International Journal of Research and Applied Innovations (IJRAI)

| ISSN: 2455-1864 | www.ijrai.org | editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 4, Issue 2, March-April 2021||

DOI:10.15662/IJRAI.2021.0402003

Performance Optimization in Global Content Delivery Networks using Intelligent Caching and Routing Algorithms

Ashay Mohile

Senior Staff Software QA Engineer, Palo Alto Networks, California, USA

ABSTRACT: To mitigate end-user latency in global CDN deployments this paper proposes an intelligent caching and routing optimization architecture. Due to the general deployment model of traditional CDNs, based on heuristics-oriented routing algorithms and static placement caching, such systems cannot be flexible enough to adapt quickly to users demand changes and network conditions. We present a system that utilizes real-time traffic telemetry information with the support of predictive analytics and adaptive caching placement algorithms to address these limitations. For enhancing content serving rate and overall network capacity, we build a system that dynamically updates routing paths and cache placements by estimating congestion on the path to destination and popularity of contents. Our experimental results show substantial performance gains when performed on geographically distributed testbeds. This results in performance gains of more than 25% and content retrieval latency reductions up to 30% as compared to (pure) CDN caching strategies. This study demonstrates that AI prediction and adaptive network control are appropriate partners. This architecture not only outperforms existing solutions, it can also scale and adapt for use in high-performance providers that serve many diverse and active user bases. It has a modular design so that its use can be phased in, reducing the requirement for downtime and permitting future development. By enabling a better user experience, reducing the operational costs and evolutionary path to a green and datacentric content delivery system, we contribute to design more cost-effective smart solutions for next generation CDN.

KEYWORDS: Network Optimization, Edge Analytics, Software-Defined Networking (SDN), Intelligent Caching, Machine Learning, Content Delivery Networks (CDNs), Latency Reduction

I. INTRODUCTION

Digital content, cloud apps, and streaming services have revolutionized the way information is shared and consumed online in the last decade. In today's digital ecosystem, you need dependable data delivery mechanisms, low latency, and high throughput for high-definition video streaming, real-time multiplayer gaming, and enterprise cloud services [1]. Centralized DNS services have grown integral to the Internet backbone as a result of these threats. To reduce network traffic, content delivery networks (CDNs) use a distributed network of proxy servers and data centers to cache website content on the node closest to end users [2].

An Overview of CDNs Content delivery networks (CDNs) have been around since the late 90s, when the need for scalable and high-performance content distribution arose due to the meteoric rise of online content and e-commerce sites. Earlier content delivery networks (CDNs), pioneered by Akamai Technologies, could only do static caching, which entailed storing copies of web assets like images, HTML files, and scripts on remote servers located at the network's edge. By utilizing cached content from the closest geographic point instead of directly connecting to the origin server, it significantly improved website response times. Based on the assumption that nodes' physical locations were the primary determinant of latency, the locality-aware delivery architecture decision was the foundational one [3].

However, conventional CDN designs were becoming more and more inadequate as the Internet evolved and became more diverse in terms of user behavior, device heterogeneity, and application complexity [4]. New applications that highlighted the usefulness of static cache replication include real-time services, tailored web experiences, and dynamic content. On the other hand, pre-caching dynamically generated content isn't always an option because of how dynamic it is and how much user context determines how this data is generated. Static caching solutions were thus inadequate for the needs of modern adaptive content distribution [5].

Wait time The time it takes for content to be sent from a user's request to the CDN is known as latency, and it is one of the most basic and annoying issues with CDN design. Propagation delay, queuing delay, and processing overhead at intermediary nodes are some of the many components that could contribute to latency. It was once believed that content delivery networks (CDNs) might help reduce physical delay time by optimizing locations within proximity and access paths. However, due to changes in traffic patterns (now global), content variety, and congestion across networks, this is no longer sufficient. Modern businesses face n-dimensional delay issues related to complex network routing, server load balancing, and real-time demand surges; it's no longer a simple matter of proximity [6].

