SCIENTIFIC Publishing Limited

Journal of Intelligent Communication
https://ojs.ukscip.com/index.php/jic

Article

On Improving Traffic Management in Small Cell Network Using a Novel Uplink Caching Framework

Mubarak Mohammed Al Ezzi Sufyan [1,*], Waheed Ur Rehman [2], Mahfoudh Al-Asaly [3], Ghassan A. A. Al-Maamari [4], Tabinda Salam [5] and AbdulRahman Al-Salehi [6]

[1] Department of Computer Information Systems, Al-Jawf Faculty, University of Saba Region, Marib, Yemen
[2] Department of Computer Science, University of Peshawar, Peshawar 25000, Pakistan
[3] Department of Information Systems, College of Computer, Qassim University, Buraydah 51174, Saudi Arabia
[4] Department of Computer Information Systems, Faculty of Computer Science & IT, University of Saba Region, Marib, Yemen
[5] Department of Computer Science, Shaheed Benazir Bhutto Women University Peshawar, Peshawar 25000, Pakistan
[6] Department of Computer Science, COMSATS University Islamabad (CUI), Islamabad Campus, Islamabad 45550, Pakistan
[*] Correspondence: mub.sufyan2015@gmail.com or mub.sufyan2025@gmail.com

Received: 4 July 2025; **Revised:** 14 August 2025; **Accepted:** 17 August 2025; **Published:** 3 September 2025

Abstract: The exponential growth of data traffic and user demand in modern communication systems has significantly increased the complexity of data streaming and management in Beyond Fifth Generation (B5G) networks. These networks face critical challenges such as network congestion, traffic load imbalance, latency, energy consumption, spectrum inefficiency, and limited storage capacity for real-time content delivery. Addressing these issues requires new architectural and conceptual approaches rather than incremental improvements to existing methods. This paper introduces a novel conceptual framework for cache-enabled uplink transmission within heterogeneous network environments comprising Macro Base Stations (MBSs), Small Cell Networks (SCNs), and mobile user devices. The proposed framework aims to optimize uplink content delivery by eliminating redundant cached data through curated content lists and employing content segmentation for distributed cache placement. The framework is organized into three interrelated components: Unified Distributed Cached Content Management at the MBS level, Content Deduplication and Segmentation at the SCN level, and Content Matchmaking at the mobile device level. Together, these components enable efficient data synchronization, enhance resource utilization, and minimize redundant data transmissions. Although this study is primarily conceptual, it establishes a strong theoretical foundation for future experimental validation. The proposed design is expected to improve traffic management efficiency, reduce energy consumption, and enhance Quality of Service (QoS) and user experience in future B5G and 6G communication environments.

Keywords: Distributed; Uplink; Caching-Enable; Hierarchy; Consolidation; Duplication; Elimination; Broadcasting

1. Introduction

In recent years, there has been a significant increase in data production and demand by end-users, leading to exponential growth in data streaming. It is expected that the number of internet users will reach 7.7 billion by 2027.

This growth poses substantial challenges for Beyond 5th Generation (B5G) networks, including increased traffic load, congestion, latency, and heightened demands on energy consumption, spectrum usage, storage, and content delivery [1–3]. In the context of Fifth Generation (5G) and B5G wireless communication, many technologies such as small cell networks (SCNs), coordinated multi-point (CoMP), and massive multiple-input multiple-output (mMIMO) technology, etc., have been integrated for enhancing the performance and capabilities of cellular networks. They played a crucial role in extending 5G network coverage, particularly in urban and densely populated areas, which required the deployment of a large cell network. The significance of increasing the network capacity, coverage, and enhancing overall performance. The strategy of network densification is seen as a crucial approach in optimizing the 5G and B5G infrastructure to meet the growing demand for high-speed with low-latency communication [4,5]. Moreover, another approach to managing the data explosion is through network caching at the edge [6,7], and then Cache-enabled small cell networks [8]. This technique involves storing cached versions of content at various locations within the network, such as Base Stations (BS), Macro Base Stations (MBS), SCNs, and Mobile devices. When these cached versions are available, the devices can access the content locally without needing to upload or download it from the core network [9]. This local availability significantly reduces communication costs, energy and bandwidth consumption, and latency [10,11]. **Figure 1** shows the caching scenario in B5G cellular network using SCNs, MBS, and the cloud.

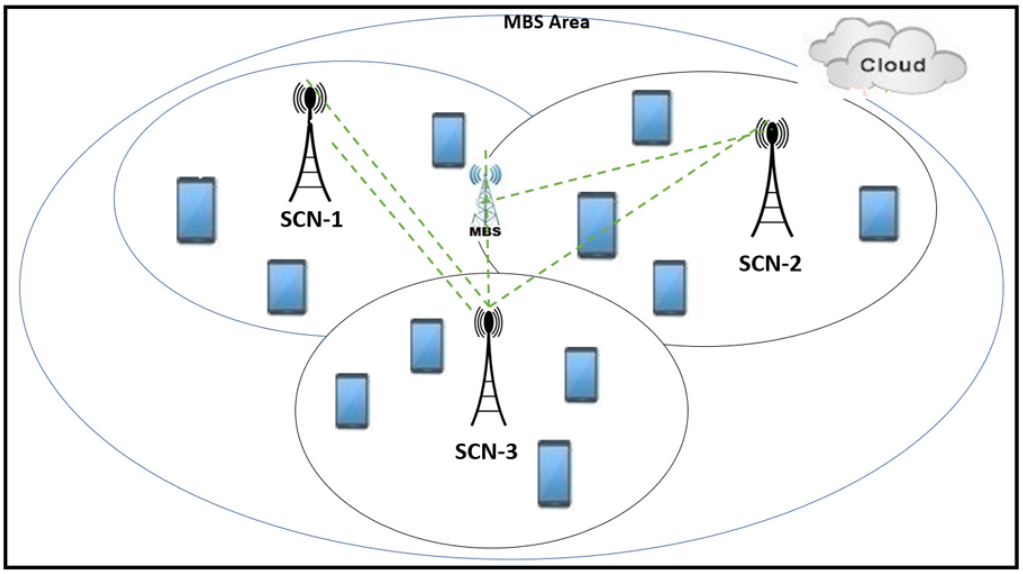

Figure 1. A caching scenario in B5G cellular network.

