-assisted Content Caching.

In the context of DT-enabled content caching, several unresolved issues and technical challenges require further research. One of the critical challenges is incentivizing intelligent devices, such as objects in the ecosystem, to create edge caching clouds [197]. Additionally, there is a need for further study on composition update and service maintenance of cache cloud in the case of complex scenarios with dynamic traffic and non-stationary object topology. The issue becomes even more acute if the system contains objects of different types.

Functional Dimension.

Completely functional DTs should, by definition, accurately replicate the physical ecosystem in its temporal, geographical, and operational aspects at all granularities. This implies that a digital counterpart should include *all parameters* of the physical system. As a result, an entirely effective DT implementation requires modeling the physical ecosystem with the vast number of state types as well as gathering and processing an enormous amount of data. However, existing methodologies and computational tools may not be able to handle this dimensionality problem [198]) in DT implementation. To overcome this issue, the functional digital replica of a corresponding physical system may be divided into multiple DTs, each representing a subclass of the characteristics and states of the biological ecosystem in specific temporal and geographical zones. Another solution might be to employ modern high-speed CPUs and more advanced big-data analysis techniques.

Security.

The use of DTs raises significant security concerns. Since a DT serves as a duplicate of a physical system, it is necessary to protect both the biological system and its DTs, as well as the links they create (i.e., communication links between physical and virtual environments). DTs, which are used to communicate with third-party programs and applications, are more vulnerable to external threats than the physical system making their security more fragile. Any security breach affecting the DTs also affects the physical systems. However, since the DT and the physical system are physically separated, identification of a security breach of a DT may be delayed. Additionally, DTs may influence and/or control their physical systems;

therefore, keeping DTs safe from hackers and intruders is crucial to ensure the continued operation of a physical system. Therefore, there is a need for improved transparency and interpretability of decisions based on DTs, since most physical assets for which DTs may be utilized need a high degree of safety and security [198, 199].

Optimal Resource Management.

A DT should continuously monitor the real-time state of the physical system and update the system's features accordingly. Network resources such as communication, processing, and caching should be supported to construct and sustain DT. Creating and maintaining the DT with the appropriate quality requires joint heterogeneous resource management – a complex problem for large networks – to identify the required volume of the network resource as well as resource deployment and allocation. Additionally, various applications may have different requirements for promptness and similarity and diverse information derived from the system states. This leads to the need for optimization with multiple objectives, which may be contradictory to one another. Also, coordinating the application requirements in the network resource allocations is a challenging problem, given the limited availability of network resources [198].

Real-Time Communication, Data Management, and Model Update. A real-time two-way connection between the actual physical system and its DT is crucial for DT technology to achieve complete physical realism. However, maintaining it is hindered by such obstacles as spatio-temporal resolution of sensor data, significant communication latency, large data volume, high data generation rate, great variety and significant trustworthiness of data, fast archival retrieval, and online data processing. Additionally, the models need to change in perfect agreement to ensure backward compatibility when the physical asset changes over time, which requires interpretable and physically consistent models. Finally, the DT should be displayed to the user in a form that seamlessly integrates with the physical asset and is straightforward to use [199].

4.2.2 Immersive Reality from Networking and Modeling Perspectives

This subsection highlights essential research directions and challenges in networks and devices with heterogeneous capabilities in immersive experience applications.

User Behavioral Data and Social XR. The tendency for users to shift their attention from one screen to another is becoming increasingly common due to the development of many screen technologies. To address this, creative solutions built on users' social interactions and behavioral data should be considered. The screen chaos issues are interconnected and can be resolved with the same solution, which is an immersive experience that requires a data-driven architecture gathering in one location all of the relevant information that the user observes. However, such integration is not currently possible due to the lack of a shared platform. Moreover, these experiences are supposed to occur all in the exact location as virtual reality demands. For example, when receiving a call while playing a game or watching a movie, the

game or movie should automatically pause so the users need not worry about halting the game or movie to accept the call. In this scenario, big data and machine learning techniques will be essential in providing consumers with an immersive experience, given that a common data-driven platform is being used [200].

Context Information and Analytics.

It has already been suggested that context information can help optimize complex immersive experience networks. Note that in-device and in-network side data are typically referred to as context information [200]. The recent acquisition of Apple of AI startup Emotient, a company that uses advanced computer vision to identify people's emotions in the context of immersive experience, suggests that context information will play an ever-more-important role in driving the success of XR. The user's emotional state and other behavioral factors need to be considered to enhance their connected and immersive experiences. This involves anticipating and addressing user disengagement by dynamically changing the delivered context to better align with user preferences, emotional states, and viewing points. To this aim, AI tools can infer from user context information and respond accordingly.

Large-Scale XR Systems.

Another area of intense interest is exploring large-scale XR networks that are characterized mainly by dynamicity. Such systems contain many different viewpoints and types of information and hence, may utilize a high level of redundancy and collective intelligence to enable the interconnected immersive experience [200].

Computing Level.

This refers to the location and level at which the in-device (i.e., headset) and in-network processing should be decoupled. Depending on the bandwidth-latency-cost-reliability trade-offs, computation for less powerful low-end devices may be offloaded to the network. In contrast, computing for more complex high-end devices might be performed locally, which is, however, limited by the heat dissipation problem.

Localization and Tracking Accuracy.

A completely immersive XR experience requires accurate localization and tracking techniques, including the locations of objects, tracking of human eyes (also known as gaze tracking) [201], gesture recognition, change in velocity, and many more.

Green XR.

The aim is to reduce power consumption in terms of storage, computing, and communication for specific users in an immersive experience. Since power consumption reduction does not take place in the virtual world, the idea of *charging* the equipment should vanish with the introduction of green interconnected XR, or at the very least, be limited. As a result, intelligent methods for wireless power transfer and charging, as well as energy harvesting, appear promising for XR equipment [200].

Privacy.

Privacy is a major concern as the users contribute and have access to a wide variety of content and viewpoints from billions of items and users. There is a need for intelligent systems

that automatically protect privacy without placing a burden on individuals to adjust their privacy settings. Novel concepts like "collective privacy" may be promising to explore [202].

Harnessing Quantum.

Quantum computing has the potential to perform certain computations far more quickly than any classical computer could ever hope to achieve. Utilizing quantumness in XR could *(i)* create a bridge between the virtual and real worlds, where the traditional concept of locality is no longer relevant; *(ii)* hahdle objects in lower dimensions by utilizing entanglement and superposition in place of serial or even parallel processing [200].

Interoperability.

Virtual media- and information-rich environments have helped different construction stakeholders understand and visualize the design effectively. However, there is still a need to streamline the workflow of architecture and construction. To address this, many software providers, including Unity, have recently attempted to bridge this gap by using middleware. These advancements are still in the early stages and need further refinement and development. Additionally, the transfer of Building Information Modeling (BIM) models and associated meta-data into the Unity game engine to offer an immersive experience has become more straightforward with the release of Unity Reflect. However, creating interactivity remains a challenging task and requires customized algorithms.