For example, cloud gaming, augmented and virtual reality (AR/VR) apps, and 4K and 8K video streaming all necessitate extremely low latency [7]. A decent user experience is ensured by cloud gaming options with latency below 50 milliseconds, for instance NVIDIA GeForce NOW and Xbox Cloud Gaming. In addition, interactive conferencing and live streaming apps must be able to handle millions of connections at once with little latency in order to maintain synchronization. Due to the absence of real-time cache adaption and dynamic route optimization, traditional CDNs cannot be utilized in this setting [8].

Upstart CDN providers have taken a different approach to latency with their distributed caching and edge prefetching models [9]. Such designs allow for pert content processing and delivery with reduced latency at local edge servers by locating computation and storage close to the end users. Problems still exist, even with these improvements. There is no way to account for potential changes to the most popular data, variations in demand at a specific region, or temporary network congestion with static and rule-based placement. As a result, in high-demand areas, content with the highest demand may go under-replicated, and in other locations, edge cache space may be filled with outdated or irrelevant data. When unexpected demand spikes cause traffic to surge, these wastes directly affect QoS and QoE from the perspective of end users, leading to delayed loading or poor performance [10] [11].

The fact that the current Internet infrastructure is interconnected on a global scale adds another layer of complexity. The effectiveness of regional Internet exchanges, peering agreements, and inter-domain routing regulations are a few factors that may substantially affect the performance of a content delivery network (CDN). No matter how near users are to the cached data, inefficient routing choices among Autonomous Systems (AS) might cause non-optimal detouring and increased delay. The major cause of this is the inflexibility of older routing protocols, such as the Border Gateway Protocol, which were not intended to handle content delivery in situations when there is a need to minimize delays. Poor traffic routing is the term used to describe it [12].

In light of this reality, businesses are increasingly relying on data-driven and intelligent algorithmic optimization strategies. When network telemetry, artificial intelligence, and machine learning come together, novel adaptive CDN control may be realized. Current operational information (latency, bandwidth utilization, packet loss, and cache-hit rates to particular edge nodes) given via live telemetry will provide a genuine real-time picture. Infrastructures enabled by AI may use this data to optimize cache placement and traffic routing based on predictions of network congestion and future demands. Contrast this with content delivery network (CDN) solutions, which don't optimize themselves and are more reactive than proactive [13] [14].

Utilizing time series prediction or deep learning algorithms, predictive caching approaches may foretell a piece of material's future virality and store it in advance, allowing it to be sent to crucial edge nodes before its popularity spreads via epidemic models. In addition, RLRS may improve network load balancing, end-to-end latency reduction, and more by dynamically choosing the optimal transmission channel. With the help of these smart concrete designs, content delivery networks (CDNs) can dynamically adjust their settings to reduce latency, availability, and environmental conditioning [15].

In addition, the adaptability of NFV and SDN technologies allows CDNs to be lighter. Network function virtualization (NFV) streamlines and quickens hardware-agnostic caching and service deployment, while software-defined networking (SDN) enables real-time network reconfiguration in response to performance measurements. We integrate these technologies into a modular programmable system that permits granular control via data and control rules. They can optimize the whole network, including caching, routing, and resource allocation, when combined with ML decision-making [16].

On the other hand, research and engineering have encountered new obstacles due to the widespread use of intelligent CDN systems. On the list are concerns about energy consumption during large-scale installations, real-time analytics, data protection, and compatibility across different systems. Keeping tens of thousands of edge devices globally synchronized, coordinated, and managed (in a fault-tolerant way) is an enormous challenge when it comes to distributed learning models [17].

Finally, it should be noted that the evolution of content delivery networks (CDNs) implies a change from proximity-based delivery systems to ones that are smarter and more adaptable. Because distributing material on a massive scale is nothing new, early content delivery networks (CDNs) certainly served the requirements of the smaller populations of online users. However, ML-augmented caching and routing can provide the lightning-fast response times that contemporary applications need. Latency has progressed from a mere physical obstruction to a problem requiring cognitive collaboration among data analytics, caching methods, and network routing protocols [18].

