Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 257 (2025) 599–606

Procedia Computer Science

www.elsevier.com/locate/procedia

The 16th International Conference on Ambient Systems, Networks and Technologies (ANT)
April 22-24, 2025, Patras, Greece

AI-Driven Load Balancing in Paravirtualized 5G Networks: Architecture, Implementation, and Future Prospects

Pierre Mobara[a], Justin Moskolaï Ngossaha[a,*], Rodrigue Djeumen Djatcha[a], Igor Tchappi[b], Samuel Bowong Tsakou[a]

[a]*Department of Mathematics and Computer Science, Faculty of Science, University of Douala, PO Box 24 157, Cameroon.*
[b]*FINATRAX, SnT, University of Luxembourg, 6 Rue Richard Coudenhove-Kalergi, 1359 Kirchberg Luxembourg*

Abstract

With the rapid proliferation of services and the challenges faced by existing deployed technologies, a need arose for an advanced solution. This need culminated in the development of 5G technology to address the limitations of previous generations. However, despite being a relatively recent innovation and still in its pre-deployment phase, 5G networks already face significant load management issues at their peripheries due to the increasing number of services and applications. How can the integration of Artificial Intelligence (AI), particularly using supervised methods, optimize workload distribution in 5G networks? This paper presents a comprehensive literature review on 5G and AI, a case study of a paravirtualized 5G network enhanced with AI agents and explores the opportunities arising from the integration of these two technologies in bandwidth management.

Keywords: 5G Networks; Artificial Intelligence (AI); Load Balancing; Paravirtualization; Quality of Service (QoS)

1. Introduction

Recent statistics highlight substantial growth in both the number of mobile subscribers and the demand for bandwidth, video, and data services. Mobile data traffic is projected to grow at an annual rate of 46% between 2017

* Corresponding author.
E-mail address: moskolaijust2000@yahoo.fr

10.1016/j.procs.2025.03.077

and 2022 [1, 18]. 5G's share of mobile data traffic will reach an estimated 34 percent by the end of 2024, an increase from 25 percent at the end of 2023. This share is forecast to grow to 80 percent in 2030 [2]. Furthermore, mobile video is anticipated to constitute nearly 79% of mobile data traffic, with an annual growth rate of 35% expected until 2025 [3]. Given this exponential increase in demand, it is unlikely that a single access technology can meet these projections. Instead, high-speed access technologies must continue to evolve, enhancing system performance and capabilities. Additionally, new communication modes will need to complement existing technologies to address emerging and challenging use cases in mobile networks.

5G represents the next evolution in mobile communication, offering high bandwidth, wider coverage, and ultra-low latency [1]. This technology enables the deployment of novel services and applications, including the Internet of Things (IoT), augmented and virtual reality, smart city networks, autonomous vehicles, and industrial automation. Beyond supporting advanced applications, 5G networks provide new revenue opportunities for service providers through the integration of technologies such as data analytics, machine learning, and artificial intelligence [1, 16]. Structured network systems exhibit two critical properties: dynamism and large-scale operation. In such systems, effective load balancing is essential to maintain optimal quality of service (QoS). To the best of our knowledge, no existing research comprehensively addresses the simultaneous balancing of storage load and index management load [2]. This study proposes a solution to address the research gap by optimizing workload distribution across the network to enhance efficiency and performance. It systematically considers key factors influencing load balancing, such as traffic dynamics, computational capacity, and service demands, ensuring adaptive resource allocation. Additionally, the approach minimizes maintenance costs in dynamic environments by leveraging automation and reducing the need for manual intervention.

To achieve this, the paper is organized as follows: Section 2 presents a comprehensive literature review on 5G, load balancing, artificial intelligence, and its applications. Section 3 outlines the proposed methodological approach and the design of an AI-based architecture. Finally, Section 4 discusses the validation of the proposed framework and its associated results.

