An Intelligent Scheduling Optimization Algorithm for Multimodal Cache Resources with Status Awareness of Metropolitan Area Network CDN Nodes

Ruirong Jiang*, Zhibiao Xiong, Junliang Wu, Jinyong Xu
School of Artificial Intelligence, Nanchang JiaoTong Institute, Nanchang, China

***Abstract*—To address the resource scheduling challenges faced by metropolitan area network content delivery networks (CDN) when carrying multimodal traffic streams such as high-definition video, virtual reality (VR), and augmented reality (AR), this study proposes an intelligent optimization algorithm for multimodal cache resource scheduling that is CDN Node State Awareness. First, to address the dynamic nature of network topology and the heterogeneity of service streams, we construct a directed graph-based metropolitan area network CDN model. This model enables real-time perception of multi-dimensional states of nodes, including CPU utilization, memory usage, remaining bandwidth, and cache occupancy. We also introduce a mechanism for quantifying the transmission demand weight and cache value of multimodal content, providing a foundational support for differentiated scheduling. Second, at the optimization enhancement layer, we design a transmission path selection strategy, a cache replacement mechanism that integrates content value and access popularity, and an adaptive scheduling structure based on node load balancing. Furthermore, a Deep Q-Network is introduced at the cloud computing decision layer. Node states and user request features are modeled as a state space, while cache placement and request allocation strategies are modeled as an action space. A multi-objective reward function integrating hit rate, response latency, and packet loss rate is designed to achieve dynamic and intelligent scheduling of multimodal cache resources. Integrating path selection, cache updates, and fault recovery mechanisms to construct an overall optimization model enhances the system's adaptive scheduling capability in complex business services. The experiment shows that the algorithm has significant advantages in multi node collaborative scheduling: within 1-8 seconds, the transmission rate of node B reaches 30Mbps and the resource utilization rate of node A is improved; The resource download time remains stable at 4.4-4.9 seconds during 24-hour operation; In the large-scale scenario of 500 user requests, cross node cache load adaptive balancing, system overhead linearly increases, and data transmission security rate reaches 99.45%, creating an efficient, reliable, and scalable intelligent scheduling system for multi-mode content distribution in metropolitan area networks.**

Keywords*—Metropolitan area network; CDN node state awareness; multimodal; cached resources; intelligent scheduling optimization; deep Q-Network; reward function

I. Introduction

In CDN nodes of metropolitan area networks, multimodal caching resources are a collection of heterogeneous resources such as storage, computing, and network bandwidth, which can collaboratively process user requests and service content, improving resource utilization and service experience [1]. Multi-modal cache resource intelligent scheduling optimization, coordinated allocation, and dynamic adjustment of heterogeneous resources, achieve the best match between requests and resources, overcome the limitations of single resource scheduling, reduce service latency globally, and improve node throughput, which is the core technology to ensure the service capacity and cost-effectiveness of metropolitan area networks.

Traditional CDN caching scheduling relies heavily on single dimensions such as content popularity, without fully considering multidimensional dynamic information such as real-time node load, resource utilization, and network status, resulting in scheduling decisions lagging behind environmental changes [2]. CDN node state awareness can collect and analyze multimodal data such as node load and link quality in real-time, providing an accurate and timely decision-making basis for intelligent scheduling, enabling the scheduling system to dynamically adapt to network and service changes. Based on this, we propose a state-aware metropolitan area network CDN node multi-mode cache resource intelligent scheduling optimization algorithm, which achieves precise on-demand resource allocation and global collaborative optimization, improves resource utilization elasticity, and service response reliability.

II. Related Work

To ensure network connectivity [3], many scholars have researched resource scheduling methods. Perotin et al. [4] proposed a multi-resource scheduling algorithm (MRSA) for scalable parallel job computing workflows, which optimizes job execution sequences by dynamically configuring job resource sets and expanding lists. However, this method has limitations in dealing with resource contention and system dynamics. If the scheduling strategy does not fully consider the real-time changes in node load, it will result in unreasonable resource allocation, affect the minimization of completion time goals, and be difficult to adapt to the burst and heterogeneous characteristics of multi-mode traffic in metropolitan area network CDN. Kloda et al. [5] formulated the cache partitioning problem as an integer quadratic constrained programming model, designing heuristic algorithms and search strategies for preemptive and non-preemptive scheduling, respectively. However, their heuristic algorithm lacks the necessary dimensions when coordinating priority-based scheduling with cache allocation. If node state information is

