DOI: 10.21917/ijct.2026.0573

INTELLIGENT EDGE CACHING IN MOBILE NETWORKS: A SIMULATION-BASED STUDY OF CONTENT SELECTION, PLACEMENT AND PERFORMANCE

Sahib Bahadar and Nabi Rehmat

School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, China

Abstract

Mobile caching has become a practical mechanism for reducing latency, improving user experience, and lowering repeated backhaul transmissions in content-centric wireless systems. This study organizes mobile caching around three core questions: how to cache, what to cache, and where to cache. It evaluates replacement and admission ideas, compares edge placement options in the evolved packet core, radio access network, small base stations, and device-to-device links, and presents a simulation-based analysis of cache size, request load, popularity skew, and prediction accuracy. The results indicate that user-preference-aware and edge-oriented strategies provide higher hit ratio, lower latency, and stronger backhaul relief than conventional non-cooperative approaches. However, these gains depend on accurate popularity estimation, cache coordination, and the overhead of maintaining edge intelligence.

Keywords:

Mobile Caching, Edge Caching, RAN Caching, Cache Replacement, 5G

1. INTRODUCTION

The rapid growth of mobile multimedia traffic has turned caching from a supporting function into a central design element of modern wireless networks. Rich media services, video streaming, short-form content, and real-time applications place simultaneous pressure on radio access links, backhaul capacity, and core-network processing. As a result, simply increasing spectral efficiency is no longer sufficient; content must also be stored closer to the point of demand so that repeated downloads do not traverse the entire network path every time [1]–[5]. Earlier Internet systems used web caching and content delivery networks to relieve congestion caused by repeated requests for the same objects. The same principle now appears in mobile systems, but with stronger constraints. Edge nodes have limited storage, local popularity changes quickly, and mobility makes it difficult to maintain accurate content placement. These challenges explain why caching remains an active topic in 5G and beyond-5G research even though the basic idea itself is well established [4]–[9]. Much of the published work discusses only one part of the problem, such as replacement policy, popularity learning, or deployment layer. This paper integrates these themes into one coherent structure by connecting cache management, content selection, and cache placement. The contribution of this study is twofold. First, it condenses the literature into a clear discussion of how to cache, what to cache, and where to cache. Second, it presents a reproducible simulation section with equations, a parameter table, and five comparative graphs to evaluate the effect of cache size, request load, popularity skew, placement layer, and prediction accuracy. The remainder of the paper is organized as follows. Section 2 reviews the relevant literature, Section 3 presents the methodology and equations, Section 4 reports simulation results, Section 5 discusses the implications, and Section 6 concludes the paper.

2. LITERATURE REVIEW

2.1 CACHE MANAGEMENT AND REPLACEMENT

The question of how to cache is traditionally answered through replacement and coordination policies. Classical policies such as least recently used (LRU), least frequently used (LFU), and first-in first-out (FIFO) are simple and implementable, but they do not fully capture the constraints of wireless environments, where cache space is small, connectivity changes quickly, and the cost of a miss depends on both bandwidth and mobility [4], [12]–[16]. Time-aware and utility-based policies improve this situation by combining recency, frequency, lifetime, size, or delivery cost into the eviction decision. In mobile and ad hoc settings, cooperative caching is especially important because local caches can act as a larger virtual cache. Coordinated schemes try to reduce unnecessary duplication and improve accessibility across neighbouring nodes, whereas uncoordinated schemes make decisions using only local state. Collaborative multicell caching and cooperative ad hoc caching further show that coordination can improve access probability when storage is distributed across cells or mobile nodes [10], [24]. The literature consistently shows that simple policies remain useful as baselines, but better performance is obtained when replacement decisions consider validity time, node cooperation, or utility rather than recency alone [12], [14]–[16].

