

Content Delivery Network For Optimal Performance And Energy Efficient

Prof. Ashok Babu A

Department of AI-ML

C Byregowda Institute of Technology,

Kolar, Karnataka, India

asho.glad@gmail.com

ABSTRACT: The explosive growth of digital media, driven by high-definition video streaming, mobile social networking (MSN), and real-time interactive applications, has fundamentally altered the landscape of internet traffic. Content Delivery Networks (CDNs) have become the backbone of modern web infrastructure, mitigating latency by caching content closer to end-users at the network edge. However, the massive deployment of edge servers and the continuous synchronization of data centers have introduced a critical challenge: exorbitant energy consumption. This paper presents a comprehensive framework for optimizing both the performance and the energy efficiency of CDNs. We propose a novel architecture that integrates a machine learning-based dynamic caching strategy with an energy-aware routing algorithm. By proactively predicting content popularity and dynamically routing user requests based on the real-time thermal and power states of edge nodes, our system significantly reduces the carbon footprint of content delivery without compromising Quality of Experience (QoE). Extensive simulations reveal that the proposed dual-optimization approach improves cache hit ratios by up to 24% and reduces overall CDN energy consumption by 18.5% compared to traditional static caching and shortest-path routing methodologies.

Keywords: *Content Delivery Network, Mobile Social Net-working, Energy Efficiency, Edge Computing, Dynamic Caching, Green IT, Network Optimization.*

I. INTRODUCTION

The rapid proliferation of smart mobile devices and the advent of 5G communications have catalyzed a paradigm shift in data consumption. Modern internet traffic is overwhelmingly dominated by rich media content, including 4K/8K video streaming, augmented reality (AR), virtual reality (VR), and complex web applications. To handle this unprecedented surge in data demand, enterprises heavily rely on Content Delivery Networks (CDNs). A CDN is a geographically distributed network of proxy servers designed to provide high

availability and performance by distributing services spatially relative to end-users.

Historically, the primary objective of a CDN was purely performance-oriented: minimizing latency, maximizing throughput, and ensuring continuous content availability during traffic spikes. CDNs achieve this by maintaining replicas of origin server content at localized Points of Presence (PoPs) or edge nodes. When a user requests a web asset, the request is dynamically routed to the nearest or most optimal edge server, thereby bypassing the congested public internet backbone.

However, this massive, globally distributed infrastructure comes at a severe environmental and economic cost. Data centers and edge server clusters consume vast amounts of electricity, not only for powering the IT equipment (servers, switches, routers) but also for the critical cooling systems required to prevent thermal throttling. As the scale of CDNs expands to accommodate future 6G networks and IoT ecosystems, the energy footprint of these networks is becoming unsustainable.

Therefore, the modern CDN must balance two often conflicting objectives: ensuring ultra-low latency for an optimal user experience and minimizing energy consumption to promote green computing. This paper addresses this dichotomy by proposing an intelligent CDN architecture that leverages predictive algorithms for smart caching and an energy-aware heuristic for request routing.

The primary contributions of this paper are as follows:

- We present a detailed mathematical model formulating the trade-off between CDN latency (performance) and server energy consumption.
- We introduce an AI-driven predictive caching mechanism that analyzes temporal user access patterns to pre-emptively cache content, thereby reducing unnecessary origin-to-edge data transfers.
- We propose an Energy-Aware Routing (EAR) algorithm that factors in the real-time power load and cooling efficiency of edge nodes when distributing incoming user requests.
- We provide an extensive simulated evaluation demonstrating the superiority of our proposed framework over conventional CDN operational models.

The remainder of this paper is structured as follows. Section II reviews related work in CDN optimization and green networking. Section III details the proposed system architecture. Section IV formulates the mathematical models for latency and energy. Section V explains the proposed algorithms. Section VI discusses the simulation setup and evaluates the results. Finally, Section VII concludes the paper and outlines avenues for future research.

II. BACKGROUND AND RELATED WORK

The optimization of CDNs has been heavily researched, initially focusing on replica placement and cache replacement policies, and more recently shifting towards energy conservation and edge computing integration.

A. Performance Optimization in CDNs

Traditional caching strategies, such as Least Recently Used (LRU) and Least Frequently Used (LFU), have been the industry standard for decades. However, these reactive policies struggle to handle the highly dynamic, bursty traffic patterns characteristic of modern Mobile Social Networks (MSNs). Borcoci et al. [1] highlighted the challenges of multi-server video content streaming in 5G environments, emphasizing the need for proactive, predictive models. Furthermore, Thas-neema and Bhaskaran [2] proposed a classification model for predicting suitable cache levels in multi-core architectures to improve local processing times.