*Corresponding author.

not effectively integrated, it may weaken the cache minimization effect and fail to meet the dynamic adaptation requirements of multimodal content for multi-dimensional resources such as node CPU and bandwidth. Ahani and Yuan [6] proposed a decomposition algorithm based on Lagrangian relaxation, which combines content recommendation values with data timeliness. However, the scheduling performance depends on the accuracy of the content recommendation model. If the model lacks real-time node cache, link quality, and other information, it will result in poor cache update decisions and difficulty in ensuring service quality for low-latency applications such as VR/AR. Hatami et al. [7] constructed a Markov decision model for resource-constrained IoT caching and designed a relaxed truncation heuristic algorithm to minimize the information age. However, the deviation between the sensor energy harvesting model and reality can affect the optimality of the strategy. In a metropolitan area network CDN, the parameters related to multi-mode service flow dynamically change with node status, and single-dimensional modeling is difficult to accurately schedule. Compared with the above work, this study innovatively combines the multidimensional state perception of CDN nodes in metropolitan area networks with deep Q-networks, and integrates multiple mechanisms such as content value, access popularity, path overhead, and load balancing to form an end-to-end intelligent scheduling framework, filling the gap in collaborative scheduling in multimodal heterogeneous business scenarios.

III. Intelligent Scheduling Optimization of Multimodal Cache Resources

A. Metropolitan Area Network CDN Model and State Awareness

By capturing multimodal content—including high-definition video, VR/AR, and big data files—from the metropolitan area network CDN model, we establish a multimodal cache resource intelligent scheduling optimization model. This model comprises a metropolitan area network CDN network modeling layer, an optimization enhancement layer, and a cloud computing decision-making layer, as illustrated in Fig. 1.

As shown in the architecture of the multimodal cache resource intelligent scheduling optimization model in Fig. 1, the modeling layer constructs a directed graph of the metropolitan area network CDN. It implements state awareness for metropolitan area network CDN nodes by monitoring real-time node status, such as CPU, memory, and bandwidth, while quantifying the transmission demand weights and caching value of multimodal content, including HD video, VR/AR, and big data files. In the optimization and enhancement layer, path optimization selects the service path with the lowest transmission overhead; cache replacement is performed based on content value and access popularity; and load balancing and fault recovery are designed by monitoring node load balancing. In the cloud computing decision-making layer, the system collaborates with the cloud computing platform to dynamically allocate cache resources, making real-time decisions based on node state awareness using a Deep Q-Network, thereby achieving intelligent scheduling optimization of multimodal cache resources.

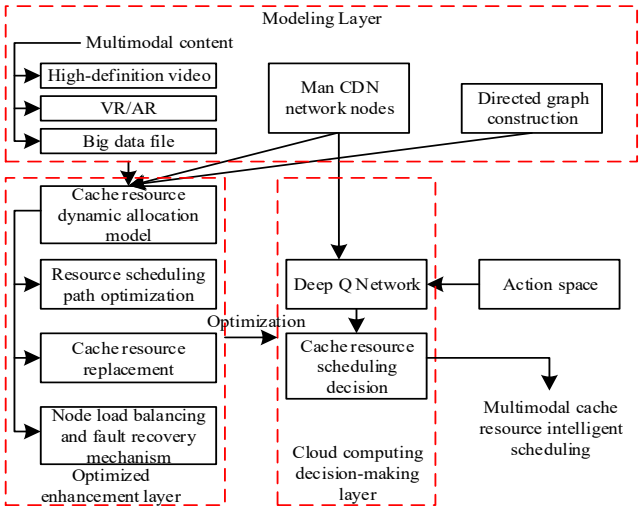

Fig. 1. Architecture of the multimodal cache resource intelligent scheduling optimization model.