In the context of uplink transmission, cache-enabled mechanisms significantly impact cellular networks. Current literature explores various approaches for improving and enhancing the cache-enabled uplink transmission in cellular networks. In this vein, Pu [12] developed an upload acceleration service framework for mobile devices using host-based Wi-Fi as the access point. That framework operated in three stages: designated flow data acceleration, a two-tier caching architecture (front-end and back-end), resulting in reduced uploading time and stable User Generated Content (UGC) transmission. Similarly, an edge-network upload cache was proposed to assist in uploading UGC [13], aiming to benefit both service providers and end-users by reducing upload time and peak traffic volume between edge networks and data centers. Another approach introduced by Tai [14], which presented a Smart Offloading Proxy (SOP) for caching content uploads during congested events, which enhanced the user experience and reduced the upload scheduling time. Moreover, the cache-enabled uplink transmission in SCNs has been placed in research, such as Zhang [15], presented a theoretical framework to alleviate wireless SCNs' burdens through cache-enabled uplink transmission, duplicate content elimination, and cache management strategies. That

caching capability at Small Base Stations (SBS) improved transmission efficiency by reducing wireless backhaul traffic congestion and access link utilization. Further, Tokunaga [16], proposed a novel upload cache architecture supporting parallel uploading of segmented files using distributed edge servers on multi-Radio Access Technology (RAT) HetNets. As well as Papazafeiropoulos [17], a MIMO network architecture with numerous Base Stations (BSs) employing cache-enabled uplink transmission was developed that improved outage probability and delivery rate with increased cache storage. On the other hand, studies addressing Energy Efficiency (EE) and cache hit rate, such as Khoshkholgh [18], have examined the impact of user association on the hit rate and EE of Femto-BS uplink caching under various Cell Association (CA) schemes. These schemes were proposed and evaluated in terms of EE, hit rate, and data rate. Contrarily, Sufyan [19], introduced the Broadcast Cache Assist Uplink (BCAU) scheme to avoid redundant content uploads by matching cached and incoming content attributes, significantly improving EE and throughput in uplink transmission over Beyond 5G Small Cell Networks (B5G-SCN). In addition, Sufyan [20] proposed a cooperative SBS cache distribution framework aimed at eliminating duplicated contents between SCNs and MBSs to support uplink transmission, with a particular focus on content matching and caching large content fractions. However, the efficiency of the distributed cache with respect to existing contents and their sizes was not addressed [20]. This limitation was later discussed by Chiotiset al. [21], where the authors tackled the issue by eliminating duplicated contents across distributed caches based on scenarios presented by Sufyan [19,20].

Furthermore, researchers studying uplink traffic over cellular networks have shifted their focus to cache placement at various locations, such as MBSs and distributed SCNs. This strategy, however, has resulted in duplicated content uploads and transmissions, as well as redundant content storage in distributed caches. Such redundancies increase energy consumption, spectral usage, and waste cache space. To address these issues, new approaches have emerged, including content deduplication, segmentation, and redistribution of cached content, with the goals of reducing communication costs and resource consumption, improving the end-user experience, and enhancing overall network performance and efficiency [19–21]. Nevertheless, enhancing user experience, reducing redundant traffic load, and optimizing cache utilization, along with managing the entire cached content, remain key challenges in current cellular networks and will continue to be critical for future network generations, particularly when considering the optimization of uplink performance in limited-capacity radio stripes [22].

The remainder of this paper is organized as follows: Section 2 presents the problem statement and main contributions, followed by a discussion of the uplink transmission framework in Section 3. Section 4 introduces the framework architecture, and Section 5 outlines future work. Finally, the conclusion is provided in Section 6.

2. Problem Statement and Contributions

Building upon previous discussions, the challenges faced by single or distributed cache include content duplication, limited knowledge about cache contents at the Mobile Device, and the effective caching of large new or existing content. Additional challenges include mobile devices being unaware of cached content, uploading duplicated content to the cache, and storing duplicate content among distributed caches. Furthermore, there is often insufficient available space for incoming content and unequal distribution of free space across the distributed cache. Although existing solutions in the literature have tackled certain challenges at one or two levels of the network architecture, they frequently fail to offer a complete resolution. Still, critical issues such as improving user experience, minimizing redundant traffic, optimizing cache utilization, and effectively managing all cached content continue to persist. This motivates the proposal of a framework aimed at tackling these challenges across all three tiers of the network: Macro Base Stations (MBS), Small Cell Networks (SCNs), and mobile devices. To this end, the study formulates several key research questions that inform the design and evaluation of the proposed uplink caching framework: How can SCN densification improve traffic management and uplink performance? How does the multi-tier network topology influence efficiency, and how can predictive models support proactive traffic management? How effective are cache-enabled SCNs in reducing latency, bandwidth usage, and communication costs, and what strategies optimize cache utilization, content placement, and duplication across tiers? How can uplink mechanisms reduce upload time, match cached content with mobile uploads, and adapt to dynamic traffic and device behavior? Finally, what are the trade-offs in cache allocation, and how can mobile participation and distributed cache management enhance energy efficiency, upload speed, and overall network performance in dense multi-tier environments? In this regard, the main aims of this framework are to reduce the communication cost and resource consumption of the cellular network by limiting the amount of data transfer. The primary objective is to eliminate duplicated cached content

among distributed cache, generate a list of unique cached contents, store unique sizable content at the MBS, and segment content for distributed cache placement, as well as perform Matching-based attributes (MA) on mobile devices. For that, the matching performs among contents in two ways: among heterogeneous attributes at the mobile tier using Dissimilarity for Attributes of Mixed Types [23], and among content chunks at the SCNs and MBS tiers using the Method of Content-Defined Chunking Algorithm [24,25] for special scenarios.

The main contributions of this work, building upon prior research [15,17,20,21], are summarized as follows:

1. Propose a unified framework for uplink transmission in Beyond 5G Small Cell Networks (B5G-SCNs) addressing next-generation network challenges and requirements.
2. Design an efficient caching mechanism to reduce redundant communication by maintaining and updating distributed cached content based on content popularity and validity.
3. Introduce a content matching system to identify cached content on mobiles, minimizing unnecessary uploads and communication costs.
4. Implement duplication elimination and segmentation of large contents with distributed placement, optimizing cache utilization, and enabling efficient content redistribution across the network.
5. Enhance network performance by improving cache hit ratio, reducing uplink traffic, and increasing overall network efficiency.

Overall, the paper proposes a streamlined framework for caching, content matching, and management in B5G-SCNs, aiming to enhance network capacity and performance. It should be noted that this work primarily presents a conceptual framework and architectural design. The theoretical foundation and operational mechanisms are detailed here, while quantitative validation through simulation is planned as the next step, as discussed in the Future Work section.