2. Literature review

2.1. 5G Networks

As its name suggests, 5G refers to the fifth generation of mobile communication standards. For the first time, it represents a globally accepted international standard, transcending regional and continental differences. 5G follows the fourth generation (4G), which was still being deployed in regions such as France as recently as 2019, while coexisting with legacy 2G and 3G networks. The evolution of mobile networks has been transformative: 2G enabled voice communication in portable devices, 3G introduced mobile internet access, and 4G significantly enhanced data speeds. 5G builds upon these advancements on a much larger scale, providing unprecedented connectivity capabilities that extend beyond traditional devices to include cars, cities, and industrial systems [4].

5G technology introduces a transformative shift in wireless communication by delivering ultra-high data speeds, minimal latency, and enhanced reliability, enabling advancements in autonomous vehicles, smart cities, and industrial automation. According to the National Frequency Agency (ANFR), it not only alleviates 4G network congestion but also drives innovation in sectors like e-health, where it supports telemedicine and telesurgery with unprecedented precision. As a catalyst for digital transformation, 5G is set to revolutionize emerging industries, reinforcing its role as a key enabler of future connectivity and technological progress [5].

Adapted mobile technologies in 5G provide seamless connectivity between people, devices, data, and applications, forming the backbone of intelligent networks. Beyond these advancements, 5G scales connectivity to unprecedented levels, enabling integration across sectors, including vehicles, cities, and industrial facilities. Its distinctive features—such as ultra-high speeds, minimal latency, and enhanced reliability—are set to revolutionize existing markets and drive innovation. Figure 1 illustrates the anticipated architecture of fifth-generation mobile networks when combined with Cloud infrastructure and Software-Defined Networking (SDN), which serves as the network's core component [4].

The 5G network architecture is built upon Cloud RAN (C-RAN) and SDN. In addition to the transport network and core network, the architecture incorporates an application cloud and an SDN controller cloud. One of the most

significant contributions of SDN is its ability to provide flexibility in network management. As Christian Bonnet, a researcher at Eurecom specializing in networks and telecommunications, explains: *"In a traditional architecture without SDN, routing rules are predefined and difficult to modify."* He further states that *"SDN introduces intelligence into the network, enabling the dynamic modification of these rules when necessary."*

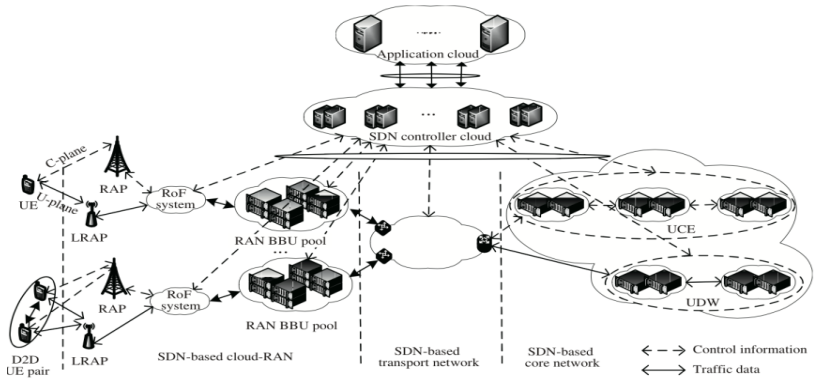

Fig. 1. 5G Architecture [4]

This enhanced flexibility allows for the segmentation of the network through the creation of customized rules that isolate data paths for specific services, ensuring efficient resource allocation and improved network performance [5].

