Sarcouncil Journal of Engineering and Computer Sciences

ISSN(Online): 2945-3585

Volume- 04| Issue- 08| 2025

Review Article **Received:** 20-07-2025| **Accepted:** 10-08-2025 | **Published:** 30-08-2025

Optimizing Cache Layers for Faster Rendering in E-commerce Platforms: A Case Study of Megacache Migration

FNU Pawan Kumar

Birla Technical Training Institute, Pilani Rajasthan, India

Abstract: The emergence of E-Commerce has resulted in growing pressures on the Web Platform, with less latency and more performance. Optimization of the cache layer is now playing an organic role in countering the rendering delays, backend load, and augmenting the user experience. This review is conducted in the light of the developments of the cache with emphasis on the multi-layer frameworks and migration of Megacache. Attention is given to the work to capture the architectures and to capture range edge caches to application caches, as well as to distributed caches, to enhance performance. Asynchronous invalidation of caches can necessitate key issues, consistency can necessitate trade in cache invalidation, consistency cache invalidation, consistency cache invalidation issues at the level of key, cache invalidation migration. A theoretical model mentioning the significance of causation between the elements hit rate, latency, consistency, and cost of achieving a cache optimization is suggested. In between, the review discusses such emerging practices as intelligent caching and predictive prefetching. The evidence indicates that the structure of optimization in the cache is not only critical towards technical efficiency, but it is a critical factor in terms of the competitiveness of businesses in the digital business enterprise. The work gives a systematic insight into the available approaches, safe, and the places in which further research must be carried out in the process of finding out the cross-platform integration of both cache/based and safe.

Keywords: Cache optimization; e-commerce performance; Megacache migration; multi-layer caching; rendering latency; distributed caching; content delivery networks; cache invalidation.

INTRODUCTION

Digitized commerce has also created a new concentrated centre on responsivity and user experiences of websites in a sense that had never been the centre of concentration before, particularly in high traffic settings where the milliseconds loss signifies a high level of customer involvement and customer conversion rates (Pati, S., & Zaki, Y. 2025). The complex operation in which so much active data, the images of the products, is involved, until it locates the most appropriate recommendations to the user, produces some interesting challenges in the exhausting and continuous display as the e-commerce platforms continue to multiply. In this framework, the caching concept will be observed to have developed as an influential approach of performance optimizing that eliminates the latency of the relaxed meaning of upholding the expression of the information carried to the user in more frequent and more immediate closeness (Podlipnig, S., & Böszörmenyi, L. 2003).

The relevance of the optimization of the cache layer is that it can balance the speed, efficiency, and system scalability. A well-designed cache layer not only makes page rendering times better but has many benefits and the benefits of such an architecture: a reduction of the load allocated to the backend, a reduction of the bandwidth usage, as well as a higher resiliency of the e-commerce platform during peak loads. This carries especially because the total selling of the E institutions in the world is likely to exceed 8 trillion dollars in 2026, and to which the needs will arise that the speed of loading the systems with flexibility will also be additional needs (Li, W. S. *et al*., 2001; Thakur, A. 2021)

Inexpensively optimized cache mechanisms, however, can lead to the effects of stale content, inconsistency on user experience, and costly downtime that have been considered poor business competitiveness (Barroso, L. A. *et al*., 2019). In spite of the admitted inclusiveness of the caching strategies, we face challenges, which are obstinate to its actualisation. These include the problem of caching invalidation that is issued as a result of the fact that the updates introduced on the dynamic content are propagated inefficiently on the distributed systems of caches (Cao, P., & Irani, S. 1997). The other challenge is that the tradeoff between cache hit rates and storage overhead has rendered the scaling decisions very complex in the multi-layered architecture (Wu, H. *et al*., 2021). Better yet, migrating across the caching platforms, such as migrating simple caching to aspect-based caching platforms, such as Megacache, is also challenging to control in terms of compatibility, data integrity, and cost-effective percentage (Sáez, S. G. *et al*., 2014).

Existing literature has conceptualized the caching applied in distributed systems and content delivery networks (CDNs), although relatively little literature has focused on systematic approaches to optimal activations of the layers of the cache in e-

***Corresponding Author:** FNU Pawan Kumar

DOI- https://doi.org/10.5281/zenodo.17211078

commerce rendering changes pipelines (Vakali, A., & Pallis, G. 2003).

This review attempts to critically examine the aggregation optimisation cache layers, which are geared towards rapid rendering in an e-commerce setting, especially on the migration process to Megacache.