References

1. O. Chukhno, N. Chukhno, G. Araniti, C. Campolo, A. Iera, and A. Molinaro, "Optimal Placement of Social Digital Twins in Edge IoT Networks," *Sensors*, vol. 20, no. 21, p. 6181, 2020.
2. O. Chukhno, N. Chukhno, G. Araniti, C. Campolo, A. Iera, and A. Molinaro, "Placement of Social Digital Twins at the Edge for Beyond 5G IoT Networks," *IEEE Internet of Things Journal (Early Access)*, 2022.
3. O. Chukhno, N. Chukhno, G. Araniti, C. Campolo, A. Iera, and A. Molinaro, "Social-Aware Orchestration in 5G+/6G-IoT Ecosystems," *(Ready for submission)*, 2023.
4. O. Chukhno, O. Galinina, S. Andreev, A. Molinaro, and A. Iera, "Content Distribution Dynamics of Edge-Aided Immersive Reality Services," *(Ready for submission)*, 2023.
5. O. Chukhno, O. Galinina, S. Andreev, A. Molinaro, and A. Iera, "Interplay of User Behavior, Communication, and Computing in Immersive Reality 6G Applications," *IEEE Communications Magazine*, vol. 60, no. 12, pp. 28–34, 2022.
6. G. Smith, "From 1982 Coca-Cola Vending Machine to Latest Trend: What the Internet of Things Means for Business," *Real Business*, vol. 15, 2015.
7. J. Coates, "Why Retail Giant Coca-Cola is Using IoT Connected Vending Machines," *Internet of Business*, 2016.
8. A. H. Alavi, P. Jiao, W. G. Buttlar, and N. Lajnef, "Internet of Things-enabled Smart Cities: State-of-the-Art and Future Trends," *Measurement*, vol. 129, pp. 589–606, 2018.
9. Juan Pedro Tomas, "Global IoT Connections to Reach 50 billion by 2030: Study." [Online] (accessed January 24, 2023) `https://enterpriseiotinsights.com/`, 2019.
10. L. Atzori, A. Iera, and G. Morabito, "Sociocast: A New Network Primitive for IoT," *IEEE Communications Magazine*, vol. 57, no. 6, pp. 62–67, 2019.
11. A. Ometov, O. Chukhno, N. Chukhno, J. Nurmi, and E. S. Lohan, "When Wearable Technology Meets Computing in Future Networks: A Road Ahead," in *Proceedings of the 18th ACM International Conference on Computing Frontiers*, pp. 185–190, 2021.
12. Ericsson, "Ever-present Intelligent Communication: Introduction – 5G and Beyond," *White paper*, Novemmber 2020.
13. E. Ekudden, "Future network trends," *Ericsson technology review*, September 2020.

14. F. Tang, X. Chen, M. Zhao, and N. Kato, "The Roadmap of Communication and Networking in 6G for the Metaverse," *IEEE Wireless Communications*, 2022.
15. GSMA, "Cloud AR/VR Whitepaper," *White paper*, 2019.
16. H. Viswanathan and P. E. Mogensen, "Communications in the 6G Era," *IEEE Access*, vol. 8, pp. 57063–57074, 2020.
17. C. V. N. Index, "Global Mobile Data Traffic Forecast Update, 2016-2021 White Paper. 2017. Cisco, San Jose, CA, USA, March 28, 2017."
18. P. Jonsson, S. Carson, S. Davis, G. Blennerud, P. Lindberg, K. Öhman, J. Travers, F. Pedersen, P. Linder, J. Sethi, P. Rinderud, J. Alonso-Rubio, and G. JL, "Ericsson Mobility Report," June 2020.
19. L. Atzori, A. Iera, G. Morabito, and M. Nitti, "The Social Internet of Things (SIOT)–When Social Networks Meet the Internet of Things: Concept, Architecture and Network Characterization," *Computer Networks*, vol. 56, no. 16, pp. 3594–3608, 2012.
20. M. Roopa, S. Pattar, R. Buyya, K. R. Venugopal, S. Iyengar, and L. Patnaik, "Social Internet of Things (SIoT): Foundations, thrust areas, systematic review and future directions," *Computer Communications*, vol. 139, pp. 32–57, 2019.
21. L. Atzori, C. Campolo, A. Iera, G. Milotta, G. Morabito, and S. Quattropani, "Sociocast: Design, Implementation and Experimentation of a New Communication Method for the Internet of Things," in *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*, pp. 662–667, IEEE, 2019.
22. M. Nitti, R. Girau, and L. Atzori, "Trustworthiness management in the social internet of things," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 5, pp. 1253–1266, 2013.
23. M. Grieves, "Origins of the Digital Twin Concept," *URL: https://www. researchgate. net/publication/307509727_ Origins_ of_ the_ Digital_ Twin_ Concept*, 2016.
24. M. Wise, "APM: Driving Value with the Digital Twin," in *Proc. GE Digit.*, pp. 1–43, 2017.
25. M. Grieves and J. Vickers, "Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems," in *Transdisciplinary Perspectives on Complex Systems*, pp. 85–113, Springer, 2017.
26. K. M. Alam and A. El Saddik, "C2PS: A Digital Twin Architecture Reference Model for the Cloud-based Cyber-Physical Systems," *IEEE Access*, vol. 5, pp. 2050–2062, 2017.
27. M. J. Kaur, V. P. Mishra, and P. Maheshwari, "The Convergence of Digital Twin, IoT, and Machine Learning: Transforming Data into Action," in *Digital twin technologies and smart cities*, pp. 3–17, Springer, 2020.
28. D. Cearley and B. Burke, "Top 10 Strategic Technology Trends for 2019: A Gartner Trend Insight Report," March 2019.
29. C. Mouradian, D. Naboulsi, S. Yangui, R. H. Glitho, M. J. Morrow, and P. A. Polakos, "A Comprehensive Survey on Fog Computing: State-of-the-Art and Research Challenges," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 416–464, 2017.

30. T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1657–1681, 2017.

31. Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

32. F. Tao, H. Zhang, A. Liu, and A. Y. Nee, "Digital Twin in Industry: State-of-the-Art," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2405–2415, 2018.

33. B. R. Barricelli, E. Casiraghi, and D. Fogli, "A Survey on Digital Twin: Definitions, Characteristics, Applications, and Design Implications," *IEEE Access*, vol. 7, pp. 167653–167671, 2019.

34. M. Nitti, V. Pilloni, G. Colistra, and L. Atzori, "The Virtual Object as a Major Element of the Internet of Things: A Survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1228–1240, 2015.

35. L. Atzori, J. L. Bellido, R. Bolla, G. Genovese, A. Iera, A. Jara, C. Lombardo, and G. Morabito, "SDN&NFV Contribution to IoT Objects Virtualization," *Computer Networks*, vol. 149, pp. 200–212, 2019.