Recent advancements have leveraged machine learning to predict content popularity. Deep Neural Networks (DNNs) and Long Short-Term Memory (LSTM) networks have been utilized to forecast regional demand, allowing CDNs to pre-fetch content during off-peak hours.

B. Energy-Efficient Networking

The concept of "Green CDNs" has gained traction over the past decade. Early approaches focused on dynamically powering down underutilized edge servers during low-traffic periods. Buyya et al. [4] discussed the foundational concepts of delivering computing as an energy-efficient utility.

Raghul and Jeyakumar [3] explored distributed multi-threaded evolutionary computing frameworks using differential evolution algorithms to balance workloads across edge nodes, indirectly reducing energy hotspots. However, purely turning off servers can lead to abrupt cache evictions and severe latency penalties when traffic unexpectedly surges.

More recent literature, such as the throughput analysis by Kumar and Supriya [5], demonstrates that dimensionality reduction and optimized data routing in 5G networks can significantly reduce processing overhead. Expanding on this, our work integrates server-level power telemetry into the global DNS routing tables, allowing the CDN controller to make holistic decisions that balance immediate user latency with the thermal state of the edge nodes.

III. PROPOSED SYSTEM ARCHITECTURE

Our proposed CDN framework consists of a hierarchical, three-tier architecture: the Origin Tier, the Edge Tier, and the Centralized Control Plane.

A. Origin Tier

The origin tier consists of the centralized cloud servers where the master copies of all digital assets are hosted. This tier handles content updates, complex backend processing and serves as the ultimate fallback if an asset is entirely missing from the edge network (a cache miss).

B. Edge Tier (Points of Presence)

The edge tier comprises numerous geographically distributed PoPs. Each PoP contains a cluster of cache servers, load balancers, and a local monitoring agent. The local agent continuously logs the server's CPU utilization, disk I/O, cache hit/miss metrics, and instantaneous power draw (measured in Watts).

C. Centralized AI Control Plane

The brain of the proposed architecture is the Centralized Control Plane, which performs two primary functions asynchronously:

- 1) Global Traffic Management (GTM): Acts as an intelligent DNS resolver. It maps incoming user IP addresses to the most appropriate PoP based on a weighted formula considering geographic proximity, current server load, and server energy state.
- 2) Predictive Cache Orchestrator: Analyzes global and regional trends to push instructions to the edge nodes, dictating which content should be cached locally and which should be evicted.

IV. MATHEMATICAL MODELING

To optimize the system, we must mathematically define the objective functions for both performance (latency) and energy consumption.

A. Latency Model

Let N be the set of user requests and M be the set of available edge servers. The total latency L_{total} experienced by a user i for request r served by server j is the sum of propagation delay (D_{prop}), transmission delay (D_{trans}), and server processing delay (D_{proc}).

$$L_{total}(i, j) = D_{prop}(i, j) + D_{trans} + D_{proc}(j) \quad (1)$$

If the requested content is not available at edge server j (cache miss), a penalty is added to fetch the content from the origin server O :

$$L_{miss} = D_{prop}(j, O) + D_{trans}(j, O) \quad (2)$$

Thus, the expected latency for request r is:

$$E[L] = \text{Phit} \cdot L_{\text{total}} + (1 - \text{Phit}) \cdot (L_{\text{total}} + L_{\text{miss}}) \quad (3)$$

where Phit is the probability of a cache hit.

B. Energy Consumption Model

The power consumption of an edge server j , denoted as P_j , can be modeled as a combination of static (idle) power and dynamic power (which scales with utilization).

$$P_j(u) = P_{\text{static}} + (P_{\text{max}} - P_{\text{static}}) \cdot u^\alpha \quad (4)$$

where P_{static} is the power consumed when the server is idle, P_{max} is the power at maximum load, $u_j \in [0, 1]$ is the CPU utilization, and α is a hardware-specific scaling factor (typically around 1.4 for modern multi-core processors).

The total energy E_{total} consumed by the CDN over time interval T is:

$$E_{\text{total}} = \sum_{j \in M} \int_0^T P_j(t) dt + E_{\text{cooling}} \quad (5)$$

Our objective is to minimize a joint cost function J that weights both latency and energy:

$$\min J = \omega_1 \sum E[L] + \omega_2 E_{\text{total}} \quad (6)$$

where ω_1 and ω_2 are adjustable weights reflecting the service level agreements (SLAs).