LITERATURE REVIEW

Table 1. No summary study features the optimization of caches**.**

References	Methodology	Key Findings	Relevance to Current Research
(Wang, C. *et al*., 2017)	Literature synthesis and conceptual analysis of CDN architecture, services, and deployment models.	Identified CDNs as essential for improving content availability, scalability, and performance. Analyzed edge server placement and content replication strategies.	Establishes a foundational understanding of CDNs, which remains relevant as modern network architectures evolve to include edge and cloud computing paradigms.
(Mahmood, M. R. *et al*., 2022)	Systematic literature review of integrated NCC approaches in 5G and beyond wireless systems.	Emphasized the necessity of integrating NCC to reduce latency and improve bandwidth efficiency. Discussed potential applications in edge computing and mobile cloud.	Critical for understanding next-generation wireless systems and the integration of communication and computational resources – a precursor to edge intelligence and 6G.
(Gupta, A. 2025)	Review of ML/AI models used in IoT environments. Focused on deep learning, reinforcement learning, and federated learning.	Identified challenges like scalability, energy consumption, and model accuracy. Proposed AI as a backbone for managing complexity in massive IoT environments.	Highly relevant for current research in 6G and intelligent edge computing, where AI-driven decision-making is critical.
(Zheng, Z. *et al*., 2020)	Analytical and case study-based evaluation of distributed architectures in e-commerce environments.	Highlighted the role of real-time data processing, microservices, and personalized recommendation systems.	Bridges the gap between system scalability and personalization, making it useful in CDN and edge-based e-commerce delivery frameworks.
(Pan, H. & Yang, X. 2021)	Comprehensive survey covering smart contract platforms (e.g., Ethereum), programming models, and limitations.	Addressed security, scalability, and interoperability as major bottlenecks. Introduced emerging solutions like formal verification and cross-chain platforms.	Relevant to blockchain-based CDN systems or edge platforms, especially where automation and trust are key concerns.

Publisher: SARC Publisher

Theoretical Model to be proposed on Cash Optimization of e-commerce platforms.

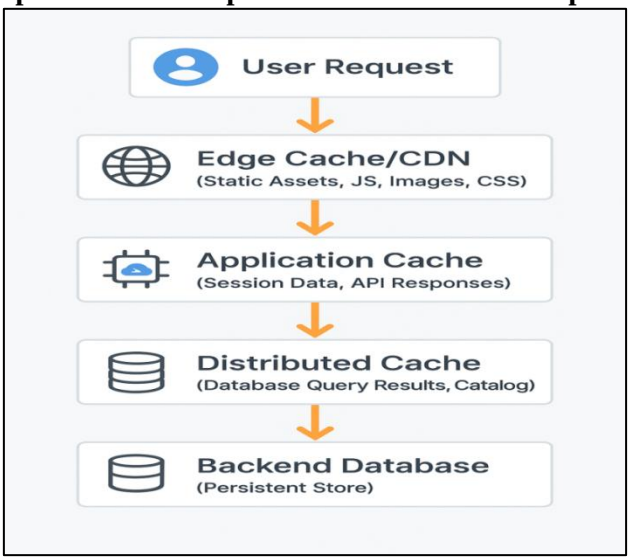

Figure 1: Multiple-layered Cache Optimization Model**.**

Description of the Model

The proposed model is the combination of multi-layer suggestions of caching strategies onto a joint network on e-commerce platforms. The role of each cache layer will be different in minimizing latency and enhancing efficiency when rendering (Zhang, X. *et al*., 2021; Stonebraker, M., & Çetintemel, U. 2018):

Edge Cache/CDN Layer

The content delivery networks result in the storage of the static content that is nearer to the users, thereby cutting the latency to the geographically spread traffic.

Application Cache Layer

Popular dynamic calls (e.g., API calls, states in a session) are stored as a cache to speed up the information application logic.

Distributed Cache Layer

Key-value systems (e.g., Redis, Memcached) save query results and product catalog on more than one node to offload the database.

Backend Database Layer

The persistent database can be used as the authoritative source, but it is accessed far less often, minimizing the costly disk operations.