36. T. He, H. Khamfroush, S. Wang, T. La Porta, and S. Stein, "It's Hard to Share: Joint Service Placement and Request Scheduling in Edge Clouds with Sharable and Non-Sharable Resources," in *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*, pp. 365–375, IEEE, 2018.

37. L. Wang, L. Jiao, T. He, J. Li, and M. Mühlhäuser, "Service Entity Placement for Social Virtual Reality Applications in Edge Computing," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pp. 468–476, IEEE, 2018.

38. Q. Fan and N. Ansari, "On Cost Aware Cloudlet Placement for Mobile Edge Computing," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 4, pp. 926–937, 2019.

39. S. K. Rao and R. Prasad, "Impact of 5G Technologies on Industry 4.0," *Wireless Personal Communications*, vol. 100, no. 1, pp. 145–159, 2018.

40. Y. Siriwardhana, P. Porambage, M. Liyanage, and M. Ylianttila, "A Survey on Mobile Augmented Reality with 5G Mobile Edge Computing: Architectures, Applications, and Technical Aspects," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 1160–1192, 2021.

41. J. W. Patton, "Protecting Privacy in Public? Surveillance Technologies and the Value of Public Places," *Ethics and Information Technology*, vol. 2, no. 3, pp. 181–187, 2000.

42. W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge Computing: Vision and Challenges," *IEEE internet of things journal*, vol. 3, no. 5, pp. 637–646, 2016.

43. G. Premsankar, M. Di Francesco, and T. Taleb, "Edge Computing for the Internet of Things: A Case Study," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1275–1284, 2018.

44. P. Porambage, J. Okwuibe, M. Liyanage, M. Ylianttila, and T. Taleb, "Survey on Multi-Access Edge Computing for Internet of Things Realization," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2961–2991, 2018.
45. A. C. Baktir, A. Ozgovde, and C. Ersoy, "How Can Edge Computing Benefit from Software-Defined Networking: A Survey, Use Cases, and Future Directions," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2359–2391, 2017.
46. K. Zhang, Y. Mao, S. Leng, Q. Zhao, L. Li, X. Peng, L. Pan, S. Maharjan, and Y. Zhang, "Energy-Efficient Offloading for Mobile Edge Computing in 5G Heterogeneous Networks," *IEEE Access*, vol. 4, pp. 5896–5907, 2016.
47. P. Mach and Z. Becvar, "Mobile edge computing: A Survey on Architecture and Computation Offloading," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 3, pp. 1628–1656, 2017.
48. F. Samie, V. Tsoutsouras, L. Bauer, S. Xydis, D. Soudris, and J. Henkel, "Computation Offloading and Resource Allocation for Low-Power IoT Edge Devices," in *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*, pp. 7–12, IEEE, 2016.
49. Q. Fan and N. Ansari, "Application Aware Workload Allocation for Edge Computing-based IoT," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 2146–2153, 2018.
50. M. G. R. Alam, M. M. Hassan, M. Z. Uddin, A. Almogren, and G. Fortino, "Autonomic Computation Offloading in Mobile Edge for IoT Applications," *Future Generation Computer Systems*, vol. 90, pp. 149–157, 2019.
51. M. Grieves, "Digital Twin: Manufacturing Excellence Through Virtual Factory Replication," *White paper*, vol. 1, pp. 1–7, 2014.
52. R. Giaffreda, "iCore: A Cognitive Management Framework for the Internet of Things," in *The Future Internet Assembly*, pp. 350–352, Springer, 2013.
53. M. Weyrich and C. Ebert, "Reference Architectures for the Internet of Things," *IEEE Software*, vol. 33, no. 1, pp. 112–116, 2015.
54. A. El Saddik, "Digital Twins: The Convergence of Multimedia Technologies," *IEEE multimedia*, vol. 25, no. 2, pp. 87–92, 2018.
55. X. Sun and N. Ansari, "EdgeIoT: Mobile Edge Computing for the Internet of Things," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 22–29, 2016.
56. M. I. Naas, P. R. Parvedy, J. Boukhobza, and L. Lemarchand, "iFogStor: An IoT Data Placement Strategy for Fog Infrastructure," in *2017 IEEE 1st International Conference on Fog and Edge Computing (ICFEC)*, pp. 97–104, IEEE, 2017.
57. L. Zhao and J. Liu, "Optimal Placement of Virtual Machines for Supporting Multiple Applications in Mobile Edge Networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 6533–6545, 2018.
58. W. Wang, Y. Zhao, M. Tornatore, A. Gupta, J. Zhang, and B. Mukherjee, "Virtual Machine Placement and Workload Assignment for Mobile Edge Computing," in *2017 IEEE 6th International Conference on Cloud Networking (CloudNet)*, pp. 1–6, IEEE, 2017.

59. S. Pasteris, S. Wang, M. Herbster, and T. He, "Service Placement with Provable Guarantees in Heterogeneous Edge Computing Systems," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pp. 514–522, IEEE, 2019.
60. D. Harutyunyan, N. Shahriar, R. Boutaba, and R. Riggio, "Latency-Aware Service Function Chain Placement in 5G Mobile Networks," in *2019 IEEE Conference on Network Softwarization (NetSoft)*, pp. 133–141, IEEE, 2019.
61. J. Xu, L. Chen, and P. Zhou, "Joint Service Caching and Task Offloading for Mobile Edge Computing in Dense Networks," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pp. 207–215, IEEE, 2018.
62. K. Poularakis, J. Llorca, A. M. Tulino, I. Taylor, and L. Tassiulas, "Joint Service Placement and Request Routing in Multi-Cell Mobile Edge Computing Networks," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pp. 10–18, IEEE, 2019.
63. J. An, X. Gui, W. Zhang, and J. Jiang, "Nodes Social Relations Cognition for Mobility-Aware in the Internet of Things," in *2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing*, pp. 687–691, IEEE, 2011.
64. K. M. Alam, M. Saini, and A. El Saddik, "Toward Social Internet of Vehicles: Concept, Architecture, and Applications," *IEEE Access*, vol. 3, pp. 343–357, 2015.
65. A. Ometov, A. Orsino, L. Militano, D. Moltchanov, G. Araniti, E. Olshannikova, G. Fodor, S. Andreev, T. Olsson, A. Iera, *et al.*, "Toward Trusted, Social-Aware D2D Connectivity: Bridging Aross the Technology and Sociality Realms," *IEEE Wireless Communications*, vol. 23, no. 4, pp. 103–111, 2016.
66. S. Andreev, J. Hosek, T. Olsson, K. Johnsson, A. Pyattaev, A. Ometov, E. Olshannikova, M. Gerasimenko, P. Masek, Y. Koucheryavy, *et al.*, "A Unifying Perspective on Proximity-based Cellular-Assisted Mobile Social Networking," *IEEE Communications Magazine*, vol. 54, no. 4, pp. 108–116, 2016.
67. L. Militano, M. Nitti, L. Atzori, and A. Iera, "Enhancing the Navigability in a Social Network of Smart Objects: A Shapley-value based Approach," *Computer Networks*, vol. 103, pp. 1–14, 2016.
68. B. Afzal, M. Umair, G. A. Shah, and E. Ahmed, "Enabling IoT Platforms for Social IoT Applications: Vision, Feature Mapping, and Challenges," *Future Generation Computer Systems*, vol. 92, pp. 718–731, 2019.
69. L. Patrono, L. Atzori, P. Šolić, M. Mongiello, and A. Almeida, "Challenges to be Addressed to Realize Internet of Things Solutions for Smart Environments," 2019.
70. L. Atzori, C. Campolo, B. Da, R. Girau, A. Iera, G. Morabito, and S. Quattropani, "Smart devices in the social loops: Criteria and algorithms for the creation of the social links," *Future Generation Computer Systems*, vol. 97, pp. 327–339, 2019.
71. F. Al-Turjman, "5G-Enabled Devices and Smart-Spaces in Social-IoT: An Overview," *Future Generation Computer Systems*, vol. 92, pp. 732–744, 2019.