In the context of a multi-modal metropolitan area network content delivery network (CDN), the network [8] can be modeled as a directed graph $G=(V,E)$, where $V=\{v_1,v_2,\cdots,v_N\}$ represents the set of nodes in the multi-modal metropolitan area network content delivery network (CDN), N denotes the total number of nodes, $E=\{e_1,e_2,\cdots,e_M\}$ denotes the set of links between nodes, M denotes the total number of links. The spatial model of the multi-modal metropolitan area network content delivery network is shown in Fig. 2.

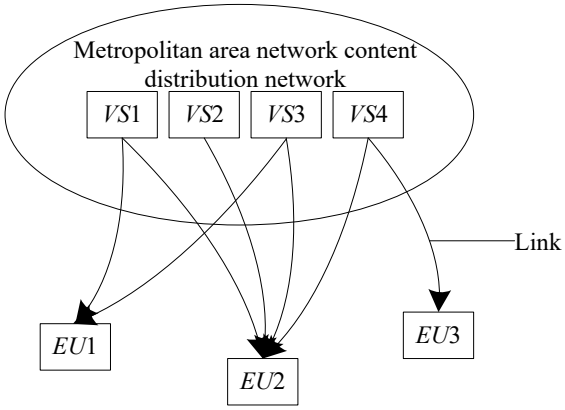

Fig. 2. Multi-modal metropolitan area network content distribution network space model.

In the multi-modal metropolitan area network content distribution network space model shown in Fig. 2, VS and EU represent the set of service nodes and the set of end users in the multi-modal metropolitan area network content distribution network, respectively, and $V=VS\bigcup EU$.

During caching and scheduling [9-11], each link $e \in E$ between nodes in the multimodal metropolitan area network, the content distribution network has a bandwidth capacity η_e and a transmission delay τ_e. The multimodal content set is represented as $C = \{c_1, c_2, \cdots, c_K\}$, where K denotes the types of multimodal content. Each type of content c_k has a modal feature vector $d_k = [Size_k, Rate_k, Deadline_k, Priority_k]$, and $Size_k$, $Rate_k$, $Deadline_k$, and $Priority_k$ represent the size, transmission rate requirement, deadline, and priority of the k the multimodal content, respectively. The state vector of the node v_n at time t is expressed as:

$$S_n(t) = [CPU_n(t), Mem_n(t), BW_n(t), Cache_n(t), Load_n(t)] \quad (1)$$

In particular, $CPU_n(t)$ represents the CPU utilization of the multimodal metropolitan area content delivery network at time t, $Mem_n(t)$ represents the memory usage of the multimodal metropolitan area content delivery network at time t, $BW_n(t)$ represents the remaining bandwidth of the multimodal metropolitan area content delivery network at time t, $Cache_n(t)$ represents the cache space usage of the multimodal metropolitan area content delivery network at time t, and $Load_n(t)$ represents the current load of the multimodal metropolitan area content delivery network at time t.

B. Modeling Multimodal Content Characteristics and Dynamic Allocation of Cache Resources in Metropolitan Area Network CND

Based on the multimodal content — including high-definition video, VR/AR, and big data files—obtained from the metropolitan area network CDN model shown in Fig. 1, a quantitative model is established to map modal features into scheduling decision criteria. The transmission demand weights ω_n for the multimodal content c_k in the metropolitan area network, CDN is calculated and expressed as:

$$\omega_n = \frac{\alpha_1 Size_n}{\max(Size)} + \frac{\alpha_2 Rate_n}{\max(Rate)} + \frac{\alpha_3 Priority_n}{\max(Priority)} + \frac{\alpha_4}{(Deadline_n + 1)} \quad (2)$$

where, α_1, α_2, α_3, and α_4 all represent weight coefficients, and $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$. $\max(\ )$ denotes the maximum value of this feature across all multimodal content in the metropolitan CDN network.

To ensure the cache scheduling effect [12-13], the cache value V_{kn} of multimodal content c_k in the metropolitan area network, the CDN network at the content delivery network node v_n is further calculated as:

$$V_{kn} = \chi_1 \xi_{kn} + \chi_2 / (\lambda_{kn} + 1) + \chi_3 B_n / Size_n \quad (3)$$

where, ξ_{kn} represents the proportion of requests for multimodal content c_k that are successfully served at the metropolitan area network CDN node v_n; λ_{kn} represents the number of network hops from the user's request to the metropolitan area network CDN node; B_n represents the network bandwidth of the metropolitan area network CDN node v_n; and χ_1, χ_2, and χ_3 all represent value coefficients.