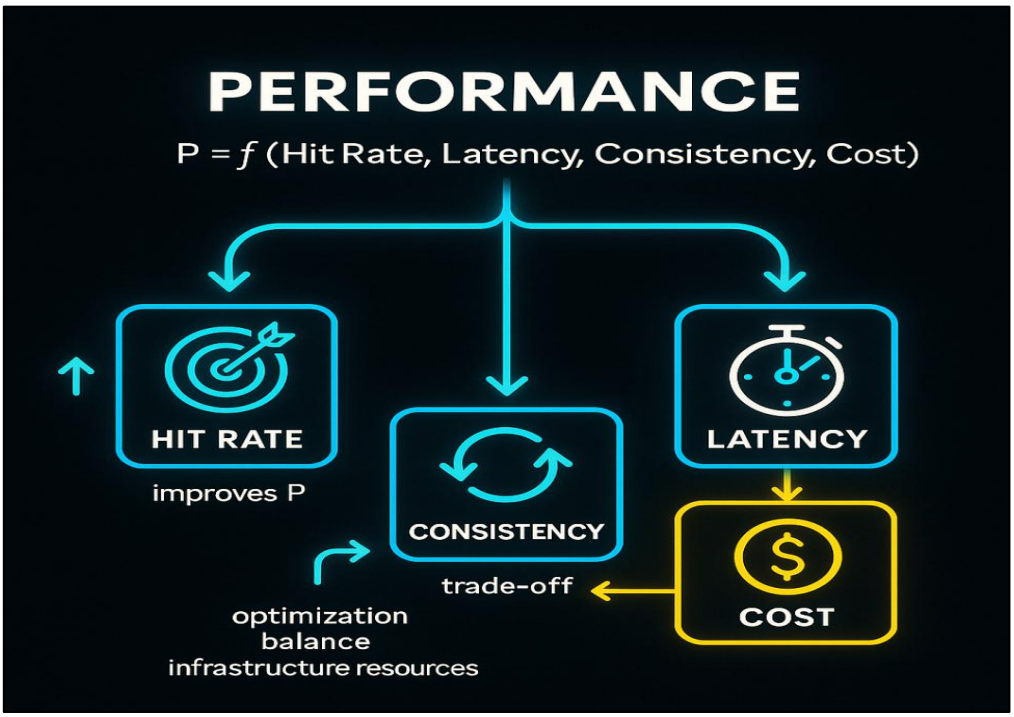

Figure 2: The Optimization Framework provides a theoretical model of optimization.

Publisher: SARC Publisher

According to the theoretical model, the performance of cache optimization (P) is determined by four significant variables: the psychological response to the cache hit, the minimization of latency, data consistency, and cost-effectiveness. It is aimed at maximizing the number of hits and minimizing the latency, maintaining the consistency, and keeping the operation costs under control (Banti, K. *et al*., 2023).

DISCUSSION

According to the latest research, it has been established that multi-tier cache structures are especially good in terms of enhancing the maximum functionality of systems with high trafficking systems. The problem of invalidating the cache due to link solving compositions, however, remains, although the element of compositions is naturally a constraint, since traces A times it is in Ukraine and is consequently perfectly consistent with distributed caches of real-time (Zhang, X. *et al*., 2021; Lim, H. *et al*., 2011). The predictive cache replacement policy and the adaptive replacement policy prove to be part of the effective strategies that can augment efficiency by addressing issues with (stale) data (Duan, S. *et al*., 2022). Movement to Advanced Caching Solutions, including frameworks, including Megacache Driven, as our Need for Elasticity and Greater Cache Hit ratios on Today's Modern Retail Platform (Stonebraker, M., & Çetintemel, U. 2018). Introducing a smart caching concept along with the application of machine learning algorithms to predictive prefetching is becoming an established approach in any attempt to relieve the dynamic nature of the retail environment with a view to reducing response times (Banti, K. *et al*., 2023). The given paradigm relies on a focus on the fact of equalizing these aspects. As an illustration, explaining how to achieve high hits in leverage of cache hit and less frequent invalidations that may cause latency and higher cost of operations unless heavily optimized (Liu, Q. *et al*., 2024) can be the most ineffective.

The form of distribution thus contrived can enhance the possibilities as well as the scaling of delays of the synchronisation among the nodes.

Real-World Application of the Proposed Model

The suggested multi-layered cache optimization model is highly appropriate to the architectural designs adopted by the modern e-commerce systems, which are strongly geared toward the ambitious low-latency user experience and high scalability to irregular traffic conditions. This model has, in effect, been applied to systems like Amazon, Alibaba, and Walmart, wherein the count and quantity of user interaction and transactions necessitate sub-second bursts of user interaction. Caching of the dynamic content, including customized recommendations, user session states, and content of shopping carts, has real-world applications in the application cache layer. Higher: Microservice - Platforms such as Shopify, Flipkart employ in-memory caches (eg, with Varnish or NGINX microcaches) to serve frequent API calls, which gradually take load over time with a low latency learning curve, saturating load in microservices with user actions and checkout requests.