72. O. Voutyras, P. Bourelos, S. Gogouvitis, D. Kyriazis, and T. Varvarigou, "Social Monitoring and Social Analysis in Internet of Things Virtual Networks," in *2015 18th International Conference on Intelligence in Next Generation Networks*, pp. 244–251, IEEE, 2015.

73. R. Girau, S. Martis, and L. Atzori, "Lysis: A Platform for IoT Distributed Applications over Socially Connected Objects," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 40–51, 2016.

74. E. Baccarelli, M. Scarpiniti, P. G. V. Naranjo, and L. Vaca-Cardenas, "Fog of Social IoT: When the Fog Becomes Social," *IEEE Network*, vol. 32, no. 4, pp. 68–80, 2018.

75. Y. Lu, S. Maharjan, and Y. Zhang, "Adaptive Edge Association for Wireless Digital Twin Networks in 6G," *IEEE Internet of Things Journal*, vol. 8, no. 22, pp. 16219–16230, 2021.

76. T. Ouyang, Z. Zhou, and X. Chen, "Follow Me at the Edge: Mobility-Aware Dynamic Service Placement for Mobile Edge Computing," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 10, pp. 2333–2345, 2018.

77. B. Gao, Z. Zhou, F. Liu, and F. Xu, "Winning at the Starting Line: Joint Network Selection and Service Placement for Mobile Edge Computing," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pp. 1459–1467, IEEE, 2019.

78. R. Morabito, "Virtualization on Internet of Things Edge Devices With Container Technologies: A Performance Evaluation," *IEEE Access*, vol. 5, pp. 8835–8850, 2017.

79. S. Muralidharan, B. Yoo, and H. Ko, "Designing a Semantic Digital Twin Model for IoT," in *2020 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1–2, IEEE, 2020.

80. S. Kekki, W. Featherstone, Y. Fang, P. Kuure, A. Li, A. Ranjan, D. Purkayastha, F. Jiangping, D. Frydman, G. Verin, *et al.*, "MEC in 5G Networks," *ETSI white paper*, vol. 28, pp. 1–28, 2018.

81. R. Landa, J. T. Araújo, R. G. Clegg, E. Mykoniati, D. Griffin, and M. Rio, "The Large-Scale Geography of Internet Round Trip Times," in *2013 IFIP Networking Conference*, pp. 1–9, IEEE, 2013.

82. T. C. Koopmans and M. Beckmann, "Assignment Problems and the Location of Economic Activities," *Econometrica: journal of the Econometric Society*, pp. 53–76, 1957.

83. A. M. Frieze and J. Yadegar, "On the Quadratic Assignment Problem," *Discrete applied mathematics*, vol. 5, no. 1, pp. 89–98, 1983.

84. W. P. Adams and T. A. Johnson, "Improved Linear Programming-based Lower Bounds for the Quadratic Assignment Problem," *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 16, pp. 43–77, 1994.

85. I. ILOG, "IBM ILOG CPLEX Optimization Studio Documentation," 2020.

86. I. ILOG, "IBM ILOG CPLEX Optimization Studio, Version 12.10.0," *Website http://www.ilog.com/products/cplex*, 2020.

87. C. Marche, L. Atzori, V. Pilloni, and M. Nitti, "How to exploit the Social Internet of Things: Query Generation Model and Device Profiles' Dataset," *Computer Networks*, p. 107248, 2020.

88. Ericsson, "Why IoT changes everything." [Online] (accessed January 24, 2023) `https://www.ericsson.com/`, 2021.

89. L. U. Khan, W. Saad, D. Niyato, Z. Han, and C. S. Hong, "Digital-Twin-Enabled 6G: Vision, Architectural Trends, and Future Directions," *IEEE Communications Magazine*, vol. 60, no. 1, pp. 74–80, 2022.

90. H. X. Nguyen, R. Trestian, D. To, and M. Tatipamula, "Digital Twin for 5G and Beyond," *IEEE Communications Magazine*, vol. 59, no. 2, pp. 10–15, 2021.

91. M. Mashaly, "Connecting the Twins: A Review on Digital Twin Technology & its Networking Requirements," *Procedia Computer Science*, vol. 184, pp. 299–305, 2021.

92. C. Campolo, G. Genovese, A. Iera, and A. Molinaro, "Virtualizing AI at the Distributed Edge Towards Intelligent IoT Applications," *Journal of Sensor and Actuator Networks*, vol. 10, no. 1, p. 13, 2021.

93. "ETSI GS MEC 003 v1.1.1. Mobile Edge Computing (MEC); Framework and Reference Architecture," March 2016.

94. X. Chen, Y. Bi, X. Chen, H. Zhao, N. Cheng, F. Li, and W. Cheng, "Dynamic Service Migration and Request Routing for Microservice in Multi-cell Mobile Edge Computing," *IEEE Internet of Things Journal*, 2022.

95. X. Liu, B. Cheng, and S. Wang, "Availability-Aware and Energy-Efficient Virtual Cluster Allocation Based on Multi-Objective Optimization in Cloud Datacenters," *IEEE Transactions on Network and Service Management*, vol. 17, no. 2, pp. 972–985, 2020.

96. L. Atzori, C. Campolo, B. Da, R. Girau, A. Iera, G. Morabito, and S. Quattropani, "Enhancing Identifier/Locator Splitting Through Social Internet of Things," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2974–2985, 2018.

97. L. Kaufman and F. Broeckx, "An Algorithm for the Quadratic Assignment Problem using Bender's Decomposition," *European Journal of Operational Research*, vol. 2, no. 3, pp. 207–211, 1978.

98. S. Sahni and T. Gonzalez, "P-Complete Approximation Problems," *Journal of the ACM (JACM)*, vol. 23, no. 3, pp. 555–565, 1976.