Table 1. The 8 5G performance indicators established by the ITU - IMT-2020 [6]

Performances / Generations	4G	5G
Maximum flow (Gbbs)	1	20
Flow rate seen by the user (Mbbs)	10	100
Spectral efficiency	1x	3x
Speed (Km/h)	350	500
Latency (ms)	10	1
Quantity of connected objects /Km2	105	106
Network energy efficiency	1x	100x
Flow over an area (Mbbs/m2)	0.1	10

Table 1 presents the key performance indicators of 5G in comparison to 4G, highlighting the significant advancements introduced by 5G in terms of speed, the number of connected devices, and latency—critical metrics in telecommunications. 5G networks will enable a wide range of transformative applications. Additionally, 5G will provide the robust connectivity required for Artificial Intelligence (AI) applications, enhancing the intelligence and efficiency of networks. These advancements will facilitate critical innovations such as 5G network slicing, cloud interconnection, virtualization, and the development of self-organizing networks. Figure 2 illustrates the benefits of 5G adoption, with the most significant advantages being industrial automation, Gigabit-level speeds, the development of smart cities, and enhanced access to cloud-based services

2.2. Load Balancing in 5G Networks and Paravirtualization

Load balancing is the strategic allocation of system resources, including network bandwidth, storage capacity, and CPU power, to optimize overall performance. In structured peer-to-peer (P2P) systems, load refers to the resource burden on individual peers, which can be categorized into index management load and storage load

[8]. Index management load balancing pertains to bandwidth utilization over time for routing messages through a peer, where each peer has a finite routing capacity, and the overhead is defined as the difference between the actual load and this capacity. Conversely, storage load balancing ensures the equitable distribution of stored objects across peers while maintaining independence from key distribution. Each peer contributes a defined storage capacity, and storage overheadarises when the total size of stored objects surpasses this capacity, potentially impacting system efficiency. Table 2 summarizes various studies on load balancing, highlighting diverse research objectives and approaches. This assessment provided valuable insights and served as a foundation for this work, drawing inspiration from recent advancements in the field.

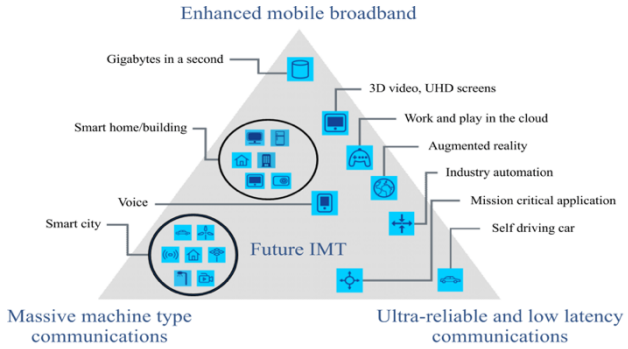

Fig. 2. 5G applications [7]

2.3. Application of artificial intelligence in 5G networks

Thanks to the vast amount of data generated by telecommunications networks and the capabilities of AI predictive models, numerous applications of AI in network management have emerged [17]. Notable examples include:

- Predictive Maintenance: AI can identify and even anticipate technical failures, enabling proactive interventions.
- Incident Response: Upon detecting or predicting an incident, AI can evaluate all possible solutions, simulate their deployment, and measure their effects. This allows human operators to make informed decisions by selecting the most effective solution.
- Virtual Assistants: AI-driven virtual assistants efficiently handle routine customer requests, while more complex issues can be escalated to human support.
- Network Optimization: AI optimizes services by analyzing various parameters that influence the quality of the user experience.
- Prediction and Anticipation: AI enables refined and personalized service offerings for customers, while also predicting and anticipating their needs.

Following the literature review, existing research highlights the significant potential of Artificial Intelligence (AI) in addressing the challenges of load balancing in 5G networks while emphasizing the need for further investigation in this area. Implementing load balancing in paravirtualized 5G networks using AI is a complex task requiring careful consideration. Node congestion remains a critical issue in modern networks due to the proliferation of emerging services. To address this challenge, several approaches were explored, and the findings indicate that combining AI with load balancing presents a scalable solution. 5G technology, equipped with Multiple Input Multiple Output (MIMO) antennas, simplifies network architectures and achieves throughput rates of up to 10 Gbps.