Moreover, the distributed cache layer, which relies on in-memory key-value stores such as Redis and Memcached, is applied to both large systems, such as eBay and Zalando, to store commonly accessed data, such as product listing data, price data, inventory, and shipping policies. This is not only advantageous with regards to scaling heading the terminals of background relational or NoSQL databases, but also enables horizontal scaling in areas and existences of data centers. As an example, the retail architecture at Amazon uses Redis clusters to rely on such a high-speed product catalog access, and thus calls on-the-fly to pre-comprehended data on which the recommendation engine and search index will rely.

At the back-end layer, the still data storage continues to represent the truth of every transaction; however, the real-time interaction has now been offloaded into the cache layers. The interaction of mellifluence fishbones those layers, and then, the layer latency, a certified like quite a certified, and then in hit rates, and then authorize gather existing data faults, and benefit is a bit even robot under control. One such is by Walmart Labs, which made an interesting move to an adaptive cache invalidation whereby it used a stream-based event sourcing mechanism to propagate any change to the various levels of the caches in real-time to ensure some data freshness without over-provisioning of the resources.

Moreover, intelligent behaviors are also infiltrating into caching, which further stresses the latest systems. newspaper creek: mythical parent of Newspaper Creek - in consideration of its one-time and current newspaper articles. For example, Etsy and Best Buy are experimenting with the ML-enabled cache layers and are also examining the

Publisher: SARC Publisher

navigation pattern of the users, and the purchasing movement to pre-define the data likely to be required to purchase surgery, so that the actual request can be received by the application on their caches.

These pragmatic applications are used not only to confirm the practical applicability of the offered model but also to provoke the elevation of the nebulous trade-offs that are dispersed in the theoretical model. As an illustration, using edge and application caches would drastically help to increase his speed; however, some steps can be taken to address the other invalidation to ensure that we do not introduce the issue of stale information. The opportunities are evident; if the sync delay is not careful, the information consistency decays. Such issues are addressed by the retail platform, where TTL (time to live) functionality can address event trigger cache busting and probabilistic refresh caching strategies and replace the refreshing goals of optimizations as identified by the model above. Comprehensively, the multi-tiered conceptual logical cache architecture is not only theoretical anymore, but in practice, in the retail world globally, it is deployed. These implementations are evidence-based proofs of concept, aimed at a compromise between speed, scale, and cost-efficiency, the high-demand settings where responsiveness expectations of users are not bargaining variables.

FUTURE DIRECTIONS

The trend of intelligent and adaptive cache optimization is on the rising curve of advancement of nowadays e-commerce platforms. One of the broad directions of future research is machine learning-style prefetching, in which the predictor model can be aware of the latencies that users are about to experience, prior to their actual occurrence, and preload the content in the cache memory so as to minimize the perceived latencies even during busy traffic sessions. The other significant trend is the combination of the caches with cross-platform, which is caused by the creation of mobile devices, the Internet of Things (IoT) systems, and immersive technologies (augmented and virtual reality). There must be a smooth performance on these heterogeneous platforms in the future version of the cache.

In addition, the demand for sustainable computing is driving the energy efficiency caching model design that looks into the trade-off between system performance and, in a more eco-friendly way, mitigates the eco-sustainability of large-scale e-Commerce infrastructure. Moreover, the users of blockchain consider blockchain-enabled cache validation as a possible trust and transparency logging/consistency logging solution in distributed cache systems, particularly when it comes to invalidation between many different nodes. Lastly, the future of caching solutions will be characterized by personalization-conscious solutions that will enable platforms to provide countless differentiating user-held user experiences, but note that such a form of optimization of the tests and content freshness factor.

CONCLUSION

The concept of optimization of the cache layer has grown to be the pillar of providing excellent performance in contemporary e-commerce infrastructure. By using the creators relying on the multiple-layer cache frameworks, the platforms can reduce the turnaround time in employing rendering, balance the workloads on the backends, and provide an unlimited user selection journey. The things on a practical level have to do with it, its challenges, and its opportunities. One theoretical model presented in this review is used to emphasize the main basis of successful optimization: hit rate times latency, consistency, and cost in a cache. Although quite a lot was done to correct the performance in extant strategies, there are still some problem areas related to the personalisation and cash invalidation synchronisation. Looking into intelligent energy-efficient and cross-platform caching models suggests the impending era of learning and practice. It is in the end less the Cache optimization in purely technical terms, not only the imperative, but in Terms of strategic silence in terms of the competitiveness of the business of e-Commerce.