2.2 CONTENT SELECTION AND ADMISSION

The question of what to cache is usually addressed through popularity-driven or user-preference-driven selection. Static popularity models often assume an independent reference process, but real content demand is dynamic and highly skewed. A small set of objects usually generates most requests, yet ephemeral content such as news, episodic videos, and trending media may become popular only for a short period [9], [17], [18]. Measurements of YouTube traffic also support the watch-global, cache-local principle, where globally popular objects and locally repeated viewing patterns both matter for cache placement [23]. This makes popularity learning a critical part of cache effectiveness. Beyond global popularity, several studies show that local preference matters. Most-popular-video strategies are easy to implement but may not represent local cell demand. Reactive and proactive user preference profile methods improve local hit ratio by incorporating the behaviour of active users in a cell site. Admission control also plays a role by avoiding excessive replication and preserving cache diversity, especially when neighbouring nodes can already serve the same object [8], [17], [18].

2.3 CACHE PLACEMENT IN MOBILE NETWORKS

The question of where to cache leads to a comparison among the evolved packet core (EPC), the radio access network (RAN), small base stations (SBSs), and device-to-device (D2D) links. EPC caching is easier to manage centrally and benefits from a larger traffic view, but it still leaves repeated transmissions on the path between the core and the radio edge. RAN and SBS caching move content closer to users and usually reduce access delay more strongly, although their storage budgets are smaller and local demand estimates are noisier [3], [4], [8], [11], [19]–[22]. Cooperative RAN caching based on local altruistic game models also indicates that coordinated cell-level placement can improve delivery efficiency when neighbouring radio nodes cooperate [11]. D2D-assisted caching pushes reuse even closer to the user by exploiting storage already available in smartphones and nearby devices. This can produce strong local offloading and spectral gains, but it also introduces coordination, availability, and energy constraints. Across the literature, no single layer dominates under all conditions. Instead, the preferred placement depends on cache size, popularity skew, cooperation level, and how much operational complexity can be tolerated [19]–[22].

3. METHODOLOGY

A deterministic numerical simulation framework was developed to compare cache replacement policies and placement layers under controlled conditions. The simulation was implemented in Python 3.11 using NumPy for numerical computation and Matplotlib for graph generation. No operator trace was used; instead, requests were synthesized from a Zipf popularity model so that all compared strategies used the same demand distribution. This design makes the experiment reproducible and suitable for comparing relative trends among policies, although it should not be interpreted as a measurement of one specific commercial network.

3.1 CONTENT POPULARITY MODEL

A library of N objects is ranked according to a Zipf distribution, which is widely used in caching studies because it captures the strong popularity skew observed in multimedia demand. The request probability of the f-th object is defined as follows:

$$p_f = \frac{f^{-\gamma}}{\sum_{k=1}^{N} k^{-\gamma}} \tag{1}$$

The parameter γ controls the degree of popularity concentration. A larger γ means that a smaller subset of files attracts more requests, which generally improves the effectiveness of limited cache storage.

3.2 PERFORMANCE METRICS

For a cache containing the set C, the aggregate hit ratio is computed as the sum of the request probabilities of the cached objects. Average latency is modeled as a weighted combination of cache-hit delay and cache-miss delay. Backhaul relief is evaluated as the percentage reduction in carried traffic relative to a no-cache baseline.

$$H = \sum_{f \in C} p_f \tag{2}$$

$$L_{\text{avg}} = H\,L_{\text{hit}} + (1-H)\,L_{\text{miss}} \tag{3}$$

$$O = \frac{T_{\text{no-cache}} - T_{\text{cache}}}{T_{\text{no-cache}}} \times 100\% \tag{4}$$

In Eq.(3), L_{hit} denotes the latency of serving a request from the selected cache layer, while L_{miss} denotes the latency of fetching the object through the backhaul and core path. Eq.(4) reports the percentage of traffic removed from the backhaul relative to the no-cache reference case.