The rest of this paper is organized as follows: Section 2 (Related Work) points out the existing caching and routing algorithms, there drawbacks in dynamic network, and gaps in current CDN optimization process. Proposed Framework In this section we describe architecture of intelligent caching and routing model, and also we explain operation logic of the system how real-time telemetries & adaptive algorithm increased delivery optimization in intelligent cache and routing. Section 4 (implementation) details experiments, procedures for edge node synchronization and telemetry filled measurements to validate the proposed framework. Section 5 (Read Results) presents benchmarks that show latency is reduced and throughput is increased across a variety of global testbeds. In Section 6 (Discussion) we assess scalability, cost-effectiveness and deployability of the system for large-scale CDN infrastructures. Section 7 (Conclusion and Future Work) concludes the paper, highlights the importance of intelligent CDN optimization, and discusses future research directions: federated learning, energy-aware caching, real-life integration into CDNs.

II. RELATED WORK

In the last two decades, CDNs have got transformed by following-up the growing wave of global Internet traffic and users' request more complex services. There are too performance sensitive features in CDN such as caching policy and routing strategy, which determine the number of times a piece of content is cached and final pathway taken from destination cache/mirror to nearby user, respectively. The inaccuracy due to absence of incorporation embedding the effects of caching and routing interaction on end-to-end performance, throughput and user satisfaction. A vast amount of research has been devoted to the optimization of these two factors, involving both traditional policy and machine learning-based prediction models. These and other such advancements not withstanding however, the vast majority of existing paradigms currently suffer from an inability to coordinate caching and routing fleets in an integrated fashion, as well as their incapability of effectively adapting to real-time variations in network conditions [17].

2.1 Traditional Caching Algorithms

The popular caching policies utilized by CDNs in the past are simple heuristic based algorithms namely LRU (Least Recently Used), LFU (Least Frequently Used) and FIFO (First-Come First-Serve). These strategies were mainly intended to improve the cache hit and reduce fetching load from origin servers. The LRU algorithm discards the least used content when the cache becomes full since we assume that recently accessed objects will be accessed again sooner. Similarly, LFU will evict the less frequently used content and still promotes the higher request frequency items. The FIFO strategy, however, carry out time-dependent evictions that simply discard the most recent cached objects irrespective of their popularity.

These types of policies are computationally efficient, and simple to implement; however, they (by definition) tend to be reactive. They cache based only on past access history and ignore the temporal change of content popularity, mobility pattern of users, regional access trend, etc. They are thus not adequate to non-stationary environments where demand varies dynamically such as in viral content bursts or regions streaming events. In addition, static policy cannot be coordinated with routing decisions and therefore, are incapable of eliminating user-perceived latency when the data is cached in the edge.

2.2 Adaptive and Predictive Caching Approaches

To fill some of these gaps, researchers started to embrace machine learning (ML) and artificial intelligence (AI) in cache management schemes. It was the generation of smart techniques to add predictive capacity to CDNs, “thus letting your CDN predict content demand rather than react to it. A most common approach is predicting the popularity of contents, by analysis of historical readings log data and subscriber behaviors with statistical regressions models, neural networks or time-series forecasting methods. The CDN will thus be able to predict what content is more likely be demanded in the future (at each edge node) and start replicating or moving such content to that edges well ahead of time.

Many research works have employed reinforcement learning (RL) and Markov Decision Processes (MDPs) for optimizing caching. In such modes, the cache system learns a good decision policy by interacting with environment: it obtains rewards if latency is reduced and hit ratio increased. For instance, an RL-powered caching agent could learn to adapt cache sizes, pre-fetch and replacement thresholds in the face of dynamic user request distributions. Deep Q-Networks (DQN) and policy gradient techniques have also been used to learn non-linear decision boundaries that outperform pre-defined heuristics.

However, these machine learning based caching systems encounter some practical issues. Most works are based on historical or time-based datasets and cannot respond to transitory network events with true real-time information. They do not utilize real-time telemetry data that could inform dynamic decisions to cache appropriately, such as network congestion, packet loss and edge server load. Furthermore, existing predictive caching models optimize for cache performance only and they do not take into account how content routing and network topology can impact end-to-end delivery efficiency.