3. Uplink Traffic Management Framework (UTMF)

The UTMF presented in this study is a comprehensive solution aimed at addressing key challenges encountered in cache-enabled uplink transmission within cellular networks. This framework operates across all three levels of the network architecture, Macro Base Stations (MBS), Small Cell Networks (SCNs), and Mobile devices to optimize the efficiency of content delivery. Its main objectives include eliminating duplicated cached contents among distributed caches, generating a curated list of cached contents, storing unique sizable content at the MBS, and segmenting content for efficient cache placement, as well as facilitating content matching at mobiles. Moreover, leveraging techniques such as Dissimilarity for Attributes of Mixed Types for attribute-based matching and the Method of Content-Defined Chunking Algorithm for chunk-based matching, which is used to perform deep matching with its own duplicated contents only. By integrating insights from various network levels and leveraging prior research, the Framework ensures efficient content management and transmission in cache-enabled uplink scenarios to improve both user experience and network performance by reducing redundant traffic, optimizing cache utilization, and efficiently managing all cached content. Based on the three levels of the network, the proposed framework comprises three main components, each performing specific tasks at each level and facilitating the exchange of input and output. These components are outlined in the following subsections.

3.1. Macro Base Station (MBS) Level

The MBS serves as the center of the cellular network, collecting relevant information from all SCNs and controlling them, as illustrated in **Figure 1**. A Mobile Device must have a list of distributed cached contents to avoid unnecessary uploads. Therefore, in this study, an MBS is designated as the marker for the final content list (MFCL). The goal of the MFCL is to create two content lists, namely, the Un-Duplicated Content List (UDCL) and the Duplicated Content List (DCL). As their names imply, the UDCL consists of unique distributed cached contents that can be used by mobiles for content matching, while the DCL contains duplicated distributed cached contents intended for evicting identical contents. The MBS performs different tasks as it prepares, consolidates, filters, and verifies cached contents to generate the UDCL and DCL. And broadcasting these lists to mobile devices, which then decide to retain or remove content according to MBS recommendations. The process is illustrated in **Figure 2**.

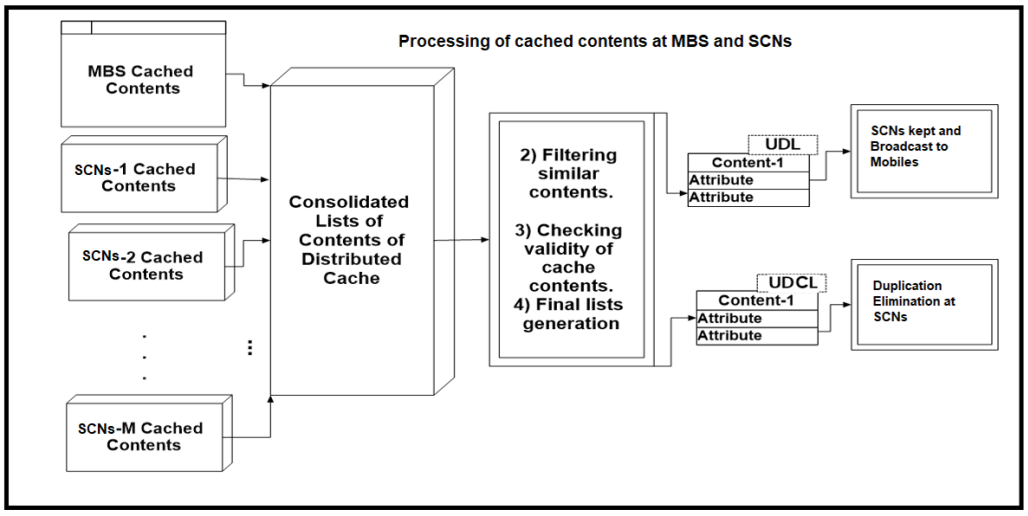

Figure 2. Unified Distributed Cached Contents Management Module (UDC^2M^2) [20].

Building on the previously described MBS operations, the data flow diagram termed Unified Distributed Cached Contents Management Module (UDC^2M^2) is represented as the first part of the proposal framework, which is illustrated in **Figure 3**.

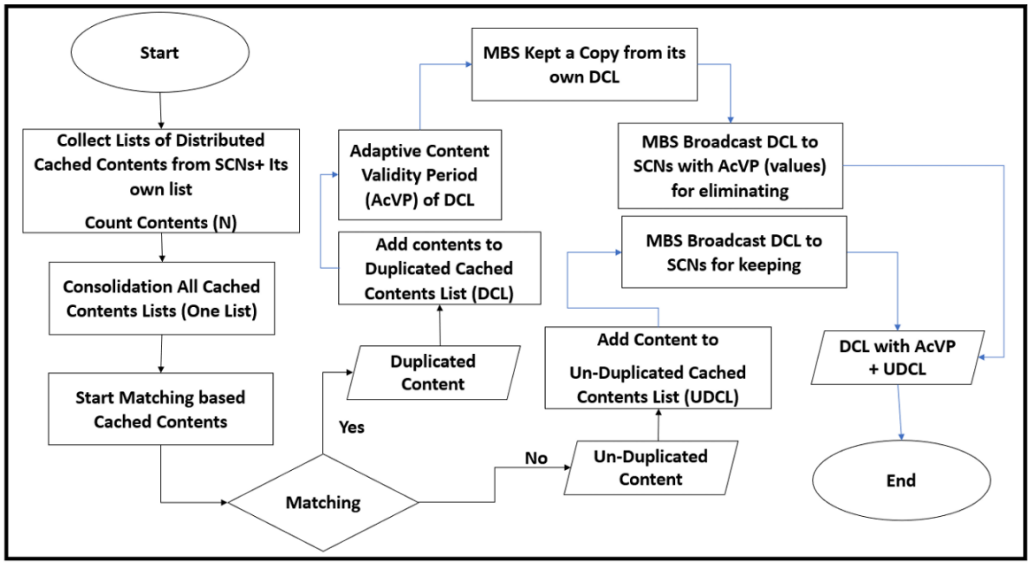

Figure 3. The diagram of the Unified Distributed Cached Contents Management Module (UDC^2M^2).

Furthermore, the main functions of the UDC^2 M^2 framework are detailed as follows:

- Gathering Cached Contents (GC^2), which can be performed as follows:
  - The MFCL gathers distributed cached contents from different SCNs and
  - Consolidates these lists with its own list of cached contents upon receiving the cached content lists from all SCNs.
- Consolidated Cached Contents Lists (C^3L), which can be performed as follows:
  - The MFCL applies a row-wise combination function to generate a consolidated list and then
  - Filters out similar content, which is received from various SCN caches.