Let the total cache capacity of the metropolitan area network content delivery network node v_n be Q_n, and the used cache is $Q_n^{used}(t)$. Then, the remaining cache is calculated as:

$$Q_n^{res}(t) = Q_n - Q_n^{used}(t) \quad (4)$$

Introducing the cache allocation decision variable κ_{kn}, when $\kappa_{kn} = 1$, it indicates that the multimodal content c_k is cached at the metropolitan area network content delivery network node; when $\kappa_{kn} = 0$, it indicates that the multimodal content c_k is not cached at the metropolitan area network content delivery network node. Based on the cache value V_{kn} calculated from Eq. (3), the objective of allocating multimodal cache resources is set to maximize the overall cache value, expressed as:

$$F = \max \sum_{k=1}^{K} \sum_{n=1}^{N} V_{kn} \cdot \kappa_{kn} \quad (5)$$

To ensure the validity of the multimodal cache resource allocation objective constructed in Eq. (5), constraint conditions are further established to constrain Eq. (5).

Considering that the capacity of actual caching resources in metropolitan area network content delivery networks is limited, a cache capacity constraint [14] is formulated, expressed as:

$$\sum_{k=1}^{K} Size_n \cdot \kappa_{kn} \le Q_n^{\max} \ , \ \forall n \quad (6)$$

where, $Q_n^{\max}$ represents the maximum capacity of the metropolitan area network content delivery network cache resources.

We formulate a multimodal content uniqueness constraint for the metropolitan area network content delivery network, expressed as:

$$\sum_{n=1}^{N} \kappa_{kn} \le 1 \ , \ \forall k \tag{7}$$

We formulate the actual metropolitan area network content delivery network node load constraint, expressed as:

$$Load_n(t) + \sum_{k=1}^{K} \mu_{kn} \cdot \kappa_{kn} \le Load_n^{\max} \ , \ \forall n \tag{8}$$

where, μ_{kn} represents the predicted request rate for the multimodal content c_k at the metropolitan area network content delivery network node v_n, and $Load_n^{\max}$ represents the maximum load capacity of the metropolitan area network content delivery network node.

Based on the construction of the multimodal cache resource allocation objectives and constraints in Eq. (5) - (8) described above, the optimal cache resource allocation value V_{best} is ultimately obtained.

C. Intelligent Scheduling of Multimodal Cache Resources Based on Deep Q-Networks

Based on the node state awareness results $S_n(t)$ obtained from the metropolitan area network content delivery network model in Section II-A, a Deep Q-Network is employed to make dynamic scheduling decisions for multimodal cache resources, thereby constructing the intelligent scheduling model for multimodal cache resources as shown in Fig. 3.

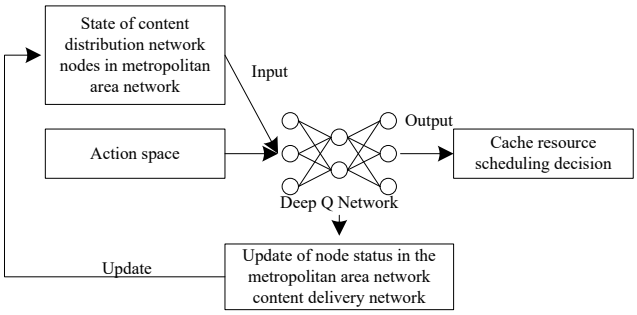

Fig. 3. Intelligent scheduling model of multimodal cache resources.

Define the state space of the metropolitan area content delivery network nodes at time t as $S(t)$, expressed as:

$$S(t) = \left[S_1(t), S_2(t), \cdots, S_n(t), \psi(t), R(t) \right] \tag{9}$$

where, $\psi(t)$ denotes the characteristics of the content queues awaiting scheduling in the metropolitan area content delivery network model at time t, and $R(t)$ denotes the user request distribution vector at time t.