AI, with its computational power and learning algorithms, facilitates intelligent decision-making through supervised models, enabling effective node decongestion. Meanwhile, paravirtualization reduces reliance on physical hardware, streamlining maintenance and lowering network installation costs. This paper focuses on the integration of 5G, paravirtualization, and AI. By leveraging the combined potential of these technologies, workloads at the 5G network edge can be efficiently balanced, ensuring enhanced service availability and optimized network performance.

3. Proposed methodological framework

3.1. Methodological approach

For this research, an exploratory approach was adopted, which proved to be the most appropriate choice for several reasons. The study of artificial intelligence (AI) management is relatively recent, and existing documentation emphasizes its ongoing evolution and anticipated changes in the coming years. As a result, an explanatory approach, which would require older and more extensive datasets, was not deemed suitable. The primary focus of this research was to highlight the contribution of AI to load management in paravirtualized 5G networks.

To collect and analyze the required data, a corporate network environment was set up, connected to a 5G router. A Synology module hosting multiple servers employing paravirtualization was installed within this environment.

The paravirtualized system consists of two main components: server networks and service applications. This setup required the implementation of two AI models: (i) an unsupervised model, responsible for managing optimal communication between the servers; (ii) a supervised model, which ensures effective load distribution based on a monitoring dashboard that evaluates the services running on each server.

Table 2. Summary of load balance studies

Authors	Studies	Goals	Results
Karl Aberer and Manfred Hauswirth, 2022 [10]	The quest for balancing peer load in structured peer-to-peer systems	Develop an equity solution between the entities of a network	Based on distributed application architecture that partitions tasks or workloads
Agrawal et al. 2024 [11]	Shortest path routing in arbitrary networks	Optimize communication in a network	real-time adaptive traffic with signal control algorithm
Los et al. 2023 [12]	Simple load balancing for distributed hash tables	Provide a simple load management solution based on hash tables	Traffic managed by the distributed hash function
Bienkowski et al., .2005 [13]	Dynamic Load Balancing in Distributed Hash Tables	Provide a dynamic load management solution based on hash tables	Traffic managed by the distributed hash function
Bhagwan, R., Savage, S., Voelker, G.M (2003) [14]	Understanding availability	Develop a workload management solution based on monitoring the availability of equipment	Based on +40% rate as overload indicator
Nogueira J., 2014 [15]	Large-scale, decentralized multicasting at the application-level Infrastructure	Allow decentralized task management to avoid overloaded nodes	*application multicast routing infrastructure*

3.2. Architecture design based on artificial intelligence

Figure 3 illustrates the proposed architecture of the 5G network load management model leveraging AI. The architecture is divided into three main layers: (i) *the 5G Router:* This component integrates the model into a 5G environment, ensuring compatibility and connectivity; (ii) *The Synology Layer:* Serving as the core of the architecture, this layer comprises two key elements such as Paravirtual Servers which handle network workloads and AI Agents, responsible for monitoring communication between servers and assessing their saturation levels to optimize load distribution; (iii) *The User Plane:* This layer provides access to the services hosted on the paravirtual servers. An AI agent dynamically intervenes to direct requests to the least-loaded server, thereby improving overall efficiency and performance.

The solution proposed outlines the interactions between the various modules involved in the communication chain. A central component of this architecture is the AI agent, which plays a critical role in supervising all servers and services to ensure balanced workload distribution across the servers.

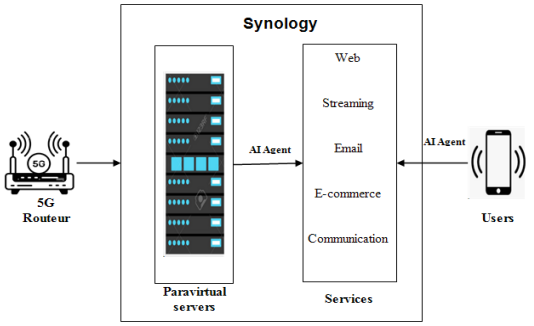

Fig. 3. AI-based load management architecture

4. Validation and discussion

4.1. Case study

The proposed architecture for implementing a virtual load management system integrates three key technologies: Cloud, SDN, and Network Functions Virtualization (NFV). Together, these elements form the foundation of the 5G core network, which is interconnected with the operator's network. To develop this architecture, a comprehensive study was conducted on the three concepts underlying the virtual core network. Figure 5 below provides a clear depiction of the components and their interactions in the creation of a virtualized network. The following Figure 4 provides a detailed description of the components, and their interactions as follows:

- Customer Portal: The administrator uses the customer portal to configure and manage corporate policies, services, and network infrastructures. Each enterprise maintains its own virtual network infrastructure.
- Moderator: The moderator introduces the selected services and service chains to enterprise customers. It also summarizes the key details regarding resource allocation for virtual machines and network configurations.
- Cloud Controller (CC): The CC serves as a standard cloud controller (e.g., OpenStack) with added support for streaming networks.
- SDN Controller (SC): The SDN controller is responsible for managing and provisioning the enterprise network topology. It maps network requirements to an appropriate set of physical and virtual networks, including the customer's CPE (Customer Premises Equipment).
- SDN-based Central Processing Element (CPE) and vSwitch: This is a streamlined version of legacy CPE, with most Network Functions (NFs) removed. Virtual Network Functions (VNFs) operate within virtual machines hosted on a hypervisor running on a server. Multiple virtual services can share the same server resources efficiently.

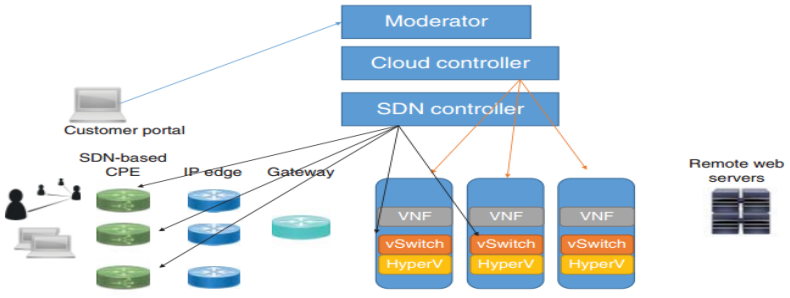

Fig. 4. Architecture Implemented

4.2. Results and performance evaluation

Load balancing remains a critical challenge, particularly concerning the efficient use of bandwidth for message routing and the optimal utilization of computing resources.

Figure 5 below illustrates the resource utilization levels, providing a comprehensive overview of system performance. It includes key metrics such as CPU usage, memory consumption, network activity, data volume, file system usage within the network, and storage domain levels. This interface serves as a performance dashboard, enabling the AI agent to make informed decisions by monitoring resource saturation levels and ensuring balanced system operation.