- Content Matching (CM):
  Matching contents involves evaluating dissimilarity between content attributes, considering various data types by using the Dissimilarity for Attributes of Mixed Types algorithm [20] because each content has its own heterogeneous attributes, such as nominal, ordinal, interval, ratio, etc., as well as using the Method of Content-Defined Chunking Algorithm for special cases, such as for sizable cached contents.
- Generating Cached Content Lists (GC^2 L):
  The matching-based attributes resulted in two cached content lists, namely, the Unduplicated Content List (UDCL) and the Duplicated Content List (DCL).
- Validity of Similar Contents (VSC), which can be performed as follows:
  - The MBS retains any duplicates in its cache and advises the SCNs to remove them.
  - Subsequently, the MFCL verifies the validity of the DCL (duplicated contents among distributed SCNs) using an Adaptive Content Validity Period (AcVP) approach [20], to determine the source SCN for the content(s).
  - The AcVP calculates the probability of retaining or evicting content based on factors such as remaining energy, cache free space, and content popularity.
- Broadcast Cached Contents List (BC^2L):
  The MBS broadcasts the two lists to all SCNs, and then, based on the value of AcVP and MFCL recommendation, each SCN will perform either
  1. keeping UDCL, and knowledge about other cached contents, or
  2. remove DCL.
- Duplicate Elimination (DE):
  An DCL would remove from target SCNs based on the recommendation of the MFCL, and the value of AcVP.

3.2. SCN Level

At the SCN level, the main tasks performed include eliminating duplicated distributed cached content(s), segmenting large cached contents, placing segments, and distributively incoming content(s). These tasks form the second part of the proposed framework, referred to as Eliminated Duplicated, Distributed Content Segmentation, and Placement Allocation (*ED²CSPA*). The details of the main functions of *ED²CSPA* are as follows:

1. **Elimination of Duplicate Cached Contents (**EDC^2 **):**
   Each SCN stores the UDCL for optional deep matching, while the DCL guides content removal based on MBS recommendations and AcVP values.
2. **Segmentation and Redistribution of Sizable Cached Content (***SRSC²* **):**
   - Retain unique cached contents after duplicates are removed.
   - Remove duplicated cached contents, and retain unique ones.
   - Split large UDCL contents into smaller segments to fit cache limits.
   - Set segmentation thresholds based on content size and available space in SBSs and neighbors [20].
   - Encode and distribute segments using RUCC [21] for efficient storage and redistribution. The *SRSC²* process is illustrated in **Figure 4**.

   Furthermore, the *SRSC²* process consists of the following steps:

   (a) Determining Target Cached Contents:
   Identify cached contents whose size exceeds/equals a certain verification threshold.
   (b) Determining Target Caches:
   Identify caches among distributed SCNs with free space exceeding a verification threshold.
   (c) Segmentation of Target Contents:
   Segment based on available free space in local and corresponding SCNs.
   (d) Re-Distribute and Store Segments:
   Distribute and store segments across target caches, and update cache contents accordingly.

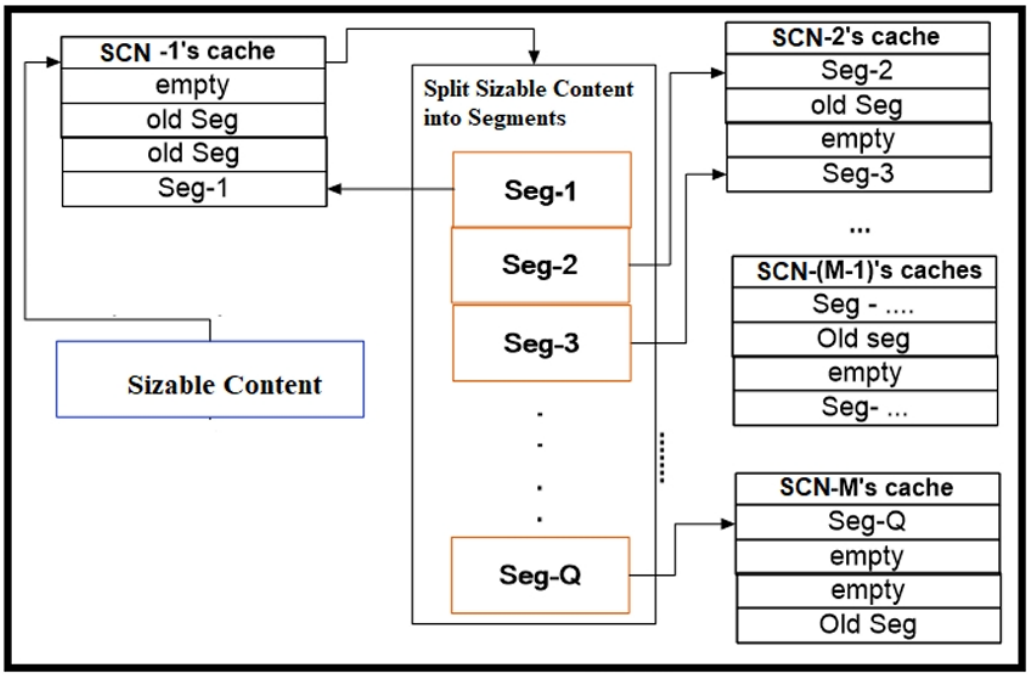

Figure 4. Segmentation and Redistribution of Sizable Cached Content ($SRSC^2$) [20].

3. **Segmentation, and Storing Distributively incoming Content (S^2DiC):**
In distributed caching, the effectiveness of the cache is often constrained by the size of the contents it manages. When the size of the new content exceeds the cache's capacity, the SCN will forward that content to MBS for storing, while if the size of the new content exceeds the cache's free space, it's advisable to partition it into smaller segments for distributed storage. As illustrated in **Figure 4**, incoming content is segmented based on its size and the available cache space of the corresponding SCNs and their neighboring nodes. This segmentation aims to optimize cache placement across multiple SCNs, thereby enhancing efficiency. The process can be executed as follows:
    - Firstly, gather necessary information, including the size of the new content, the hash key, and available space in target SCNs.
    - Secondly, segment the content into Q segments and distribute those segments among SCNs based on available space.
    - Finally, encode the segments using Maximum Distance Separable (MDS) code and create a map of encoded packets [19,20].
4. **Broadcasting, and Continuous Updating (BCU):**
Each SCN broadcasts the UDCL to all mobiles under its footprint for future use in matching. The main benefits of broadcasting the list of cached contents to the mobiles are as follows: To perform matching and eliminate duplicate content at the mobile before the real transmission of the actual content rather than at an SCN, which reduces upload traffic, latency of data uploading time, communication cost, average access time to the SCN cache, in addition to decreasing the mobile session duration of connecting to an SCN while reducing the energy spent by the mobiles and SCN, and reducing the massive number of mobiles, which are connected to an SCN. Moreover, the contents of the distributed cache undergo continuous updates, with their popularity changing due to factors such as upload times, downloads, shares, views, etc. Consequently, the UDCL is regularly updated and broadcast to all SCNs and mobile devices at specified intervals. This process ensures consistency of content between the distributed cache and the mobile devices.