3.3 SIMULATION SETUP

The simulation compares LRU, LFU, TLRU, and preference-aware caching under different cache sizes. It also compares EPC, RAN, SBS, and D2D placement under fixed budgets. Request rate, popularity skew, and prediction accuracy are varied independently so that the results section can isolate the effect of each design parameter. The graphs were generated in four steps: first, a library of 5000 ranked objects was created; second, request probabilities were computed using Eq.(1); third, cache contents were selected according to each policy and placement assumption; and fourth, hit ratio, average latency, and backhaul relief were computed using Eq.(2)-Eq.(4). Policy differences were represented through controlled utility factors around the same Zipf demand model so that the comparison remained consistent across all figures. The Table.1 summarizes the parameter set used to generate the graphs.

Table.1. Simulation Parameters in the Representative Evaluation

Parameter	**Value**
Content library size (N)	5000 objects
Number of active users	1000 users
Cache budget	2%-20% of library
Popularity skew (Zipf γ)	0.6-1.4
Replacement policies	LRU, LFU, TLRU, P-UPP
Placement layers	EPC, RAN, SBS, D2D
Request arrival rate	100-1000 requests/s
Core/backhaul miss delay	120 ms
Representative hit delays	EPC: 45 ms, RAN: 18 ms, D2D: 10 ms
Prediction accuracy range	50%-95%

4. RESULTS

This section presents five graphs generated from the simulation procedure described in Section 3. The figures highlight the main comparative trends for replacement policy, placement layer, popularity concentration, and prediction accuracy.

4.1 CACHE HIT RATIO VERSUS CACHE SIZE

The Fig.1 shows that the hit ratio increases monotonically as the available cache budget grows. However, the gain depends strongly on policy choice. LRU offers the lowest performance because it reacts only to recency, while LFU benefits from repeated popularity concentration. TLRU performs better than both due to its time-awareness, especially when short-lived content appears. The best curve is obtained by the preference-aware policy, which combines local demand information with proactive placement and therefore captures both popularity and user context.

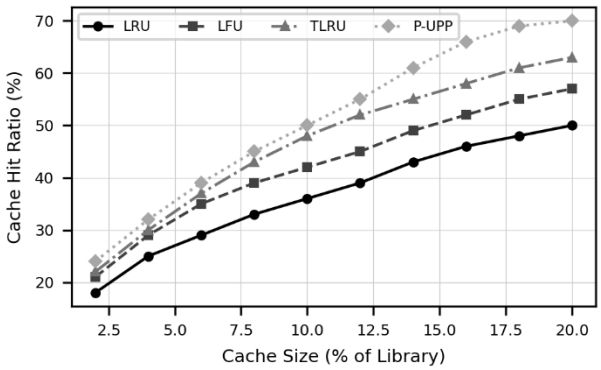

Fig.1. Cache hit ratio versus cache size for representative replacement policies

4.2 AVERAGE ACCESS LATENCY VERSUS REQUEST LOAD

The Fig.2 compares access delay under increasing request arrival rate. Without caching, latency rises quickly because every request traverses the full network path. EPC caching lowers latency, but RAN and D2D-assisted caching reduce it more strongly because reusable content is stored nearer to the user. The figure also illustrates that the relative advantage of edge caching becomes larger at heavier load, when backhaul congestion begins to dominate end-to-end service time.