99. E. L. Lawler, "The Quadratic Assignment Problem," *Management science*, vol. 9, no. 4, pp. 586–599, 1963.

100. "Universal Mobile Telecommunications System (UMTS); Radio Frequency (RF) system scenarios (Release 13)," tech. rep., 3GPP TR 25.942, January 2016.

101. A. Mei and J. Stefa, "SWIM: A Simple Model to Generate Small Mobile Worlds," in *IEEE INFOCOM 2009*, pp. 2106–2113, IEEE, 2009.

102. H. Zhao, J. Wang, F. Liu, Q. Wang, W. Zhang, and Q. Zheng, "Power-Aware and Performance-guaranteed Virtual Machine Placement in the Cloud," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 6, pp. 1385–1400, 2018.

103. Z. Ma, S. Shao, S. Guo, Z. Wang, F. Qi, and A. Xiong, "Container Migration Mechanism for Load Balancing in Edge Network under Power Internet of Things," *IEEE Access*, vol. 8, pp. 118405–118416, 2020.

104. A. Beloglazov and R. Buyya, "Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation of Vrtual Machines in Cloud Data Centers," *Concurrency and Computation: Practice and Experience*, vol. 24, no. 13, pp. 1397–1420, 2012.

105. R. Ford, M. Zhang, M. Mezzavilla, S. Dutta, S. Rangan, and M. Zorzi, "Achieving Ultra-Low Latency in 5G Millimeter Wave Cellular Networks," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 196–203, 2017.

106. M. Fischetti and A. Lodi, "Local Branching," *Mathematical programming*, vol. 98, no. 1, pp. 23–47, 2003.

107. E. Danna, E. Rothberg, and C. Le Pape, "Exploring Relaxation Induced Neighborhoods to Improve MIP Solutions," *Mathematical Programming*, vol. 102, no. 1, pp. 71–90, 2005.

108. X. Xu, X. Liu, Z. Xu, F. Dai, X. Zhang, and L. Qi, "Trust-Oriented IoT Service Placement for Smart Cities in Edge Computing," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4084–4091, 2019.

109. C. Campolo, A. Iera, A. Molinaro, and G. Ruggeri, "MEC Support for 5G-V2X Use Cases through Docker Containers," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, IEEE, 2019.

110. A. Dogra, R. K. Jha, and S. Jain, "A Survey on Beyond 5G Network with the Advent of 6G: Architecture and Emerging Technologies," *IEEE Access*, vol. 9, pp. 67512–67547, 2020.

111. K. Samdanis and T. Taleb, "The Road Beyond 5G: A Vision and Insight of the Key Technologies," *IEEE Network*, vol. 34, no. 2, pp. 135–141, 2020.

112. A. Narayanan, A. S. De Sena, D. Gutierrez-Rojas, D. C. Melgarejo, H. M. Hussain, M. Ullah, S. Bayhan, and P. H. Nardelli, "Key Advances in Pervasive Edge Computing for Industrial Internet of Things in 5G and Beyond," *IEEE Access*, vol. 8, pp. 206734–206754, 2020.

113. D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, D. Niyato, O. Dobre, and H. V. Poor, "6G Internet of Things: A Comprehensive Survey," *IEEE Internet of Things Journal*, 2021.

114. A. Orsino, A. Samuylov, D. Moltchanov, S. Andreev, L. Militano, G. Araniti, and Y. Koucheryavy, "Time-dependent Energy and Resource Management in Mobility-Aware D2D-Empowered 5G Systems," *IEEE Wireless Communications*, vol. 24, no. 4, pp. 14–22, 2017.

115. M. Asad, A. Basit, S. Qaisar, and M. Ali, "Beyond 5G: Hybrid End-to-End Quality of Service Provisioning in Heterogeneous IoT Networks," *IEEE Access*, vol. 8, pp. 192320–192338, 2020.

116. X. Li, R. Zhang, and L. Hanzo, "Optimization of Visible-Light Optical Wireless Systems: Network-Centric Versus User-Centric Designs," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 3, pp. 1878–1904, 2018.

117. V. Ranjbar, A. Girycki, M. A. Rahman, S. Pollin, M. Moonen, and E. Vinogradov, "Cell-free mMIMO Support in the O-RAN Architecture: A PHY Layer Perspective for 5G and Beyond Networks," *IEEE Communications Standards Magazine*, vol. 6, no. 1, pp. 28–34, 2022.

118. W. Sun, M. Ai, X. Duan, and M. Shu, "Research on 6G Network Adaptability Index System," in *2022 IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, pp. 417–422, IEEE, 2022.

119. E. Carter, P. Adam, D. Tsakis, S. Shaw, R. Watson, and P. Ryan, "Enhancing Pedestrian Mobility in Smart Cities using Big Data," *Journal of Management Analytics*, vol. 7, no. 2, pp. 173–188, 2020.

120. L. Liu, A. Biderman, and C. Ratti, "Urban Mobility Landscape: Real Time Monitoring of Urban Mobility Patterns," in *Proceedings of the 11th International Conference on Computers in Urban Planning and Urban Management*, pp. 1–16, Citeseer, 2009.

121. A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, and R. Banchs, "Urban Cycles and Mobility Patterns: Exploring and Predicting Trends in a Bicycle-based Public Transport System," *Pervasive and Mobile Computing*, vol. 6, no. 4, pp. 455–466, 2010.

122. S. Gao, "Spatio-Temporal Analytics for Exploring Human Mobility Patterns and Urban Dynamics in the Mobile Age," *Spatial Cognition & Computation*, vol. 15, no. 2, pp. 86–114, 2015.

123. A. Sedighi, S. M. Ulman, and M. A. Nussbaum, "Information Presentation Through a Head-Worn Display ("Smart Glasses") has a Smaller Influence on the Temporal Structure of Gait Variability during Dual-Task Gait Compared to Handheld Displays (Paper-based System and Smartphone)," *PLoS one*, vol. 13, no. 4, p. e0195106, 2018.

124. S. Y. Han, M.-H. Tsou, E. Knaap, S. Rey, and G. Cao, "How Do Cities Flow in an Emergency? Tracing Human Mobility Patterns During a Natural Disaster with Big Data and Geospatial Data Science," *Urban Science*, vol. 3, no. 2, p. 51, 2019.

125. Q. Wang and J. E. Taylor, "Patterns and Limitations of Urban Human Mobility Resilience under the Influence of Multiple Types of Natural Disaster," *PLoS one*, vol. 11, no. 1, p. e0147299, 2016.

126. J. Novák and L. Sỳkora, "A City in Motion: Time-Space Activity and Mobility Patterns of Suburban Inhabitants and the Structuration of the Spatial Organization of the Prague Metropolitan Area," *Geografiska Annaler: Series B, Human Geography*, vol. 89, no. 2, pp. 147–168, 2007.