Let $A(t)$ denote the action space in the Deep Q-Network at time t, i.e., the decision vector for multimodal cache resources scheduling, expressed as:

$$A(t) = \left[v_{tar,1}(t), v_{tar,2}(t), \cdots, v_{tar,n}(t), \cdots, v_{tar,N}(t) \right] \tag{10}$$

where, $v_{tar,n}(t)$ denotes the n th target node for multimodal cache resource scheduling.

Each multimodal cache resource scheduling action involves assigning the first N requests from the current request queue to the corresponding nodes [15]. The designed reward function is expressed as:

$$\vartheta(t) = \gamma_1 \sum hit / \sum req - \gamma_2 \overline{t_{del}} - \gamma_3 \overline{loss} + \gamma_4 \sum V_{best} / \max(V) \tag{11}$$

where, hit represents the number of requests for multimodal content that hit the target node in the metropolitan area network content delivery network, req represents the total number of requests, $\overline{t_{del}}$ represents the average response delay, $\overline{loss}$ represents the average packet loss rate, and γ_1, γ_2, γ_3, and γ_4 all represent reward weights.

Finally, intelligent scheduling of multimodal cache resources is implemented using a Deep Q-Network [16-17], and the output scheduling decision is expressed as:

$$\pi_t = Q(s_t; a_t) \tag{12}$$

where, $Q(s_t; a_t)$ represents the Q-value when the multimodal cache resources scheduling action a_t is taken under the metropolitan area network content delivery network node state s_t.

D. Optimization of Intelligent Scheduling for Multimodal Cache Resources

This study adopts Deep Q-Network instead of other deep reinforcement learning algorithms, such as PPO or DDPG, mainly for three reasons. First, the action space of this problem consists of discrete cache placement and request allocation decisions, which DQN handles naturally, whereas DDPG and PPO are better suited for continuous control tasks. Second, DQN has a simple structure and stable training, and its experience replay mechanism effectively breaks data correlation, making it appropriate for online scheduling scenarios. Third, compared with policy gradient methods, DQN achieves higher sample efficiency under the same sample size, which is particularly critical for CDN scheduling that requires real-time response. The training convergence of DQN has been verified in the experiments. As shown in Fig. 7 later in this study, the reward value stabilizes after approximately 2,000 steps, confirming the reliability of the model. The Deep Q-Network model constructed in Subsection C enables intelligent scheduling of multimodal cache resources. To further improve the effectiveness of this intelligent scheduling, the constructed multimodal cache resource intelligent scheduling model is optimized.

1) Optimization of multimodal cache resource scheduling paths

Let the set of transmission paths for a user request r_q from the source node v_s of the multimodal metropolitan area network content delivery network to the destination node v_{ob} be denoted as L_{tr}. The transmission overhead for each path $l \in L_{tr}$ is calculated as follows:

$$U(l) = \sum_{e \in l} \left(Size_q / U_e \right) + \sum_{v \in l} del_v \quad (13)$$

where, e denotes the user request path e, del_v denotes the processing delay of the multimodal metropolitan area network content delivery network node v, and the optimal path l_{best} for selecting the multimodal cache resource scheduling path is expressed as:

$$l_{best} = \arg\min_{l \in L_{tr}} U(l) \quad (14)$$

In the process of selecting the optimal path for multimodal cache resources scheduling, rationality constraints must be considered. The link bandwidth constraint is expressed as:

$$\sum_{q:e \in L_q} Rate_q \le U_e \ , \ \forall e \quad (15)$$

where, $Rate_q$ represents the transmission rate of the cached resource requested by the user, and U_e represents the transmission overhead of the cached resource for the user request path e.

2) Replacement and update strategies for multimodal cached resources

A cache replacement strategy is designed based on the content value and access popularity of multimodal cached resources. Let the replacement weight of the multimodal cached resource content c_k at the multimodal metropolitan area network content delivery network node v_n be ϖ_{kn}^{rep}, expressed as:

$$\varpi_{kn}^{rep} = \theta_1 / V_{kn} + \theta_2 t_{cache,kn} + \theta_3 \left(1 - \varphi_{kn}\right) \quad (16)$$

where, $t_{cache,kn}$ represents the cache duration of the multimodal cached resource content, φ_{kn} represents the recent access frequency, and θ_1, θ_2, and θ_3 represent the replacement weight coefficients, respectively.