Figure 5 shows the data flow of the Eliminated Duplicated, Distributed Content Segmentation and Placement Allocation (ED^2CSPA), which is performed at the SCN Level.

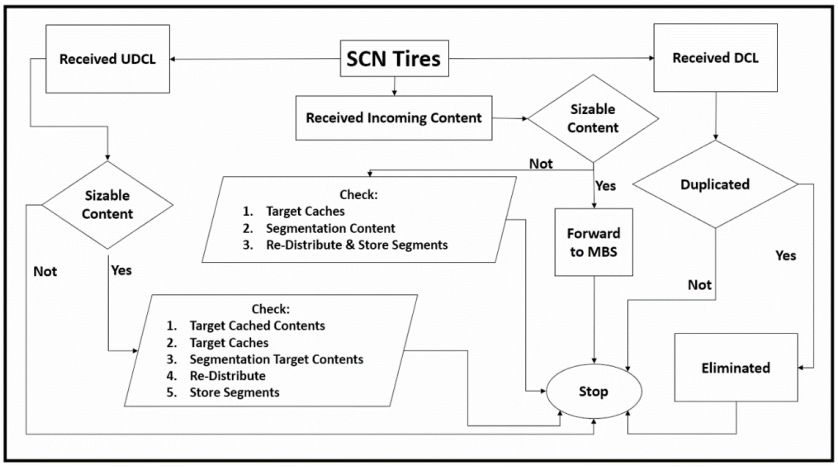

Figure 5. Eliminated Duplicated, Distributed Content Segmentation and Placement Allocation (ED^2CSPA).

3.3. Mobile Device Level

Due to the limited storage capacity at the SCN cache, coupled with constraints on the uplink transmission power of Mobile devices, along with increased communication cost and resource consumption, and added delay and latency in content delivery due to duplicate mobile data uploads, mobile devices often struggle to access information about distributed cached contents. This challenge is addressed by each SCN broadcasting the UDCL to mobile devices within its coverage area. Once mobile devices acknowledge the distributed cached contents and have data to upload, they proceed with the following tasks:

Firstly, the mobile device has content to upload, which is called incoming content (*iC*). In addition to generating a list of its attributes (*GiC*). That list includes heterogeneous attributes such as interval, string, numerical, etc.

Secondly, consolidating the (*GiC*) with the UDCL as (*GiC +UDCL*) using a row-wise combination function to generate a unified list, then initiating the matching process.

Thirdly, employing dissimilarity matching to perform Matching-based attributes (MA) among the consolidated list.

Finally, deciding whether to upload dissimilar content or discard similar content before actual transmission, thereby reducing upload traffic, decreasing the online time of connected Mobile Device, and minimizing resource consumption.

Those processes are termed Matchmaking Content at Mobile Level (MCML) and are crucial components of the proposed framework. The details of the MCMT are depicted in **Figure 6**.

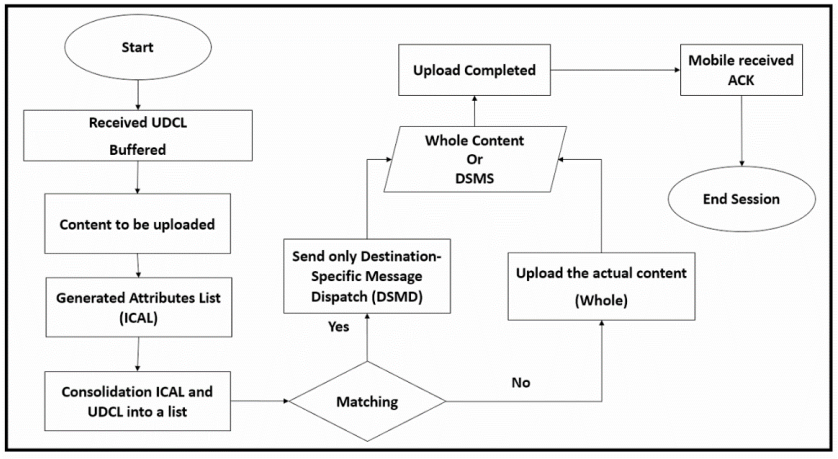

Figure 6. Matchmaking Content at Mobile Level (MCML).

4. UTMF Architecture

The architecture of UTMF for enhancing uplink traffic management through enabling caching of SCNs is designed to address challenges in cache-enabled uplink transmission in cellular networks. This framework operates hierarchically across MBS, SCNs, and Mobile Devices to optimize content delivery and storage. It aims to eliminate duplicate cached content, create curated lists of unique content, segment content for efficient cache placement, and facilitate content matching on mobile devices. The details of the framework have been discussed above in section (3). The summary of the description of the functionality is shown in **Table 1**. Moreover, **Figure 7** illustrates the overall UTMF.

Table 1. Functioning of UTMF Architecture.

Function	Description
	MBS Level ***(UDC^2M^2)***
Central Role	The MBS acts as the central controller, coordinating the management of cached contents among various distributed caches.
GC^2	Collects and consolidates content from SCNs with MBS cached contents.
GC^2L	Filters out duplicates and performs attribute-based matching.
CM	Uses the Dissimilarity for Attributes of Mixed Types algorithm to generate the UDCL and DCL, which are used to manage and eliminate redundant data.
VSC	Uses an Adaptive Content Validity Period (AcVP) to decide which content to keep and remove.
DE	Removes duplicates based on MBS recommendations and AcVP values.
BC^2L	Sends updated content lists to SCNs to manage cache efficiently.
	SCNs Levels ***(ED^2CSPA)***
EDC^2	Uses DCL to remove duplicates based on MBS suggestions.
$SRSC^2$	Segments sizable cached contents to optimize storage across SCNs.
S^2DiC	Segments and distributes new content to optimize cache space.
BCU	Regularly updates and broadcasts cached content lists to mobile devices to maintain consistency and reduce upload traffic and latency.
	Mobile Device Levels ***(MCMT)***
GAL	The Generating Attributes List (GAL) is used to create a list of attributes for new content.
Gic + UDCL	Combines new content attributes with the UDCL for matching.
MA	To determine if any duplicated contents exist, a heterogeneous attribute is used for matching.
Decision	Decides to upload dissimilar content or discard similar content to reduce upload traffic and energy consumption.