127. L. Atzori, A. Iera, and G. Morabito, "SIoT: Giving a Social Structure to the Internet of Things," *IEEE Communications Letters*, vol. 15, no. 11, pp. 1193–1195, 2011.

128. A. Hamrouni, A. Khanfor, H. Ghazzai, and Y. Massoud, "Context-Aware Service Discovery: Graph Techniques for IoT Network Learning and Socially Connected Objects," *IEEE Access*, 2022.

129. L. Atzori, C. Campolo, B. Da, A. Iera, G. Morabito, P. P. Esnault, and S. Quattropani, "Social-IoT Enabled Identifier/Locator Splitting: Concept, Architecture, and Performance Evaluation," in *2018 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, 2018.

130. N. Chukhno, O. Chukhno, S. Pizzi, A. Molinaro, A. Iera, and G. Araniti, "Unsupervised Learning for D2D-Assisted Multicast Scheduling in mmWave Networks," in *2021 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–6, IEEE, 2021.

131. N. Chukhno, S. Trilles, J. Torres-Sospedra, A. Iera, and G. Araniti, "D2D-based Cooperative Positioning Paradigm for Future Wireless Systems: A survey," *IEEE Sensors Journal*, 2021.

132. A. Ali, O. Galinina, and S. Andreev, "System-level Dynamics of Highly Directional Distributed Networks," *IEEE Wireless Communications Letters*, vol. 10, no. 7, pp. 1523–1527, 2021.

133. M. Lecci, F. Chiariotti, M. Drago, A. Zanella, and M. Zorzi, "Temporal Characterization of XR Traffic with Application to Predictive Network Slicing," *arXiv preprint arXiv:2201.07043*, 2022.

134. F. d. Carvalho, L. Morgado, and R. Machado, "eXtended New Reality," *InnovAction*, vol. 2020, no. 5, pp. 46–55, 2020.

135. H. H. H. Mahmoud, A. A. Amer, and T. Ismail, "6G: A Comprehensive Survey on Technologies, Applications, Challenges, and Research Problems," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 4, p. e4233, 2021.

136. F. Gabin, G. Teniou, N. Leung, and I. Varga, "5g multimedia standardization," *Journal of ICT Standardization*, vol. 6, no. 1, pp. 117–136, 2018.

137. K. M. Stanney, H. Nye, S. Haddad, K. S. Hale, C. K. Padron, and J. V. Cohn, "eXtended Reality (XR) Environments," *Handbook of human factors and ergonomics*, pp. 782–815, 2021.

138. Huawei iLab, "Cloud VR Network Solution," *Cloud VR White Paper*, 2018.

139. Huawei iLab, "Cloud VR User Experience and Evaluation," *Cloud VR White Paper*, 2019.

140. A. U. Rahman, G. Ghatak, and A. De Domenico, "An Online Algorithm for Computation Offloading in Non-Stationary Environments," *IEEE Communications Letters*, vol. 24, no. 10, pp. 2167–2171, 2020.

141. N. Chen, R. Gurlek, D. Lee, and H. Shen, "Can Customer Arrival Rates be Modelled by Sine Waves? Improving Service Operations with Queuing Data Analytics," *Improving service operations with queuing data analytics (January 16, 2021)*, 2021.

142. K. K. Leung, W. A. Massey, and W. Whitt, "Traffic Models for Wireless Communication Networks," *IEEE Journal on selected areas in Communications*, vol. 12, no. 8, pp. 1353–1364, 1994.

143. D. Qiu and R. Srikant, "Modeling and Performance Analysis of BitTorrent-like Peer-to-Peer Networks," *ACM SIGCOMM computer communication review*, vol. 34, no. 4, pp. 367–378, 2004.

144. A. Pyattaev, O. Galinina, S. Andreev, M. Katz, and Y. Koucheryavy, "Understanding Practical Limitations of Network Coding for Assisted Proximate Communication," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 2, pp. 156–170, 2014.

145. C. Sergiou and V. Vassiliou, "Estimating Maximum Traffic Volume in Wireless Sensor Networks using Fluid Dynamics Principles," *IEEE Communications Letters*, vol. 17, no. 2, pp. 257–260, 2013.

146. C. Perfecto, M. S. Elbamby, J. Park, J. Del Ser, and M. Bennis, "Mobile XR over 5G: A Way Forward with mmWaves and Edge," *arXiv preprint arXiv:1905.04599*, 2019.

147. J. Park, S. Samarakoon, M. Bennis, and M. Debbah, "Wireless Network Intelligence at the Edge," *Proceedings of the IEEE*, vol. 107, no. 11, pp. 2204–2239, 2019.

148. M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, *et al.*, "Above the Clouds: a Berkeley View of Cloud Computing. 2009," *University of California at Berkeley*, 2016.

149. J. Leirpoll, D. Osborn, P. Murphy, and A. Edwards, "VR Video Editing," in *The Cool Stuff in Premiere Pro*, pp. 757–792, Springer, 2017.

150. M. Ghoshal, Z. J. Kong, Q. Xu, Z. Lu, S. Aggarwal, I. Khan, Y. Li, Y. C. Hu, and D. Koutsonikolas, "An In-Depth Study of Uplink Performance of 5G mmWave Networks," in *Proceedings of the ACM SIGCOMM Workshop on 5G and Beyond Network Measurements, Modeling, and Use Cases*, pp. 29–35, 2022.

151. A. I. Sulyman, A. T. Nassar, M. K. Samimi, G. R. MacCartney, T. S. Rappaport, and A. Alsanie, "Radio Propagation Path Loss Models for 5G Cellular Networks in the 28 GHz and 38 GHz Millimeter-Wave Bands," *IEEE Communications Magazine*, vol. 52, no. 9, pp. 78–86, 2014.

152. H. Wymeersch and G. Seco-Granados, "Radio Localization and Sensing–Part II: State-of-the-art and Challenges," *IEEE Communications Letters*, 2022.

153. Y. He, D. Wang, F. Huang, and R. Zhang, "An MEC-Enabled Framework for Task Offloading and Power Allocation in NOMA Enhanced ABS-Assisted VANETs," *IEEE Communications Letters (Early Access)*, 2022.

154. A. Forde and J. Daniel, "Pedestrian Walking Speed at Un-Signalized Midblock Crosswalk and its Impact on Urban Street Segment Performance," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 8, no. 1, pp. 57–69, 2021.

155. Oculus, "Introducing Oculus Air Link, a Wireless Way to Play PC VR Games on Oculus Quest 2, Plus Infinite Office Updates, Support for 120 Hz on Quest 2, and More." [Online] (accessed January 24, 2023) `https://www.oculus.com/blog/`, 2021.

156. C. C. Sekhar, S. S. Ch, and G. N. Rao, "Future Reality is Immersive Reality," *International Journal of Recent Technology and Engineering*, vol. 7, no. 4, pp. 302–309, 2018.