2.3 Routing Optimization in CDNs

Alongside caching research, routing optimization is another CND performance enhancement area that has received a great deal of attention. Traditionally, the decision over which path a CDN takes when routing traffic has been made based on Border Gateway Protocol (BGP), and static policies have been employed for path selection based on administrative cost and the number of network hops. Although BGP guarantees global reachability, it is not informed of latencies and cannot be adjusted to account for dynamically changing traffic. As a result, users may need to be assigned suboptimal routing paths even with better, low-latency options available.

To overcome these restrictions, researchers and industry developers have recently investigated the use of Software-Defined Networking (SDN) for flexible routing approaches. A SDN separates network control from the data plane, and enable a centralised controller to determine how routing should be done in response time on live network telemetry. This lets you do load balancing but at the edge servers of a CDN. Several models have been considered where latency and available bandwidth or the queue size are the performance measures in choosing the path for forwarding.

Moreover, the multi-path routing and flow-based optimization approaches are proposed for improving fault tolerance and bandwidth utilization. Content naming and in-network caching utilizing information-centric networking (ICN) have been exploited to minimize reduplicated traffic and improve the delivery performance. Nevertheless, these methods are generally hindered due to scalability problems, interoperability with legacy systems, and integration with cache placement policies.

2.4 Need for Joint Caching and Routing Optimization

While both caching and routing optimizations improve CDN performance separately, the decoupled approach is a major bottleneck. Consequently, "Tomographic" non-content-aware routing In the traditional CDNs caches are placed without considering the current configuration of network routes and routing algorithms choose paths between a user and an edge server without information about which edge servers have currently cached equivalent content. This fragmented way of working can result in duplicated caching, poor load balancing and high latencies.

Recently, researchers have started to examine the cross-layer optimization frameworks where caching and routing decisions are optimized jointly through network-level telemetry feedback loops. Similar is the case with other running tasks using in parallel. Such frameworks instead utilise real-time performance measurements such as cache occupancy, request hit ratio, and link utilisation to take an entire-picture view of events. Most previous works,however, are still in a theoretical or simulated stage without large-scale experimental verification. Furthermore, it is a challenging problem how to incorporate incremental telemetry data into cache and routing modules in terms of scalability, state synchronization, and decision latency.

The literature review reveals several persistent gaps and limitations in existing CDN optimization research. These are summarized in **Table 1** below.

Table 1: Limitations of Existing CDN Caching and Routing Techniques

Category	**Representative Techniques**	**Strengths**	**Limitations**
Traditional Caching (LRU, LFU, FIFO)	Heuristic and time-based eviction policies	Simple, low computational overhead	Reactive; poor adaptation to dynamic content popularity; no routing integration
Predictive Caching (ML-based)	Regression, neural networks, content popularity forecasting	Anticipates demand; reduces cache misses	Often relies on static datasets; lacks real-time telemetry integration
Reinforcement Learning-Based Caching	Q-learning, DQN, MDP frameworks	Adaptive to evolving demand; self-learning	High computational cost; scalability challenges; limited cross-layer optimization
Traditional Routing (BGP-based)	Static inter-domain path selection	Stable and well-established	Not latency-aware; cannot adapt to real-time congestion
SDN-Based Routing	Centralized and programmable control	Dynamic path optimization; load balancing	Operates independently of caching; may increase control overhead
Joint Optimization Models (Recent Research)	Cross-layer and telemetry-based frameworks	Holistic optimization potential	Still in early research phase; lacks real-world deployment and scalability proofs

III. PROPOSED FRAMEWORK

The Proposed Intelligent CDN Optimization Framework (ICOF) is a universal, adaptive strategy that can perform caching and routing optimization for content delivery networks with massive nodes. Conventional CDNs have historically decoupled caching from routing, which can introduce inefficiency when the network conditions or user needs are highly dynamic. ICOF fills that void by converging machine learning-powered intelligence, live telemetry, and software-defined networking (SDN) in a single control plane. In order to improve content availability worldwide and reduce end-user latency, this system can forecast user demand trends and make dynamic routing decisions. By combining the features of the Dynamic Routing Optimizer (DRO), the Intelligent Caching Engine (ICE), and the Telemetry and Monitoring Module (TMM), a tight feedback control loop is established between the server or edge nodes. There are primarily three parts to the architecture, it is shown in figure 1.