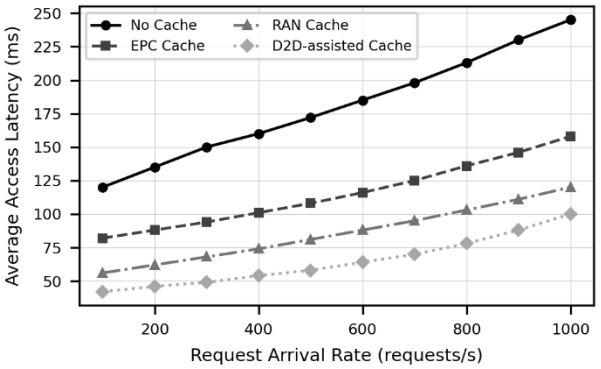

Fig.2. Average access latency versus request load for different placement layers

4.3 BACKHAUL REDUCTION VERSUS POPULARITY SKEW

The Fig.3 plots backhaul relief as the Zipf parameter increases. A larger popularity skew means that a smaller subset of objects accounts for more requests, so all caching layers become more effective. EPC caching provides moderate offload because it intercepts repeated flows at the core. RAN, SBS, and D2D strategies deliver stronger reduction because they eliminate a larger fraction of repeated fetches before the traffic reaches the backhaul. The figure confirms that popularity concentration is one of the strongest enablers of practical caching gains.

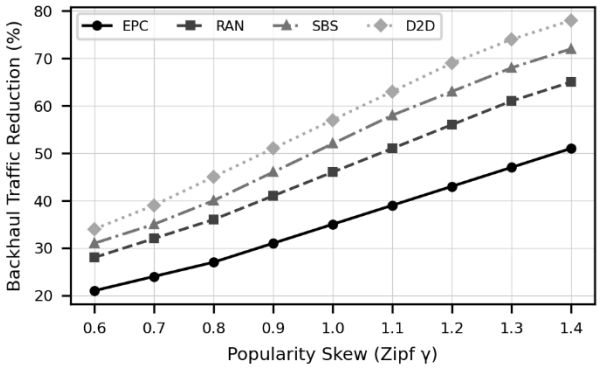

Fig.3. Backhaul traffic reduction versus popularity skew

4.4 PLACEMENT COMPARISON UNDER A FIXED CACHE BUDGET

The Fig.4 summarizes the trade-off between latency and backhaul reduction under a fixed 10% total cache budget. EPC caching still provides measurable relief and is easier to manage, but the edge-oriented layers achieve a better operating point. SBS and D2D placement produce lower mean delay because the delivery path is shorter, while their offload values are also higher. This result is consistent with the literature: the closer content is placed to users, the better the latency outcome, provided that coordination overhead remains manageable.

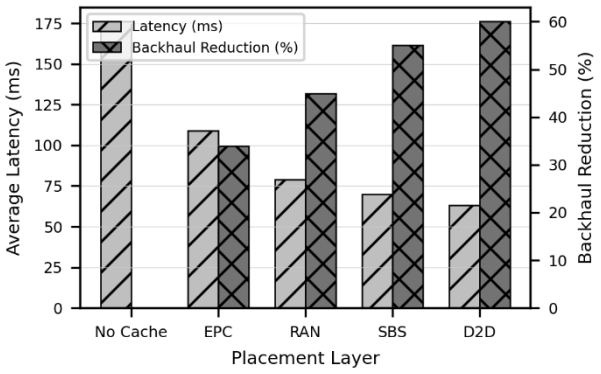

Fig.4. Placement comparison under a fixed total cache budget

4.5 EFFECT OF PREDICTION ACCURACY ON USER-PREFERENCE-BASED CACHING

The Fig.5 studies the role of prediction accuracy in popularity-aware and preference-aware methods. The MPV baseline changes only slightly with accuracy because it relies mostly on global popularity. R-UPP improves as cell-specific user behaviour becomes easier to estimate, and P-UPP achieves the highest gains when prediction quality is strong enough to justify prefetching. The graph therefore shows why intelligent caching must be paired with reliable demand estimation; otherwise, proactive placement may waste scarce edge capacity.