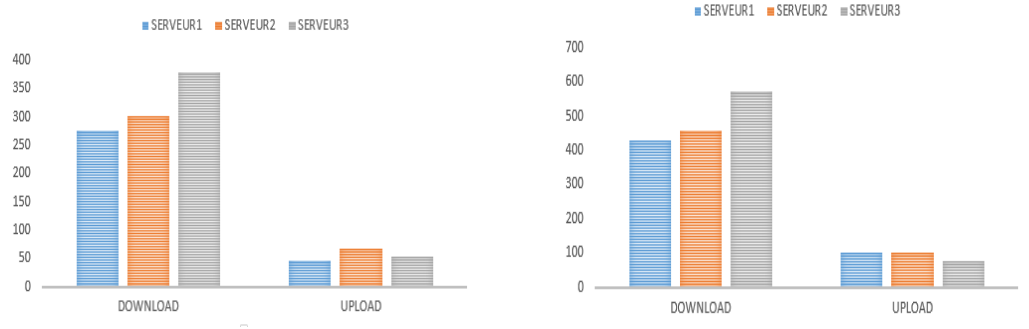

Fig. 5. Maximum throughput seen in Mbps before and after AI integration

Figures 5 compare the performance of a 5G network with and without AI integration. Without AI, download and upload speeds were around 100 Mbps and 20 Mbps, respectively, under normal conditions, and increased to 400 Mbps and 50 Mbps during maximum throughput tests. With AI integration, these speeds improved to 180 Mbps for downloads and 40 Mbps for uploads under normal conditions, and to 600 Mbps and 100 Mbps during maximum tests. AI integration resulted in a throughput improvement of approximately 5.5%, highlighting its significant role in enhancing the performance and efficiency of 5G networks.

4.3. Discussion and future directions

This work successfully implemented a paravirtualized 5G system integrated with AI to optimize server load management using Synology. By incorporating essential components of the 5G core network—vCloud for hosting virtual servers, SDN for application management, and NFV for network communication—an efficient and scalable architecture was established. The AI-driven supervised learning approach monitored server traffic based on load threshold indicators, enabling intelligent decision-making for workload distribution. Additionally, service cloning through clustering ensured fault tolerance, allowing seamless service migration to a relay server in the event of failure, thereby maintaining system continuity and minimizing downtime. The simplified system architecture effectively reduced resource consumption while ensuring optimal service management, addressing critical challenges related to load balancing in virtualized environments.

Despite these achievements, challenges emerged, particularly regarding limited documentation and implementation constraints, which were successfully navigated. The system demonstrated the feasibility of AI-driven load management for 5G networks, improving availability, reliability, and performance. Future work will focus on expanding the study to accommodate large-scale deployments and evaluating system performance under high traffic conditions with thousands of concurrent users. Enhancements will include advanced AI techniques, such as deep learning and reinforcement learning, for dynamic optimization. Cybersecurity integration will be a priority, incorporating intrusion prevention and encryption mechanisms to safeguard network security. Additionally, adaptations for next-generation 6G networks will be explored to meet future demands for lower latency and increased bandwidth, while predictive fault tolerance systems will be developed to enable proactive issue resolution.

4. Conclusion and future directions

This study presented the implementation of a paravirtualized 5G system integrated with AI to optimize load balancing in virtualized environments. By combining technologies such as SDN, NFV, and vCloud, the proposed architecture ensured efficient workload distribution and service continuity through clustering mechanisms, while supervised AI enabled real-time monitoring and decision-making. Despite the challenge of limited access to deployed 5G networks, the simulated environment provided valuable insights into the system's feasibility and performance. Moving forward, future work will focus on scaling the system for larger deployments involving thousands of concurrent users, integrating advanced AI techniques such as reinforcement learning and deep learning, and implementing robust cybersecurity measures to protect against emerging threats. Additionally, the architecture will be adapted to meet the increasing demands of next-generation 6G networks, ensuring lower latency.

Acknowledgements

Special acknowledgements to Intelligent & Sustainable Decision Making (ISDM) Research Team of the Faculty of Sciences of the University of Douala.

References

[1] Cisco Visual Networking Index. Global mobile data traffic forecast update, 2016–2021 white paper. Cisco: San Jose, CA, USA, 7:180, 2023.

[2] Bazelon, C., Sanyal, P., Taylor, R., Peretz, L. S., Sullivan, M., Selfe, N., & Christenson, P. (2023). Forecasting Wireless Broadband Capacity Shortfalls. *Available at SSRN 4528778*..

[3] Muaaz, M., Chelli, A., & Pätzold, M. (2021). Wi-Fi-based human activity recognition using convolutional neural network. In Innovative and Intelligent Technology-Based Services For Smart Environments-Smart Sensing and Artificial Intelligence (pp. 61-67). CRC Press.

[4] Al-Dujaili, M. J., & Al-dulaimi, M. A. (2023). Fifth-generation telecommunications technologies: Features, architecture, challenges and solutions. Wireless Personal Communications, 128(1), 447-469.