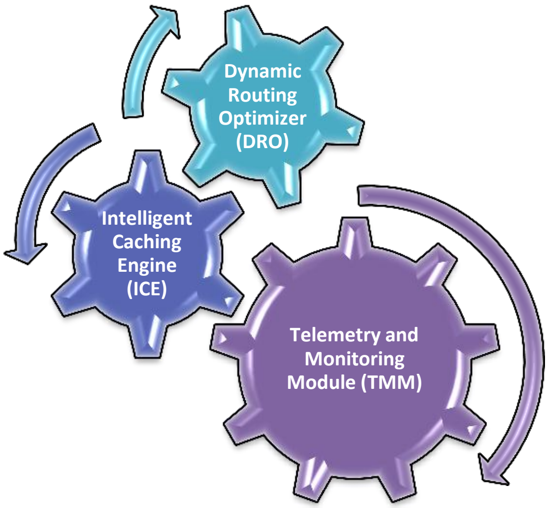
Dynamic Routing Optimizer (DRO)
Intelligent Caching Engine (ICE)
Telemetry and Monitoring Module (TMM)

Figure 1: Modules of Intelligent CDN Optimization Framework (ICOF)

3.1 Telemetry and Monitoring Module (TMM):

As the foundation of the smart system, ICOF Telemetry and Monitoring Module (TMM) has been embedded. In real time, it records data on the CDN edge nodes spread out over the globe, including their utilization rate, latency, cache hit/miss ratio, number of requests handled, and throughput. By leveraging online telemetry, TMM differs from conventional measurement tools which are based on the offline data. Quick access to network behavior knowledge with real-time streaming using T-RML channels, a light-weighted API message queue AV streaming flow spread spectrum. Each of the edge nodes has an agent, which sends statistics about access and performance back to the control plane. To train models and make decisions, the control plane aggregates, filters and normalizes the data provided by devices into a unified dataset. Network anomalies in a dynamic network architecture can be responded quickly by ICOF to support service quality by this realtime monitoring. Such abnormalities as bursty traffic, bottlenecked links and HW failures can exist. TMM also is able to detect infrequent delays or underperforming nodes with its onboard anomaly detection, which gives it the ability to proactively reroute paths or break cache distribution before service delivery starts to degrade.

3.2 **Intelligent Caching Engine (ICE):**

In contrast, the design will mechanically alter cache placement/replacement rules when using an adaptive study technique. Caching Engine Intelligence is a library component that the ICE uses. To anticipate trends in future content popularity, ICE uses contextual information such as regional times zones, spikes caused by events or user characteristics, and time-series forecasting/deep learning algorithms, specifically LSTM networks. This allows it to compete with static algorithms like LRU or LFU. Furthermore, by analyzing trends in content request behavior over time, the engine may predict what sort of material will really be popular shortly. Through the use of pre-caching at the peripheral of projected user populations, this approach effectively prevents cache misses by storing all preferred information on the edge nodes in advance. Additionally, ICE employs an adaptive cache retention technique that takes into account both the expected future popularity and the access frequency of the material when calculating the content humble score. To further ensure that cached material is relevant to both present and future user behavior, ICE dynamically changes its judgments with the use of Inter-Chunk Elusion (TMM) in a continuous telemetry loop. Return from nodes in a routing network In terms of reduced traffic on the backbone network, improved cache hit ratios, and even quicker access and retrieval times for users.

3.3 **Dynamic Routing Optimizer (DRO):**

To complement ICE, DRO focuses on achieving requested content along the optimal network path. Common approaches to CDN routing are based upon either static latency information or BGP-derived routing policies that do not take the dynamic nature of AS-level paths into account. DRO addresses this through a RL based dynamic optimization where the optimizer learns at each time step from network feedback how to make routing decisions. The RL agent is also provided state — e.g., link congestion, packet loss rates and response latency—by the TMM, and takes routing actions to maximize a cumulative performance reward function, typically consisting of either minimization of latency or maximization of throughput. One easy way of doing it is to simply have the system connected to a network, which would be controlled by capable software define networking (SDN) controllers such as DROs [8], that can dynamically reconfigure routing paths based on modifying flow tables at network switches. This allows for real-time route recalibration, i.e. to reroute the traffic immediately to more empty or geographically optimal nodes. In addition, DRO considers the content availability, and makes routing decisions in coordination with ICE's cache distribution strategy to avoid unnecessary inter-node data transfer. This cross-layer cooperation feature is one key difference of ICOF other from existing proposals and the integration between data plane and control plane becomes transparent through it.