157. F. Hu, Y. Deng, H. Zhou, T. H. Jung, C. B. Chae, and A. H. Aghvami, "A Vision of an XR-Aided Teleoperation System toward 5G/B5G," *IEEE Communications Magazine*, vol. 59, no. 1, pp. 34–40, 2021.

158. J. H. Hollman, R. H. Brey, R. A. Robb, T. J. Bang, and K. R. Kaufman, "Spatiotemporal Gait Deviations in a Virtual Reality Environment," *Gait & posture*, vol. 23, no. 4, pp. 441–444, 2006.

159. F. Menegoni, G. Albani, M. Bigoni, L. Priano, C. Trotti, M. Galli, and A. Mauro, "Walking in an Immersive Virtual Reality," *Annual Review of Cybertherapy and Telemedicine 2009*, pp. 72–76, 2009.

160. K. Pfeuffer, M. J. Geiger, S. Prange, L. Mecke, D. Buschek, and F. Alt, "Behavioural Biometrics in VR: Identifying People from Body Motion and Relations in Virtual Reality," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019.

161. Z. Y. Chan, A. J. MacPhail, I. P. Au, J. H. Zhang, B. M. Lam, R. Ferber, and R. T. Cheung, "Walking with Head-Mounted Virtual and Augmented Reality Devices: Effects on Position Control and Gait Biomechanics," *PLoS one*, vol. 14, no. 12, p. e0225972, 2019.

162. M. A. Bühler and A. Lamontagne, "Circumvention of Pedestrians while Walking in Virtual and Physical Environments," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 9, pp. 1813–1822, 2018.

163. P. Wodarski, J. Jurkojć, J. Polechoński, A. Bieniek, M. Chrzan, R. Michnik, and M. Gzik, "Assessment of Gait Stability and Preferred Walking Speed in Virtual Reality," *Acta of bioengineering and biomechanics*, vol. 22, no. 1, 2020.

164. E. Al-Yahya, H. Dawes, L. Smith, A. Dennis, K. Howells, and J. Cockburn, "Cognitive Motor Interference While Walking: A Systematic Review and Meta-Analysis," *Neuroscience & Biobehavioral Reviews*, vol. 35, no. 3, pp. 715–728, 2011.

165. T. Mustonen, M. Berg, J. Kaistinen, T. Kawai, and J. Häkkinen, "Visual Task Performance using a Monocular See-through Head-Mounted Display (HMD) While Walking," *Journal of Experimental Psychology: Applied*, vol. 19, no. 4, p. 333, 2013.

166. I. F. Akyildiz and H. Guo, "Wireless eXtended Reality (XR): Challenges and New Research Directions," *ITU J. Future Evol. Technol*, vol. 3, pp. 1–15, 2022.

167. M. Giordani, 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.

168. M. Agiwal, H. Kwon, S. Park, and H. Jin, "A Survey on 4G-5G Dual Connectivity: Road to 5G Implementation," *IEEE Access*, vol. 9, pp. 16193–16210, 2021.

169. S. Ranjan, P. Jha, A. Karandikar, and P. Chaporkar, "A Flexible IAB Architecture for Beyond 5G Network," *arXiv preprint arXiv:2201.13029*, 2022.

170. S. Chandrashekar, A. Maeder, C. Sartori, T. Höhne, B. Vejlgaard, and D. Chandramouli, "5G Multi-RAT Multi-Connectivity Architecture," in *2016 IEEE International Conference on Communications Workshops (ICC)*, pp. 180–186, IEEE, 2016.

171. M.-T. Suer, C. Thein, H. Tchouankem, and L. Wolf, "Multi-Connectivity as an Enabler for Reliable Low Latency Communications—An Overview," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 156–169, 2019.

172. N. Chukhno, O. Chukhno, S. Pizzi, A. Molinaro, A. Iera, and G. Araniti, "Efficient Management of Multicast Traffic in Directional mmWave Networks," *IEEE Transactions on Broadcasting*, 2021.

173. N. Chukhno, O. Chukhno, D. Moltchanov, A. Molinaro, Y. Gaidamaka, K. Samouylov, Y. Koucheryavy, and G. Araniti, "Optimal Multicasting in Millimeter Wave 5G NR with Multi-beam Directional Antennas," *IEEE Transactions on Mobile Computing*, 2021.

174. O. Chukhno, N. Chukhno, O. Galinina, S. Andreev, Y. Gaidamaka, K. Samouylov, and G. Araniti, "A Holistic Assessment of Directional Deafness in mmWave-based Distributed 3D Networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 9, pp. 7491–7505, 2022.

175. M. Gerasimenko, D. Moltchanov, M. Gapeyenko, S. Andreev, and Y. Koucheryavy, "Capacity of Multi-Connectivity mmWave Systems with Dynamic Blockage and Directional Antennas," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3534–3549, 2019.

176. S. Tripathi, N. V. Sabu, A. K. Gupta, and H. S. Dhillon, "Millimeter-wave and Terahertz Spectrum for 6G Wireless," in *6G Mobile Wireless Networks*, pp. 83–121, Springer, 2021.

177. A. Hyre, G. Harris, J. Osho, M. Pantelidakis, K. Mykoniatis, and J. Liu, "Digital Twins: Representation, Replication, Reality, and Relational (4Rs)," *Manufacturing Letters*, vol. 31, pp. 20–23, 2022.

178. C. Li, Y. Zhang, X. Gao, and Y. Luo, "Energy-Latency Tradeoffs for Edge Caching and Dynamic Service Migration based on DQN in Mobile Edge Computing," *Journal of Parallel and Distributed Computing*, vol. 166, pp. 15–31, 2022.

179. P. Bellavista, A. Corradi, L. Foschini, and D. Scotece, "Differentiated Service/Data Migration for Edge Services Leveraging Container Characteristics," *IEEE Access*, vol. 7, pp. 139746–139758, 2019.

180. F. Tütüncüoğlu and G. Dán, "Optimal Pricing for Service Caching and Task Offloading in Edge Computing," in *2022 17th Wireless On-Demand Network Systems and Services Conference (WONS)*, pp. 1–8, IEEE, 2022.

181. A. Reznik, L. Murillo, Y. Fang, *et al.*, "Cloud RAN and MEC: A perfect pairing," *ETSI White paper*, 2018.

182. Huawei Technologies Co., Ltd., "Cloud VR Bearer Networks: Huawei iLab VR Technology White Paper," *Huawei, Tech. Rep.*, 2017.

183. F. Farina, D. Fontanelli, A. Garulli, A. Giannitrapani, and D. Prattichizzo, "Walking Ahead: The Headed Social Force Model," *PloS one*, vol. 12, no. 1, p. e0169734, 2017.

184. 3GPP, "Study on Channel Model for Frequencies from 0.5 to 100 GHz (Rel. 14)," 3GPP TR 38.901 V14.1.1, July 2017.