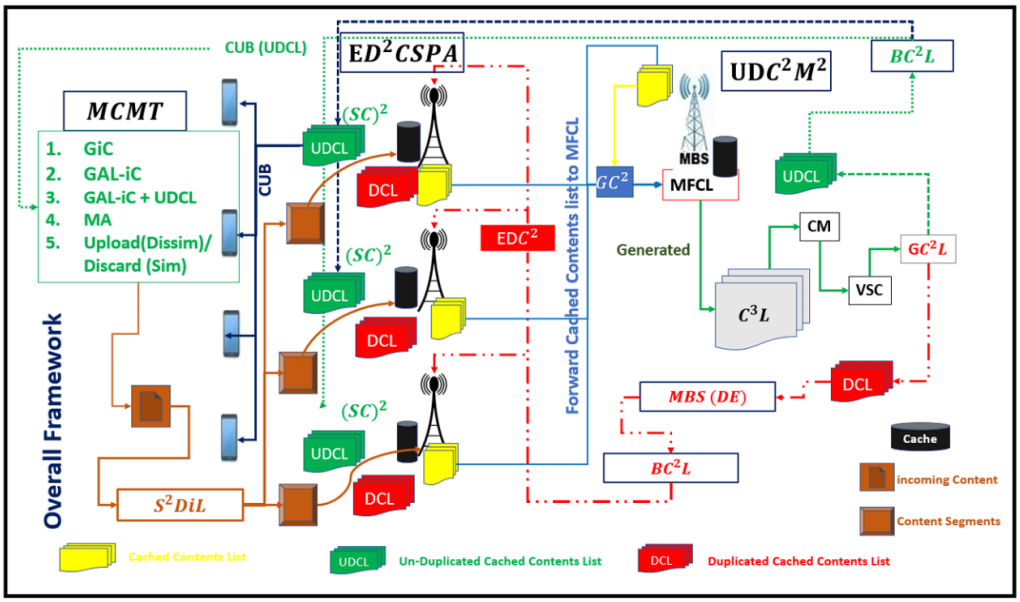

Figure 7. Overall, the Uplink Traffic Management Framework (UTMF).

The benefits to the end users can be summarized as follows:

1. Improved Battery Power: Users conserve energy by not uploading data if it is already cached.

2. Reduced Network Congestion: Leads to an improved quality of experience.
3. Fast and Efficient Data Upload: Results in lower service costs.
4. Reduced Online Time: Users spend less time connected to the SCN.
5. Shorter Response Time: Data processing locally results in quicker responses.

A UTMF improves user experience and network performance by integrating insights from various network levels, ensuring efficient content management and transmission. The overall framework, as shown in **Figure 8,** is the network scenario with the considering framework.

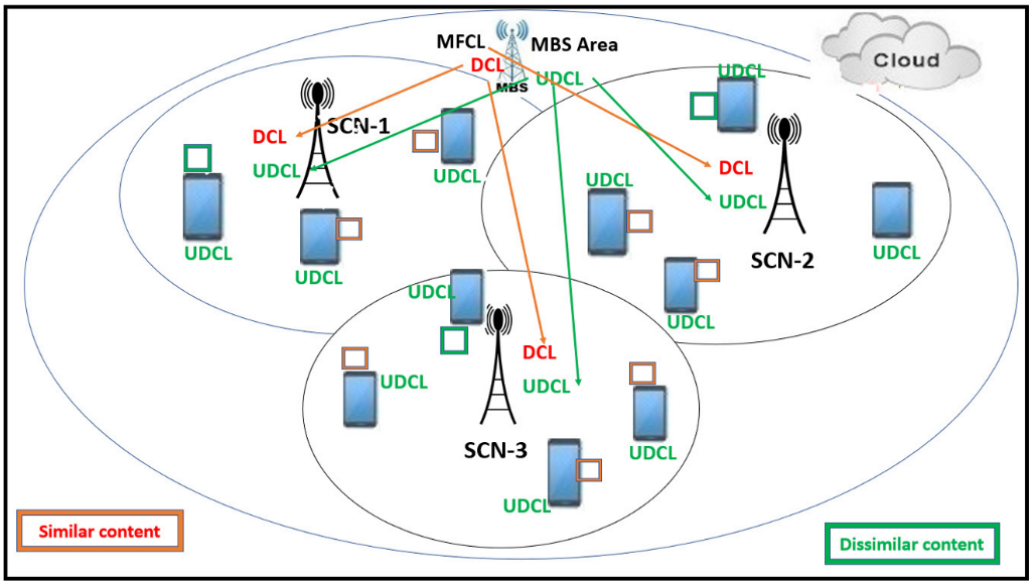

Figure 8. Network Scenario Considering the Proposed Framework.

5. Future Work

To further enhance the efficiency and intelligence of the proposed Uplink Traffic Management Framework (UTMF), future work may explore the integration of advanced Machine Learning (ML), Deep Learning (DL), and Artificial Intelligence (AI) techniques at various levels of the architecture. These technologies can improve decision-making, prediction accuracy, adaptive behavior, and autonomous management in cache-enabled uplink communication. The following directions are proposed.

5.1. AI-Enhanced Content Similarity and Matching

At the MBS and Mobile levels, the current matching mechanism relies on rule-based algorithms and dissimilarity measures. In future work:

- Deep learning models (e.g., Siamese Networks, Transformers) can be employed to learn complex and nonlinear representations of heterogeneous content attributes, enabling context-aware content similarity evaluation.
- Graph Neural Networks (GNNs) can model relationships between content pieces based on metadata, usage context, or user interaction patterns to improve duplicate detection and reduce false positives.

5.2. Intelligent Cache Replacement and Prediction at SCNs

Cache performance is heavily influenced by content popularity and availability.

- Implement reinforcement learning (RL) or multi-armed bandit algorithms to learn optimal cache replacement policies dynamically based on popularity trends, energy levels, and cache capacity.

- Use time-series forecasting models (e.g., Long Short-Term Memory (LSTM), Prophet) to predict future content demand or popularity, thereby proactively allocating and pre-fetching cache segments.

5.3. AI-Driven Adaptive Content Segmentation

The segmentation and distribution of sizable content segments can benefit from ML techniques that dynamically adjust thresholds and target caches:

- Apply clustering algorithms (e.g., K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN)) on SCN resource profiles (free space, energy, usage frequency) to group SCNs and identify the best candidates for segment placement.
- Develop a deep reinforcement learning (DRL) agent to optimize segment sizes and placements in real time for load balancing and minimal latency.

5.4. Mobility and Behavior-Aware Uplink Optimization

User mobility patterns and behaviors greatly influence uplink traffic:

- Incorporate AI models for mobility prediction (e.g., using Recurrent Neural Networks (RNNs) or attention-based models) to forecast user movement and proactively adapt UDCL dissemination.
- Use Federated Learning on mobile devices to train local models for predicting data upload intentions or content similarity, ensuring privacy-preserving personalization.