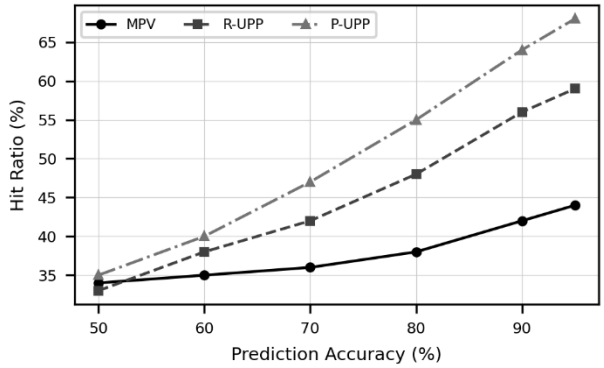

Fig.5. Effect of prediction accuracy on user-preference-based caching

5. DISCUSSION

The results section supports three main conclusions. First, policy design matters even when the cache budget is fixed. Simply increasing storage is useful, but larger gains come from combining replacement with local demand information. Second, placement closer to the radio edge produces a stronger delay reduction than centralized caching, although this benefit must be balanced against smaller local caches and higher coordination cost. Third, popularity skew and prediction quality act as gain multipliers; when popular content is concentrated and demand estimation is accurate, the same hardware budget yields a much larger performance improvement. These findings also clarify the practical limits of mobile caching. Edge caches are attractive only when operators can tolerate the additional control complexity required for synchronization, consistency, and admission decisions. Likewise, D2D caching offers the best locality but depends on device availability and user participation. For this reason, the most realistic deployment strategy is usually hierarchical: large stable caches remain in the core, while smaller adaptive caches are placed in the RAN, SBS, or device layer to absorb local demand bursts. The numerical simulation focuses on major design trends; future work can extend it with trace-driven workloads, user mobility, energy consumption, and coordination overhead analysis.

6. CONCLUSION

Mobile caching should be understood through three linked questions: how to cache, what to cache, and where to cache. The literature confirms that replacement policy, content admission, and placement layer must be considered together because each one changes the final cache hit ratio, delay, and backhaul load. The simulation results show that hit ratio improves with cache size, edge-oriented placement reduces delay, popularity skew increases backhaul relief, and user-preference-aware methods outperform purely popularity-based baselines when prediction quality is sufficient. In summary, mobile caching remains a high-impact mechanism for 5G and beyond, but its practical success depends on balancing storage, coordination, and demand prediction.

REFERENCES

[1] B. Assila, A. Kobbane and M. El Koutbi, “A Survey on Caching in 5G Mobile Network”, *WCOS*, Vol. 16, pp. 1-7, 2016.

[2] E. Borcoci, “Content Distribution in Wireless/5G Environments”, Available at https://www.iaria.org/conferences2015/filesICWMC15/InfoWare_2015_Content_5G_v1.3.pdf, Accessed in 2015.

[3] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang and W. Wang, “A Survey on Mobile Edge Networks: Convergence of Computing, Caching and Communications”, *IEEE Access*, Vol. 5, No. 1, pp. 6757-6779, 2017.

[4] X. Wang, M. Chen, T. Taleb, A. Ksentini and V. Leung, “Cache in the Air: Exploiting Content Caching and Delivery Techniques for 5G Systems”, *IEEE Communications Magazine*, Vol. 52, No. 2, pp. 131-139, 2014.

[5] G. Paschos, E. Bastug, I. Land, G. Caire and M. Debbah, “Wireless Caching: Technical Misconceptions and Business Barriers”, *IEEE Communications Magazine*, Vol. 54, No. 8, pp. 16-22, 2016.

[6] M.A. Maddah-Ali and U. Niesen, “Fundamental Limits of Caching”, *IEEE Transactions on Information Theory*, Vol. 60, No. 5, pp. 2856-2867, 2014.

[7] G. Pallis and A. Vakali, “Insight and Perspectives for Content Delivery Networks”, *Communications of the ACM*, Vol. 49, No. 1, pp. 101-106, 2006.

[8] H. Ahlehagh and S. Dey, “Video-Aware Scheduling and Caching in the Radio Access Network”, *IEEE/ACM Transactions on Networking*, Vol. 22, No. 5, pp. 1444-1462, 2014.