REFERENCES

1. Pati, S., & Zaki, Y. "Evaluating the Efficacy of Next. js: A Comparative Analysis with React. js on Performance, SEO, and Global Network Equity." *Companion Proceedings of the ACM on Web Conference 2025*. (2025)
2. Podlipnig, S., & Böszörmenyi, L. "A survey of web cache replacement strategies." *ACM Computing Surveys (CSUR)* 35.4 (2003): 374-398.
3. Li, W. S., Selçuk Candan, K., Hsiung, W. P., Po, O., & Agrawal, D. "Engineering high performance database-driven e-commerce web

Publisher: SARC Publisher

sites through dynamic content caching." *International Conference on Electronic Commerce and Web Technologies*. Berlin, Heidelberg: Springer Berlin Heidelberg, (2001).
4. Thakur, A. "Trends and analysis of e-commerce market: a global perspective." *International Journal of Applied Marketing and Management* 6 (2021): 11-22.
5. Barroso, L. A., Hölzle, U., & Ranganathan, P. "The datacenter as a computer: Designing warehouse-scale machines." *Springer Nature*, (2019)
6. Cao, P., & Irani, S. "{Cost-Aware}{WWW} Proxy Caching Algorithms." *USENIX Symposium on Internet Technologies and Systems (USITS 97)*. (1997)
7. Wu, H., Fan, Y., Wang, Y., Ma, H., & Xing, L. "A comprehensive review on edge caching from the perspective of total process: Placement, policy and delivery." *Sensors* 21.15 (2021): 5033.
8. Sáez, S. G., Andrikopoulos, V., Leymann, F., & Strauch, S. "Evaluating caching strategies for cloud data access using an enterprise service bus." *2014 IEEE International Conference on Cloud Engineering*. IEEE, (2014).
10. Vakali, A., & Pallis, G. "Content delivery networks: Status and trends." *IEEE Internet Computing* 7.6 (2003): 68-74.
11. Wang, C., He, Y., Yu, F. R., Chen, Q., & Tang, L. "Integration of networking, caching, and computing in wireless systems: A survey, some research issues, and challenges." *IEEE Communications Surveys & Tutorials* 20.1 (2017): 7-38.
12. Mahmood, M. R., Matin, M. A., Sarigiannidis, P., & Goudos, S. K. "A comprehensive review on artificial intelligence/machine learning algorithms for empowering the future IoT toward 6G era." *IEEE access* 10 (2022): 87535-87562.
13. Gupta, A. "Simplifying Distributed Systems for e-Commerce Personalization." *Journal Of Multidisciplinary* 5.7 (2025): 919-929.
14. Zheng, Z., Xie, S., Dai, H. N., Chen, W., Chen, X., Weng, J., & Imran, M. "An overview on smart contracts: Challenges, advances and platforms." *Future Generation Computer Systems* 105 (2020): 475-491.
15. Pan, H., & Yang, X. "Management model on electronic commerce data based on cloud computing." *Wireless Communications and Mobile Computing* 2021.1 (2021): 5510002.
16. Zhang, X., Qi, Z., Min, G., Miao, W., Fan, Q., & Ma, Z. "Cooperative edge caching based on temporal convolutional networks." *IEEE Transactions on Parallel and Distributed Systems* 33.9 (2021): 2093-2105.
17. Lim, H., Fan, B., Andersen, D. G., & Kaminsky, M. "SILT: A memory-efficient, high-performance key-value store." *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles*. (2011).
18. Duan, S., Wang, D., Ren, J., Lyu, F., Zhang, Y., Wu, H., & Shen, X. Distributed artificial intelligence empowered by end-edge-cloud computing: A survey." *IEEE Communications Surveys & Tutorials* 25.1 (2022): 591-624.
19. Stonebraker, M., & Çetintemel, U. ""One size fits all" an idea whose time has come and gone." *Making databases work: the pragmatic wisdom of Michael Stonebraker*. (2018): 441-462.
20. Banti, K., Louta, M., & Baziana, P. "Data quality in human-centric sensing-based next-generation IoT systems: A comprehensive survey of models, issues, and challenges." *IEEE Open Journal of the Communications Society* 4 (2023): 2286-2317.
21. Liu, Q., Yang, Y., Du, D., Xia, Y., Zhang, P., Feng, J., ... & Chen, H. "Harmonizing efficiency and practicability: optimizing resource utilization in serverless computing with JIAGU." *2024 USENIX Annual Technical Conference (USENIX ATC 24)*. (2024).

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:
Kumar, P. "Optimizing Cache Layers for Faster Rendering in E-commerce Platforms: A Case Study of Megacache Migration" *Sarcouncil Journal of Engineering and Computer Sciences* 4.8 (2025): pp 1033-1038.

Publisher: SARC Publisher