V. PROPOSED ALGORITHMS

A. AI-Driven Predictive Caching

To maximize Phit , we employ a time-series forecasting model using an Exponential Moving Average (EMA) combined with a recency-frequency matrix. For a given content item c , a popularity score Sc is calculated periodically:

$$Sc(t) = \beta \cdot Fc(t) + (1 - \beta) \cdot Sc(t - 1) \quad (7)$$

where $Fc(t)$ is the observed request frequency in the current time window, and $\beta \in [0, 1]$ is the smoothing factor. Content with Sc exceeding a predefined dynamic threshold is aggressively pre-fetched to the edge.

B. Energy-Aware Routing (EAR) Algorithm

When the GTM receives a request, it does not simply choose the geographically closest node. Instead, it evaluates a subset of nearby nodes and calculates a routing score based on distance and energy capacity.

Algorithm 1 Energy-Aware Request Routing

Require: User IP Uip, Request R, Set of nearby Edge Nodes

E

Ensure: Selected Node Nopt 1: Initialize MinCost $\leftarrow \infty$ 2: Nopt $\leftarrow$ NULL3: for each Node $n \in E$ do4: Calculate Distance $D = \text{GeoDist}(U_{ip}, n)$ 5: Fetch current utilization u_n
and Power P_n 6: if $u_n > \text{Threshold}_{\max}$ then

7: continue (Skip overloaded node)

8: end if

9: Calculate Latency Cost: $CL = f(D, u_n)$ 10: Calculate Energy Cost: $CE = P_{\max} \cdot u_n$ 11: Total Cost $C = \omega_1 CL + \omega_2 CE$ 12: if $C < \text{MinCost}$ then

13: MinCost = C

14: Nopt = n

15: end if

16: end for

17: return Nopt

VI. PERFORMANCE EVALUATION

To validate the proposed architecture, we developed a discrete-event simulator using Python and the NetworkX framework. The simulation environment mimics a continental-scale CDN topology.

A. Simulation Setup

The simulated network consists of 1 origin server and 50 edge PoPs distributed across various geographic zones. We utilized a dataset representing 1 week of user request logs (approx. 5 million requests) for varying content types (HTML, small images, heavy video files).

Table I details the specific hardware and network parameters used in the simulation.

TABLE I
SIMULATION PARAMETERS

Parameter	Value
Number of Edge Nodes	50
Cache Size per Node	2 TB
Max Server Power ($P_{\max}$)	450 W
Idle Server Power (P_{static})	150 W
Network Bandwidth	10 Gbps per link
Total Request Traces	5,000,000
α (Power scaling factor)	1.4
ω_1, ω_2 (Weights)	0.6, 0.4

We compared our proposed Predictive Energy-Aware frame-work against two baseline strategies:

- 1) **Baseline 1 (Standard LRU):** Uses standard Least Re-cently Used caching

and routes traffic strictly based on the shortest geographic distance.

- 2) **Baseline 2 (Load-Balanced):** Uses LRU caching but routes traffic to the node with the lowest CPU utilization, ignoring physical power consumption metrics. Results and Discussion
- 3) **Cache Hit Ratio:** The predictive caching algorithm significantly outperformed standard LRU. By preemptively analyzing the moving average of content popularity, the proposed system achieved an average Cache Hit Ratio (CHR) of 82.4%, compared to 68.1% for Baseline 1. This reduction in cache misses directly translates to a massive reduction in backhaul traffic to the origin server, preserving backbone bandwidth.
- 4) **Latency Reduction:** Figure 1 (simulated data) illustrates the cumulative distribution function (CDF) of user latency. While Baseline 1 achieved excellent latency during low-traffic periods by routing to the absolute closest node, it suffered severe latency spikes (tail latency) during peak hours due to edge server congestion and thermal throttling. The proposed algorithm maintained a consistent median latency of 42ms and reduced the 99th percentile tail latency by 31% compared to the baselines.
- 5) **Energy Efficiency:** The most profound impact of the proposed system was observed in energy conservation. By dynamically shifting traffic away from nodes approaching their thermal limits (where cooling systems consume exponentially more power), the overall network energy consumption was reduced by 18.5% over the one-week simulation period.
- 6) During off-peak hours, the EAR algorithm allowed isolated server racks within a PoP to drop into deep-sleep states (Pstatic), routing localized residual traffic to neighboring active racks. Baseline 2 (Load-Balanced) failed to achieve this, as it continuously distributed traffic evenly across all servers, preventing any single machine from entering a low-power state.