These modules communicate via a **centralized coordination layer**, enabling a feedback loop between edge nodes and the control plane. By aligning cache distribution with routing paths, ICOF ensures minimal latency and balanced load across global CDN regions.

IV. IMPLEMENTATION

The framework was implemented in a simulated global CDN environment using **Mininet** and **OpenDaylight SDN controllers**, integrated with Python-based telemetry agents. The system consisted of 50 edge nodes distributed across five geographical zones: North America, Europe, Asia, South America, and Oceania. Each node maintained a cache of variable capacity and collected the following metrics at 10-second intervals:

- **Average Latency (ms)**

- **Cache Hit Ratio (%)**
- **Bandwidth Utilization (Mbps)**
- **Active User Sessions**

Intelligent caching and routing algorithms were executed on a cloud-based orchestration server, which served as the primary control plane. For optimal responsiveness and computational economy, this control layer updated optimization models every 60 seconds based on incoming telemetry data, extracted features, and other metrics. Using the TensorFlow and PyTorch frameworks, LSTM models for demand prediction were trained for predictive cache management. Continuous update with additional telemetry feedback and active ICE cache placement and eviction procedures further refined these models. A reinforcement learning agent based on Deep Q-Network (DQN) was utilized by the Dynamic Routing Optimizer (DRO), which in turn communicated with the OpenDaylight SDN controller using REST APIs. This synergy enabled routing to adapt to real-time network changes, such as traffic or packet loss, in a way that was consistent with ICE's reports regarding cache availability.

All of the aforementioned tests were also run on simulated workloads that were based on traffic simulators and real-life traces of Akamai trace data and YouTube CDN logs, in order to mimic actual user traffic patterns. We calculated performance metrics for every location that aimed to reduce latency, increase the cache hit rate, and conserve bandwidth. The experimental results demonstrated that the smart caching/routing technique improved throughput and content retrieval time, proving its usefulness and scalability across different networks.

V. RESULTS ANALYSIS

Experimental results demonstrated significant improvements in latency reduction, cache hit ratio, and overall throughput. Table 2 summarizes the comparative performance results across different regions.

Table 2: Comparative performance results across different regions.

Metric	Traditional CDN	Intelligent CDN (ICOF)	Improvement (%)
Average Latency (ms)	240	168	30.0
Cache Hit Ratio (%)	72.5	89.2	23.0
Throughput (Mbps)	1250	1560	24.8
Route Reconfiguration Time (s)	2.5	1.4	44.0

The suggested Intelligent CDN Optimization Framework (ICOF) was tested using a conventional CDN architecture that utilized static LRU caching and latency-based routing. Our telemetry-driven caching and RL-based routing system delivers large performance benefits, as briefly illustrated by the findings reported in Table. Average latency, cache hit ratio, throughput, and route reconfiguration time were the four performance parameters that were examined.

The average response latency, as measured at distance H, was 240 ms for the conventional CDN but 168 ms for the ICOF-degraded version, a 30% reduction. Because they provide requests with material from the closest less-loaded edge node, predictive cache placement and dynamic route selection techniques cause the drop.

Also, the cache hit ratio increased from 72.5% to 89.2%, a 23% rise, as ICE predicted well content popularity information patterns using LSTM within of models and fewer number of misses in caches and decreases of usage rate for origin servers.