[5] Kitindi, E. J., Fu, S., Jia, Y., Kabir, A., & Wang, Y. Wireless network virtualization with SDN and C-RAN for 5G networks: Requirements, opportunities, and challenges. IEEE Access, 5, 19099-19115.

[6] Ferry Grijpink, Alexandre Ménard, Halldor Sigurdsson, and Nemanja Vucevic. Network sharing and 5g: A turning point for lone riders. McKinsey, Feb, 2024

[7] Mumtaz, S., Huq, K. M. S., Rodriguez, J., Ghosh, S., Ugwuanyi, E. E., Iqbal, M., ... & Georgakopoulos, P. (2021, December). Self-organization towards reduced cost and energy per bit for future emerging radio technologies-sonnet. In 2017 IEEE Globecom Workshops (GC Wkshps) (pp. 1-6). IEEE.

[8] Karl Aberer, Philippe Cudré-Mauroux, Anwitaman Datta, Zoran Despotovic,Manfred Hauswirth, Magdalena Punceva, and Roman Schmidt. P-Grid : Aself-organizing structured P2P system. Special Section on Peer to Peer Data Management, SIGMOD Record, 32(3) :29–33.

[9] Sameer Ajmani, Dwaine E. Clarke, Chuang-Hue Moh, and Steven Richman.Conchord : Cooperative SDSI certificate storage and name resolution. In 1st International Workshop on Peer-to-Peer Systems (IPTPS'02), LNCS 2429, Springer, pages 141–154, Massachusetts, USA.

[10] Aberer, K., & Hauswirth, M. (2002, March). An Overview of Peer-to-Peer Information Systems. In WDAS (Vol. 14, No. 2002, pp. 171-188).

[11] Agrawal, K., Kuszmaul, W., Wang, Z., & Zhao, J. (2024, June). Distributed Load Balancing in the Face of Reappearance Dependencies. In Proceedings of the 36th ACM Symposium on Parallelism in Algorithms and Architectures (pp. 321-330).

[12] Los, D., Sauerwald, T., & Sylvester, J. (2023). Balanced allocations with heterogeneous bins: The power of memory. In Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA) (pp. 4448-4477). Society for Industrial and Applied Mathematics.

[13] Bienkowski, M., Korzeniowski, M., der Heide, F.M.a. (2005). Dynamic Load Balancing in Distributed Hash Tables. In: Castro, M., van Renesse, R. (eds) Peer-to-Peer Systems IV. IPTPS 2005. LNCS, vol 3640. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11558989_20.

[14] Bhagwan, R., Savage, S., & Voelker, G. M. (2003). Understanding availability. In Peer-to-Peer Systems II: Second International Workshop, IPTPS 2003, Berkeley, CA, USA, February 21-22, 2003. Revised Papers 2 (pp. 256-267). Springer Berlin Heidelberg.

[15] Nogueira, J. (2014). A large-scale and decentralised application level multicast infrastructure.

[16] Tchappi, I., Bottaro, A., Gardes, F., & Galland, S. (2021). Towards an Online Agent Based Collision Avoidance by Mobile Edge Computing. In Advances in Practical Applications of Agents, Multi-Agent Systems, and Social Good. The PAAMS Collection: 19th International Conference, PAAMS 2021, Salamanca, Spain, October 6–8, 2021, Proceedings 19 (pp. 279-290). Springer International Publishing.

[17] Tchappi, I. H., Galland, S., Kamla, V. C., & Kamgang, J. C. (2018). A brief review of holonic multi-agent models for traffic and transportation systems. Procedia computer science, 134, 137-144.

[18] Diderot, C. D., Bernice, N. W. A., Tchappi, I., Mualla, Y., Najjar, A., & Galland, S. (2023). Intelligent transportation systems in developing countries: Challenges and prospects. Procedia Computer Science, 224, 215-222.