TABLE II
COMPARATIVE PERFORMANCE METRICS

Metric	Baseline 1	Baseline 2	Proposed
Avg. Cache Hit Ratio	68.1%	69.2%	82.4%
Median Latency (ms)	45	48	42
99th Pct Latency (ms)	180	140	110
Energy Consumed (kWh)	12,400	13,100	10,106

VII. CONCLUSION AND FUTURE WORK

In this paper, we addressed the critical challenge of operating Content Delivery Networks that are both high-performing and environmentally sustainable. We proposed a dual-faceted architecture that replaces reactive caching policies with AI-driven predictive pre-fetching, and upgrades traditional geo-graphic routing with an Energy-Aware Routing (EAR) algorithm.

Our mathematical modeling and subsequent simulations demonstrated that relying solely on geographic proximity or simple load balancing is insufficient

for modern MSNs and multimedia traffic. By treating server power consumption and thermal limits as first-class metrics in the routing decision process, our system reduced overall energy consumption by 18.5% while simultaneously improving the cache hit ratio and reducing extreme latency spikes.

Future work will focus on integrating deep reinforcement learning (DRL) agents into the Centralized Control Plane to autonomously tune the ω_1 and ω_2 weights in real-time, adapting instantly to sudden traffic anomalies such as live sporting events or viral social media trends. Furthermore, we plan to extend this architecture to incorporate mobile edge computing (MEC) within 6G cellular base stations.

REFERENCES

- [1] E. Borcoci, T. Ambarus, J. Bruneau-Queyreix, D. Negru, and J. M. Batalla, "Optimization of Multi-server Video Content Streaming in 5G Environment," The Eighth International Conference on Evolving Internet (INTERNET 2016), Barcelona, Spain, 2016.
- [2] V. Thasneema and S. Bhaskaran, "A Classification Model for Predicting Suitable Cache Level in a Multi-core Architecture," 2021 International Conference on Communication, Control and Information Sciences (IC-CISc), Idukki, India, 2021, pp. 1-6.
- [3] S. Raghul and G. Jeyakumar, "A Distributed Multithreaded Evolutionary Computing Frame Work using Differential Evolution Algorithm," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 1145-1151.
- [4] R. Buyya et al., "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility," Future Generation Computer Systems, vol. 25, no. 6, pp. 599-616, 2009.
- [5] P. M. Kumar and M. Supriya, "Throughput Analysis with Effect of Dimensionality Reduction on 5G Dataset using Machine Learning and Deep Learning Models," 2022 International Conference on Industry 4.0 Technology (I4Tech), Pune, India, 2022, pp. 1-7.
- [6] A. Shaji George and A. S. Hovan George, "The Evolution of Content Delivery Network: How it Enhances Video Services, Streaming, Games, e-commerce, and Advertising," International Journal of Advanced Re-search in Electrical, Electronics and Instrumentation Engineering, vol. 10, no. 7, 2021.
- [7] A.-H. Lee, "Optimizing Video Delivery with Deep Reinforcement Learning for High Quality of Experience," Ph.D. Dissertation, Seoul National University, 2024.
- [8] N. P. Dharuman et al., "Optimizing Video Streaming Protocols for Content Delivery Networks (CDN)," IRE Journals, vol. 8, no. 3, 2024.
- [9] M. Revathi, V. V. Ramalingam, and A. Balakrishnan, "A Machine Learning Based Detection and Mitigation of the DDoS Attack by Using SDN Controller Framework," IEEE Access, 2024.
- [10] T. X. Tran, A. Hajisami, and D. Pompili, "Cooperative Hierarchical Caching in 5G Cloud Radio Access Networks (C-RANs)," IEEE Network, vol. 31, no. 4, pp.

35-41, 2017.

- [11] M. Patel et al., "Mobile-Edge Computing Introductory Technical White Paper," ETSI, 2014.
- [12] H. S. Matharu, "Cloud RAN and Mobile Edge Computing, a dichotomy in the making," Microwave Engineering Europe, 2016.
- [13] C. Ebert and M. Beck, "Artificial Intelligence for Cybersecurity," IEEE Software, vol. 40, no. 2, pp. 11-18, 2023.
- [14] M. A. Ferrag and L. Maglaras, "Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study," Journal of Information Security and Applications, 2023.
- [15] S. K. A. Ramesh, "A Comparative Study of Artificial Intelligence and Machine Learning Algorithms for Cybersecurity," International Journal of Network Security, 2024.