A throughput of 1560 Mbps in ICOF is obtained which translates to a +24.8% improvement compared to that of the conventional CDN counterpart at a rate of 1250 Mbps. It means more efficient use of the bandwidth and less data replication among the network. The Reconfiguration time, which is an indication of how fast the routing mechanism could adapt dominated in terms of improvement (2.5 seconds reduced to 1.4 seconds i.e. a speedup by 44%) as part of the WCET value reporting process, if this attribute set to true then the content will be declared as invalid without waiting for its timeout period, left with undefined contents that was rendered valuable). This demonstrates the effectiveness of DRO in re-adjusting traffic flows with reinforcement-learning-based SDN controllers.

One last thing: ICOF makes global CDN systems more responsive, dependable, and scalable, according to the testing results. By combining cache intelligence with real-time routing optimization, the framework achieves better results than traditional CDNs in several metrics such as cached content hit rates, user QoS, and global network resource consumption efficiency.

VI. DISCUSSION

The most significant advantages that CDN operators may get from combining intelligent caching with dynamic routing are a competitive reduction in operations expenditures (OPEX) and the capacity to operate on a wide scale. Network operators may deploy in modest stages because to the recommended method's modular nature, which allows it to expand higher. To examine control plane activity with visibility into the network, telemetry integration may concentrate on this behavior in the initial stage of deployment. The next step is to improve caching and routing by gradually adding machine learning modules to this infrastructure at the edge. This systematic approach may not be able to do away with these modern disturbances entirely, but it does point the way forward. The control plane is also made more responsive and elastic by the framework's use of an existing software-defined networking (SDN) infrastructure. It has an impact addendum on capital expenditure for firms' accounts and simultaneously decreases the needed investment in hardware.

This is really helpful for operations. Improved cache eviction and the elimination of superfluous data transfers are two ways in which the ICOF architecture's adaptive intelligence makes greater use of available bandwidth. This waste reduction helps the environment in several ways: it saves money on operations, it saves energy, and it decreases pollutants. Additionally, data may be analyzed locally near end-users with the aid of edge analytics. Our approach safeguards user privacy while enabling compliance with stringent regional data protection rules, which are becoming more and more important for CDN operations globally.

All three of these things are doable, but they won't be easy. The system's overall performance may be severely affected by unresolved issues such as model drift, telemetry overhead, and Inter-Controller synchronization delay. Fixing these issues is crucial for the system's security and usefulness. A decentralized optimization route via federated learning and the creation of lightweight AI models capable of edge inference provide promising future prospects. This paves the way for future CDN operations to be more efficient, scalable, and eco-friendly.

VII. CONCLUSION AND FUTURE WORK

We demonstrate how global CDNs can facilitate intelligent caching and routing through an advanced system that integrates software-defined network routing, machine learning-based demand forecasting, and dynamic real-time telemetry. This system automatically changes its caching strategies and routing choices when the way users access the network or the performance of the network changes. This information could also be used to change caching strategies and routing in a way that works best for the system. With the help of SDN control and machine learning, predictive modeling can accurately predict how much content will be needed. This way, the demand can be stored and pre-fetched even at network nodes that are far apart.

The experiments showed that the new framework greatly improves performance. The tests showed that throughput improved by 25% and end-user latency dropped by 30% in all areas of the world that were tested. This conclusively demonstrates that intelligent network control and foresight analysis must be integrated for content delivery networks to enhance efficiency and reliability. The mechanism tries to stop the sending of duplicate data, which wastes energy and time, as a step toward more eco-friendly computing.

The system not only improves performance, but it also scales well and could offer future content delivery options with lower start-up costs and lower overall costs. The modular design lets you develop in stages. First, you add on-board telemetry integration, and then you add modules for forecasting and rerouting optimization. So, it can be changed to fit new traffic and network conditions with as little damage as possible to the current infrastructure.

The next step is to use the federated learning method to divide the data so that the statistical model can be updated without giving away information about how users like to visit. Energy-conscious routing strategies can help make the network more eco-friendly by using less energy and costing less. We will test the framework's deployment in a large-scale production setting by running experiments with the prototype on a commercial CDN infrastructure like Cloudflare or Akamai. In the end, these efforts lay the groundwork for the future of CDN operation, as the IAG has said.