[9] M. Leconte, G. Neglia, A. Carra and P. Michiardi, “Placing Dynamic Content in Caches with Small Population”, *Proceedings of International Conference on Computer Communications*, Vol. 23, pp. 1-9, 2016.

[10] A. Gharaibeh, A. Mohamed, A. Khreishah, I. Khalil and J. Wu, “A Provably Efficient Online Collaborative Caching Algorithm for Multicell-Coordinated Systems”, *IEEE Transactions on Mobile Computing*, Vol. 15, No. 8, pp. 1863-1876, 2016.

[11] H. Li, Y. Huo, T. Jing and X. Cheng, “Cooperative RAN Caching based on Local Altruistic Game for Single and Joint Transmissions”, *IEEE Communications Letters*, Vol. 21, No. 4, pp. 853-856, 2017.

[12] P.T. Joy and K.P. Jacob, “Cache Replacement Policies for Cooperative Caching in Mobile Ad Hoc Networks”, *International Journal of Computer Science*, Vol. 9, No. 3, pp. 1-6, 2012.

[13] B. Zheng, J. Xu and D.L. Lee, "Cache Invalidation and Replacement Strategies for Location-Dependent Data in Mobile Environments", *IEEE Transactions on Computers*, Vol. 51, No. 10, pp. 1141-1153, 2002.

[14] G. Cao, L. Yin and C.R. Das, "Cooperative Cache-based Data Access in Ad Hoc Networks", *Computer*, Vol. 37, No. 2, pp. 32-39, 2004.

[15] M. Bilal and S.G. Kang, "Time Aware Least Recent Used (TLRU) Cache Management Policy in ICN", *Proceedings of International Conference on Advanced Communication Technology*, Vol. 32, pp. 528-532, 2014.

[16] M. Bilal and S.G. Kang, "A Cache Management Scheme for Efficient Content Eviction and Replication in Cache Networks", *IEEE Access*, Vol. 5, pp. 1692-1701, 2017.

[17] C. Bernardini, T. Silverston and F. Olivier, "MPC: Popularity-based Caching Strategy for Content Centric Networks", *Proceedings of International Conference on Communications*, Vol. 8, pp. 3619-3625, 2013.

[18] M. Cha, H. Kwak, P. Rodriguez, Y.Y. Ahn and S. Moon, "Analyzing the Video Popularity Characteristics of Large-scale User Generated Content Systems", *IEEE/ACM Transactions on Networking*, Vol. 17, No. 5, pp. 1357-1370, 2009.

[19] K. Hamidouche, W. Saad and M. Debbah, "Many-to-Many Matching Games for Proactive Social-Caching in Wireless Small Cell Networks", *Proceedings of International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, Vol. 11, pp. 569-574, 2014.

[20] N. Golrezaei, A. Molisch, A. Dimakis and G. Caire, "Base-Station Assisted Device-to-Device Communications for High-throughput Wireless Video Networks", *IEEE Transactions on Wireless Communications*, Vol. 13, No. 7, pp. 3665-3676, 2014.

[21] S. Andreev, O. Galinina, H. Tabassum, M. Gerasimenko, S.K. Singh, E. Choi, S. Ali, M.J. Dohler 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.

[22] S. Kim, "5G Network Communication, Caching and Computing Algorithms based on the Two-Tier Game Model", *ETRI Journal*, Vol. 40, No. 1, pp. 61-71, 2018.

[23] M. Zink, K. Suh, Y. Gu and J. Kurose, "Watch Global, Cache Local: YouTube Network Traffic at a Campus Network", *Proceedings of International Conference on Multimedia Computing and Networking*, Vol. 4, pp. 1-12, 2008.

[24] L. Yin and G. Cao, "Supporting Cooperative Caching in Ad Hoc Networks", *IEEE Transactions on Mobile Computing*, Vol. 5, No. 1, pp. 77-89, 2006.