5.5. Energy-Efficient AI Models for Mobile Devices

Mobile-level matchmaking can be enhanced through lightweight AI models:

- Develop TinyML models that run efficiently on low-power mobile devices for content attribute extraction, similarity matching, and upload decision-making.
- Incorporate energy-aware model pruning or quantization techniques to reduce AI model size while preserving inference accuracy.

5.6. Anomaly Detection and Security Integration

To ensure secure and reliable caching:

- Use unsupervised anomaly detection (e.g., autoencoders, Isolation Forest) to detect abnormal content uploads or access patterns that may indicate cache poisoning or attacks.
- Combine AI with blockchain for trustworthy cache list validation, especially in distributed or multi-operator SCN environments.

5.7. End-to-End Simulation and Evaluation Framework

Future research could involve the development of a digital twin-based simulation framework powered by AI to:

- Simulate large-scale user behavior, cache dynamics, and network conditions.
- Evaluate the impact of different ML/DL algorithms on performance metrics such as hit rate, latency, energy savings, and network throughput.

By embedding intelligent, learning-based mechanisms at every level of the UTMF architecture, future enhancements can make the system autonomous, adaptive, and context-aware, enabling it to better handle dynamic environments, resource constraints, and ever-evolving content consumption patterns. The summary of the future works is shown in **Table 2**.

Table 2. Future Work Directions Using AI, ML, and DL for UTMF Enhancement.

Research Direction	Description and Techniques
AI-Enhanced Content Similarity and Matching	• Use deep learning models (e.g., Siamese Networks, Transformers) for context-aware content similarity evaluation. • Employ GNNs to model metadata and user interaction for better duplicate detection.
Intelligent Cache Replacement and Prediction at SCNs	• Apply reinforcement learning or multi-armed bandit algorithms for dynamic cache replacement. • Use time-series forecasting (e.g., LSTM, Prophet) to predict content demand and proactively pre-fetch content.
AI-Driven Adaptive Content Segmentation	• Use clustering algorithms (e.g., K-Means, DBSCAN) on SCN profiles to optimize content segment placement. • Develop deep reinforcement learning (DRL) agents to optimize segment sizes and real-time distribution.
Mobility and Behavior-Aware Uplink Optimization	• Use RNNs or attention-based models for predicting user mobility patterns to improve UDCL dissemination. • Integrate Federated Learning to predict upload behavior on mobile devices without compromising privacy.
Energy-Efficient AI Models for Mobile Devices	• Develop TinyML models for low-power content matching and upload decision-making. • Use model pruning or quantization to reduce computational load while maintaining accuracy.
Anomaly Detection and Security Integration	• Apply unsupervised learning (e.g., autoencoders, Isolation Forest) for detecting cache poisoning or abnormal uploads. • Combine AI and blockchain for secure cache list validation in distributed SCNs.
End-to-End Simulation and Evaluation Framework	• Develop a digital twin-based AI simulation framework for modeling large-scale cache-enabled networks. • Evaluate ML/DL algorithms against metrics like hit rate, latency, energy consumption, and throughput.

6. Conclusions

This paper proposed a comprehensive Uplink Traffic Management Framework (UTMF) for Beyond 5G (B5G) networks designed to address the pressing challenges of content duplication, limited cache awareness, and inefficient content storage. By adopting a multi-tiered architecture involving MBS, SCNs, and mobile devices, the framework enables the consolidation of cached content, the elimination of redundancy, segmentation for efficient placement, and content matching at the mobile level. These mechanisms collectively improve cache utilization, reduce unnecessary data uploads, alleviate uplink congestion, lower communication cost, and enhance the quality of experience for end-users through faster uploads and reduced device energy consumption. While the framework demonstrates significant potential benefits, several limitations must be acknowledged, which also define avenues for future research. First, the framework is presented as a conceptual model. Its performance evaluation was not based on real-world deployments or large-scale simulations. The complexity of real-world environments, including interference variability and heterogeneous hardware, may present challenges not captured in this design phase. Second, initial cache management policies and segmentation thresholds, though designed to be dynamic, may require more sophisticated, self-tuning algorithms to adapt to rapid fluctuations in content popularity and traffic patterns. Third, the impact of high user mobility and dynamic spectrum allocation, both critical in dense SCN environments, was not explicitly modeled in the current design, potentially affecting the scalability and accuracy of cache placement strategies. Finally, while the framework improves energy efficiency by avoiding redundant uploads, the computational overhead introduced at SCNs and mobile devices for content matching and management needs to be quantitatively assessed in future work. In this regard, the future work will focus on integrating AI- and ML-driven adaptive caching strategies, mobility-aware content dissemination, and intelligent spectrum allocation. In addition, the real-world prototyping and large-scale experimental validation will also be essential to assess energy trade-offs, deployment costs, and long-term scalability. By incorporating these enhancements, the UTMF can evolve into a more robust, adaptive, and deployable solution for next-generation B5G and 6G networks.

Author Contributions

M.M.A.E.S. provided the main idea for this work, wrote the main manuscript, designed the figures, and enhanced the novelty of the paper. M.M.A.E.S. is the main author of this paper. W.U.R. acted as the general supervisor and provided insights for the paper. M.A.-A. assisted in writing the manuscript and improving its language. G.A.A.A.-M. and T.S. provided further insights. A.A.-S. contributed additional insights, checked the latest references, and helped with the formatting. All authors have read and approved the final manuscript.

Funding

This work received no financial support.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no competing interests.

Abbreviations

Acronym	Description
B5G	Beyond 5th Generation
MBS	Macro Base Station
SCNs	Small Cell Networks
CoMP	Coordinated Multipoint
mMIMO	massive Multiple-Input Multiple-Output
5G	Fifth Generation
UGC	User Generated Content
SOP	Smart Offloading Proxy
SBS	Small Base Station
RAT	Radio Access Technology
BS	Base Station
EE	Energy Efficiency
CA	Cell Association
BCAU	Broadcast Cache Assist Uplink
B5G-SCNs	Beyond 5G Small Cell Networks
MFCL	Final Content List
UDCL	Un-Duplicated Content List
DCL	Duplicated Content List
$UDC^2\ M^2$	Unified Distributed Cached Contents Management Module
GC^2	Gathering Cached Contents
$C^3\ L$	Consolidated Cached Contents Lists
CM	Content Matching
GC^2 L	Generating Cached Content Lists
VSC	Validity of Similar Content
AcVP	Adaptive Content Validity Period
$BC^2\ L$	Broadcast Cached Contents List
$ED^2\ CSPA$	Eliminated Duplicated, Distributed Content Segmentation and Placement Allocation
EDC^2	Elimination Duplication Cached Contents
$SRSC^2$	Segmentation and Redistribution of Sizable Cached Content
S^2DiC	Segmentation and Storing Distributively Incoming Content
MDS	Maximum Distance Separable