185. S. Andreev, O. Galinina, A. Pyattaev, J. Hosek, P. Masek, H. Yanikomeroglu, and Y. Koucheryavy, "Exploring Synergy between Communications, Caching, and Computing in 5G-grade Deployments," *IEEE Communications Magazine*, vol. 54, no. 8, pp. 60–69, 2016.

186. M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A Tutorial on Beam Management for 3GPP NR at mmWave Frequencies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 173–196, 2018.

187. V. Petrov, M. Gapeyenko, S. Paris, A. Marcano, and K. I. Pedersen, "Extended Reality (XR) over 5G and 5G-Advanced New Radio: Standardization, Applications, and Trends," *arXiv preprint arXiv:2203.02242*, 2022.

188. F. Hu, Y. Deng, W. Saad, M. Bennis, and A. H. Aghvami, "Cellular-connected Wireless Virtual Reality: Requirements, Challenges, and Solutions," *IEEE Communications Magazine*, vol. 58, no. 5, pp. 105–111, 2020.

189. W. Zhang, "Cloud X: New Services in 5G Era," *Huawei, Tech. Rep.*, 2018.

190. M. de Jager, F. J. Weissing, P. M. Herman, B. A. Nolet, and J. van de Koppel, "Lévy Walks Evolve Through Interaction Between Movement and Environmental Complexity," *Science*, vol. 332, no. 6037, pp. 1551–1553, 2011.

191. G. M. Viswanathan, E. Raposo, and M. Da Luz, "Lévy Flights and Superdiffusion in the Context of Biological Encounters and Random Searches," *Physics of Life Reviews*, vol. 5, no. 3, pp. 133–150, 2008.

192. H. Murakami, C. Feliciani, and K. Nishinari, "Lévy Walk Process in Self-Organization of Pedestrian Crowds," *Journal of the Royal Society Interface*, vol. 16, no. 153, p. 20180939, 2019.

193. G. M. Viswanathan, S. V. Buldyrev, S. Havlin, M. Da Luz, E. Raposo, and H. E. Stanley, "Optimizing the Success of Random Searches," *nature*, vol. 401, no. 6756, pp. 911–914, 1999.

194. A. M. Reynolds and C. J. Rhodes, "The Lévy Flight Paradigm: Random Search Patterns and Mechanisms," *Ecology*, vol. 90, no. 4, pp. 877–887, 2009.

195. M. E. Wosniack, M. C. Santos, E. P. Raposo, G. M. Viswanathan, and M. G. Da Luz, "The Evolutionary Origins of Lévy Walk Foraging," *PLoS computational Biology*, vol. 13, no. 10, p. e1005774, 2017.

196. A. M. Reynolds, "Current Status and Future Directions of Lévy Walk Research," *Biology Open*, vol. 7, no. 1, p. bio030106, 2018.

197. K. Zhang, J. Cao, S. Maharjan, and Y. Zhang, "Digital Twin Empowered Content Caching in Social-Aware Vehicular Edge Networks," *IEEE Transactions on Computational Social Systems*, 2021.

198. M. Vaezi, K. Noroozi, T. D. Todd, D. Zhao, G. Karakostas, H. Wu, and X. Shen, "Digital Twins from a Networking Perspective," *IEEE Internet of Things Journal*, 2022.

199. A. Rasheed, O. San, and T. Kvamsdal, "Digital Twin: Values, Challenges and Enablers from a Modeling Perspective," *Ieee Access*, vol. 8, pp. 21980–22012, 2020.

200. E. Bastug, M. Bennis, M. Médard, and M. Debbah, "Toward Interconnected Virtual Reality: Opportunities, Challenges, and Enablers," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 110–117, 2017.

201. A. L. Gardony, R. W. Lindeman, and T. T. Brunyé, "Eye-Tracking for Human-Centered Mixed Reality: Promises and Challenges," in *Optical Architectures for Displays and Sensing in Augmented, Virtual, and Mixed Reality (AR, VR, MR)*, vol. 11310, pp. 230–247, SPIE, 2020.

202. A. C. Squicciarini, M. Shehab, and F. Paci, "Collective Privacy Management in Social Networks," in *Proceedings of the 18th International Conference on World Wide Web*, pp. 521–530, 2009.

Olga Chukhno is an Early Stage Researcher within H2020 MSCA ITN/EJD A-WEAR project and a PhD student at Mediterranea University of Reggio Calabria, Italy and Tampere University, Finland. She received M.Sc. (2019) in Fundamental Informatics and Information Technologies and B.Sc. (2017) in Business Informatics from RUDN University, Russia. Her current research interests include wireless communications, social networking, edge computing, and wearable applications.

The rise of the intelligent information world presents significant challenges for the telecommunication industry in meeting the service-level requirements of future applications and incorporating societal and behavioral awareness into the Internet of Things (IoT) objects. Social Digital Twins (SDTs), or Digital Twins augmented with social capabilities, have the potential to revolutionize digital transformation and meet the connectivity, computing, and storage needs of IoT devices in dynamic Fifth-Generation (5G) and Beyond Fifth-Generation (B5G) networks.

This research focuses on enabling dynamic social-aware B5G networking. The main contributions of this work include (i) the design of a reference architecture for the orchestration of SDTs at the network edge to accelerate the service discovery procedure across the Social Internet of Things (SIoT); (ii) a methodology to evaluate the highly dynamic system performance considering jointly communication and computing resources; (iii) a set of practical conclusions and outcomes helpful in designing future digital twin-enabled B5G networks.

Specifically, we propose an orchestration for SDTs and an SIoT-edge framework aligned with the Multi-access Edge Computing (MEC) architecture ratified by the European Telecommunications Standards Institute (ETSI). We formulate the optimal placement of SDTs as a Quadratic Assignment Problem (QAP) and propose a graph-based approximation scheme considering the different types of IoT devices, their social features, mobility patterns, and the limited computing resources of edge servers. We also study the appropriate intervals for re-optimizing the SDT deployment at the network edge. The results demonstrate that accounting for social features in SDT placement offers considerable improvements in the SIoT browsing procedure. Moreover, recent advancements in wireless communications, edge computing, and intelligent device technologies are expected to promote the growth of SIoT with pervasive sensing and computing capabilities, ensuring seamless connections among SIoT objects.

We then offer a performance evaluation methodology for eXtended Reality (XR) services in edge-assisted wireless networks and propose a fluid approximation to characterize the XR content evolution. The approach captures the time and space dynamics of the content distribution process during its transient phase, including time-varying loads, which are influenced by arrival, transition, and departure processes. We examine the effects of XR user mobility on both communication and computing. The results demonstrate that both communication and computing planes are the key barriers to meeting the requirement for real-time transmissions. Furthermore, due to the trend toward immersive, interactive, and contextualized experiences, new use cases affect user mobility and, therefore, system performance.

Beyond Fifth-Generation, Social Internet of Things, Digital Twining, Wireless Networks, Edge Computing

Tampere University