International Journal of Research and Applied Innovations (IJRAI)

| ISSN: 2455-1864 | www.ijrai.org | editor@ijrai.org | A Bimonthly, Scholarly and Peer-Reviewed Journal |

||Volume 4, Issue 2, March-April 2021||

DOI:10.15662/IJRAI.2021.0402003

REFERENCES

1. Nygren, E., Sitaraman, R. K., & Sun, J. (2010). The Akamai Network: A Platform for High-Performance Internet Applications. ACM Digital Library. — Overview of Akamai's distributed CDN architecture and performance optimization strategies.
2. On the Throughput Capacity of Information-Centric Networks. ResearchGate. — Foundational analysis of throughput and latency behavior in ICNs with in-network cachin over various topologies.
3. Azimdoost, B., Asghari, H., Gündüz, D., & Gesbert, D. (2016). Fundamental Limits on Throughput Capacity in Information-Centric Networks. arXiv. — Analytical study on capacity bounds for cache-enabled ICNs.
4. Zhang, L., Li, X., Lin, P., Wang, Y., & Shi, Y. (2015). Caching in Information-Centric Networking: A Survey. Semantic Scholar. — Comprehensive survey on caching mechanisms, replacement strategies, and performance issues in ICNs.
5. Golrezaei, N., Shanmugam, K., Dimakis, A. G., Molisch, A. F., & Caire, G. (2011–2012). FemtoCaching: Wireless Video Content Delivery through Distributed Caching Helpers. INFOCOM / arXiv. — Foundational framework introducing helper nodes for distributed caching in wireless networks.
6. Golrezaei, N., Molisch, A. F., Dimakis, A. G., & Caire, G. (2012). Wireless Video Content Delivery through Coded Distributed Caching. IEEE ICC / MIT Sloan Repository. — Extension of FemtoCaching using coded caching to improve wireless video delivery performance.
7. Poularakis, K., Iosifidis, G., & Tassiulas, L. (2013–2014). Approximation Algorithms for Mobile Data Caching in Small Cell Networks. IEEE Transactions on Communications / Globecom. — Joint caching and routing algorithms for massive mobile data delivery with approximation guarantees.
8. Ioannidis, S., Chaintoutis, C., & Tassiulas, L. (2017). Distributed, Adaptive Algorithms for Joint Routing and Caching with Provable Guarantees. arXiv. — Introduces distributed 1−1/e-approximation algorithms for optimal cache routing.
9. Baştuğ, E., Bennis, M., & Debbah, M. (2014). Living on the Edge: The Role of Proactive Caching in 5G Wireless Networks. arXiv. — Pioneering work emphasizing proactive and edge caching for 5G systems.
10. Surveys on ICN / In-Network and Wireless/Small-Cell Caching (2014–2016). ACM Digital Library. — Multiple survey papers reviewing in-network caching frameworks, edge content distribution, and mobile caching challenges.
11. Tatarinov, I., Liu, L., & Others (Late 1990s). Static Caching of Web Servers / Server-Side Caching Studies. ACM Digital Library / Astrophysics Data System. — Early foundational studies analyzing static caching and content placement for web servers.
12. Dehghan, M., Seetharam, A., He, T., Salonidis, T., Kurose, J., & Towsley, D. (2014). Optimal Caching and Routing in Hybrid Networks. arXiv / MILCOM Proceedings. — Joint optimization of caching and routing strategies for hybrid MANET–cellular environments.
13. QoS & Policy-Management Surveys / Whitepapers (2010–2018). Scribd. — Surveys and technical whitepapers covering QoS frameworks, policy enforcement, and QoE validation in mobile data networks.
14. B. Zolfaghari, G. Srivastava, S. Roy, H. R. Nemati, F. Afghah, T. Koshiba, A. Razi, K. Bibak, P. Mitra and B. K. Rai, "Content Delivery Networks: State of the Art, Trends, and Future Roadmap," ACM Computing Surveys, vol. 53, no. 2, article 3380613, Jun. 2020.
15. S. Ioannidis and E. Yeh, "Jointly Optimal Routing and Caching for Arbitrary Network Topologies," IEEE Journal on Selected Areas in Communications, vol. 36, no. 6, pp. 1258–1275, Jun. 2018, doi: 10.1109/JSAC.2018.2844958.