Acronym	Description
BCU	Broadcasting and Continuous Updating
iC	incoming Content
GiC	Generating a list of attributes
GiC+UDCL	Consolidating this list with the UDCL
MA	Matching-based attributes
MCML	Matchmaking Content at Mobile Level
GAL	Generating Attributes List
ML	Machine Learning
DL	Deep Learning
AI	Artificial Intelligence
GNNs	Graph Neural Networks
RL	Reinforcement Learning
LSTM	Long Short-Term Memory
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DRL	Deep Reinforcement Learning
RNNs	Recurrent Neural Networks

References

1. Teodorescu, C.A.; Durnoi, A.N.C.; Vargas, V.M. The Rise of the Mobile Internet: Tracing the Evolution of Portable Devices. *Proc. Int. Conf. Bus. Excell.* **2023**, *17*, 1645–1654.
2. Godlovitch, I.; Martins, S.S.; Gries, C.; et al. *Study on Wholesale Mobile Connectivity, Trends and Issues for Emerging Mobile Technologies and Deployments*; WIK-Consult GmbH: Bad Honnef, Germany, 2023; pp. 1–185.
3. Al-Shareeda, M.A.; Hergast, D.; Manickam, S. Review of Intelligent Healthcare for the Internet of Things: Challenges, Techniques and Future Directions. *J. Sen. Netw. Data Commun.* **2024**, *4*, 1–10.
4. Sufyan, A.; Khan, K.B.; Khashan, O.A.; et al. From 5G to Beyond 5G: A Comprehensive Survey of Wireless Network Evolution, Challenges, and Promising Technologies. *Electronics* **2023**, *12*, 2200.
5. Zhang, M. Research on English Classroom Teaching Programs in Colleges and Universities Based on Wireless Communication Technology Support in the Context of 5G. *Int. J. Inf. Commun. Technol. Educ.* **2024**, *20*, 1–17.
6. Wang, X.; Chen, M.; Taleb, T.; et al. Cache in the Air: Exploiting Content Caching and Delivery Techniques for 5G Systems. *IEEE Commun. Mag.* **2014**, *52*, 131–139.
7. Bastug, E.; Bennis, M.; Debbah, M. Living on the Edge: The Role of Proactive Caching in 5G Wireless Networks. *IEEE Commun. Mag.* **2014**, *52*, 82–89.
8. Bastug, E.; Bennis, M.; Kountouris, M.; et al. Cache-Enabled Small Cell Networks: Modeling and Tradeoffs. *J. Wireless Com. Netw.* **2015**, *41*, 1–15.
9. Elshaer, H.; Boccardi, F.; Dohler, M.; et al. Downlink and Uplink Decoupling: A Disruptive Architectural Design for 5G Networks. In Proceedings of the 2014 IEEE Global Communications Conference, Austin, TX, USA, 8–12 December 2014; pp. 1798–1803.
10. Katz, M.; Pirinen, P.; Posti, H. Towards 6G: Getting Ready for the Next Decade. In Proceedings of the 2019 16th International Symposium on Wireless Communication Systems (ISWCS), Oulu, Finland, 27–30 August 2019; pp. 714–718.
11. Mattera, D.; Tanda, M. Windowed OFDM for Small-Cell 5G Uplink. *Phys. Commun.* **2020**, *39*, 100993.
12. Pu, Y.; Nakao, A. A Deployable Upload Acceleration Service for Mobile Devices. In Proceedings of the International Conference on Information Network 2012, Bali, Indonesia, 1–3 February 2012; pp. 350–353.
13. Zhu, Y.; Nakao, A. Upload Cache in Edge Networks. In Proceedings of the 2012 IEEE 26th International Conference on Advanced Information Networking and Applications, Fukuoka, Japan, 26–29 March 2012; pp. 307–313.
14. Tai, H.-T.; Chung, W.-C.; Wu, C.-J.; et al. SOP: Smart Offloading Proxy Service for Wireless Content Uploading over Crowd Events. In Proceedings of the 2015 17th International Conference on Advanced Communication Technology (ICACT), PyeongChang, Republic of Korea, 1–3 July 2015; pp. 659–662.
15. Zhang, Z.; Chen, Z.; Xia, B. Cache-Enabled Uplink Transmission in Wireless Small Cell Networks. In Proceedings of the 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 20–24 May 2018; pp. 1–6.

16. Tokunaga, K.; Kawamura, K.; Takaya, N. High-Speed Uploading Architecture Using Distributed Edge Servers on Multi-RAT Heterogeneous Networks. In Proceedings of the 2016 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN), Rome, Italy, 13–15 June 2016; pp. 1–2.
17. Papazafeiropoulos, A.; Ratnarajah, T. Modeling and Performance of Uplink Cache-Enabled Massive MIMO Heterogeneous Networks. *IEEE Trans. Wireless Commun.* **2018**, *17*, 8136–8149.
18. Khoshkholgh, M.G.; Leung, V.C. Impact of Cell Association on Energy-Efficiency and Hit Rate of Femto-Caching. *IEEE Trans. Mobile Comput.* **2020**, *21*, 1004–1017.
19. Sufyan, M.M.A.-E.; Rehman, W.U.; Salam, T.; et al. Duplication Elimination in Cache-Uplink Transmission over B5G Small Cell Network. *EURASIP J. Wireless Commun. Netw.* **2021**, *1*, 1–24.
20. Sufyan, M.M.A.E.; Rehman, W.U.; Salam, T.; et al. Distributed Uplink Cache for Improved Energy and Spectral Efficiency in B5G Small Cell Network. *PLoS ONE* **2022**, *17*, e0268294.
21. Ur Rehman, W.; Sufyan, M.M.A.E.; Salam, T.; et al. Cooperative Distributed Uplink Cache over B5G Small Cell Networks. *PLoS ONE* **2024**, *19*, e0299690.
22. Chiotis, I.; Moustakas, A.L. Uplink Performance Optimization of Limited-Capacity Radio Stripes. *IEEE Trans. Wireless Commun.* **2024**, *23*, 12382–12395.
23. He, Y.; Wang, M.; Yu, J.; et al. Research on the Hybrid Recommendation Method of Retail Electricity Price Package Based on Power User Characteristics and Multi-Attribute Utility in China. *Energies* **2020**, *13*, 2693.
24. Meyer, D.T.; Bolosky, W.J. A Study of Practical Deduplication. *ACM Trans. Storage* **2012**, *7*, 1–20.
25. Widodo, R.N.; Lim, H.; Atiquzzaman, M.J. A New Content-Defined Chunking Algorithm for Data Deduplication in Cloud Storage. *Future Gener. Comput. Syst.* **2017**, *71*, 145–156.