When the cache of a multimodal cache resource is insufficient, the content of the resource with the highest replacement weight is replaced, thereby performing a replacement and update of the multimodal cache resource, expressed as:

$$c_{rep} = \arg\max \varpi_{kn}^{rep} \quad (17)$$

3) Load balancing and fault recovery mechanisms for multimodal metropolitan area network content delivery network nodes

Introducing the load balancing degree ρ_n of a multimodal metropolitan area network content delivery network node v_n, the calculation is expressed as:

$$\rho_n = 1 - \left| Load_n - \overline{Load} \right| / \overline{Load} \quad (18)$$

where, $\overline{Load}$ represents the average load of the entire multimodal metropolitan area network content delivery network. During intelligent scheduling of Multimodal Cache Resources, the system tends to select metropolitan area network content delivery network nodes with higher bandwidth ρ_n. During the cache activation process [18-20], when a metropolitan area network content delivery network node fails, a recovery mechanism is triggered to migrate the content from the failed node to a standby node v_{st}, expressed as:

$$v_{st} = \arg\max \left(BW_v cachefree_v \right) \quad (19)$$

where, $cachefree_v$ represents the remaining cache utilization.

The additional overhead incurred during the process of migrating content from the failed metropolitan area network content delivery network node to the standby node v_{st} is calculated as follows:

$$U_{mig} = \sum Size_k / \min\left(BW_{n \to st}, BW_{st} \right) \quad (20)$$

4) Integrated optimization of intelligent scheduling for multimodal cache resources

Integrating the aforementioned optimization techniques, this study formally defines the overall multimodal caching resource intelligent scheduling optimization problem as the following mathematical formulation:

Decision variables:

$\kappa_{kn} \in \{0,1\}$, indicating whether multimodal content is cached (allocated) to the node under the content type;

$\iota_{kn} \in \{0,1\}$, indicating whether the user request r_q selects the transmission path l;

$\nu_{kn} \in \{0,1\}$, indicating whether the user request r_q is assigned to the node v_n.

Objective function (minimizing total transmission and caching cost):

$$F_{total} = \min\left(\omega_1 \overline{t_{del}} + \omega_2 \overline{loss} - \omega_3 \xi_{kn} + \omega_4 U_{total}\right) \quad (21)$$

where, U_{total} represents the total transmission and caching overhead of the intelligent scheduling of multimodal cache resources, calculated as:

$$U_{total} = \sum_{r_q} \sum_{l} t_{r_q l} U\left(r_q\right) + \sum_{k,n} \kappa_{kn} U_{sto,kn} \quad (22)$$

where, $U_{sto,kn}$ represents storage overhead. By solving the formulated overall multimodal cache resource intelligent scheduling optimization objective, the optimal multimodal cache resource intelligent scheduling strategy is obtained.

The optimization problem is subject to the following constraints: cache capacity constraint as defined in Eq. (6), content uniqueness constraint as defined in Eq. (7), node load constraint as defined in Eq. (8), and link bandwidth constraint as defined in Eq. (15). Additionally, the consistency between path selection and node assignment requires that $Rate_q = 1$ equals 1 only when the endpoints of q include n and Eq. (8) evaluates to 1. This formulation constitutes an integer linear programming problem, which is solved approximately via a Deep Q-Network to meet real-time scheduling requirements.

IV. Experimental Analysis

To validate the effectiveness of the proposed algorithm in performing intelligent scheduling optimization of multimodal cache resources under state-aware conditions at metropolitan area network CDN nodes, a simulation environment was constructed for a specific real-world metropolitan area network. In this simulation, the proposed algorithm was applied to perform state awareness and resource scheduling for multiple township CDN nodes under the jurisdiction of this metropolitan area network. These nodes receive and process continuous video streams generated by high-definition cameras in townships, as well as interactive immersive data transmitted by drones equipped with AR/VR devices. The algorithm dynamically monitors the real-time status of cache, bandwidth, and computing load of each node, intelligently optimizes cache resource allocation, content placement, and distribution paths. The actual topology is shown in Fig. 4.

Fig. 4 shows the topology of intelligent cache resource scheduling optimization, based on which the specific model and parameters of the devices used were analyzed and determined. The results are shown in Table I.

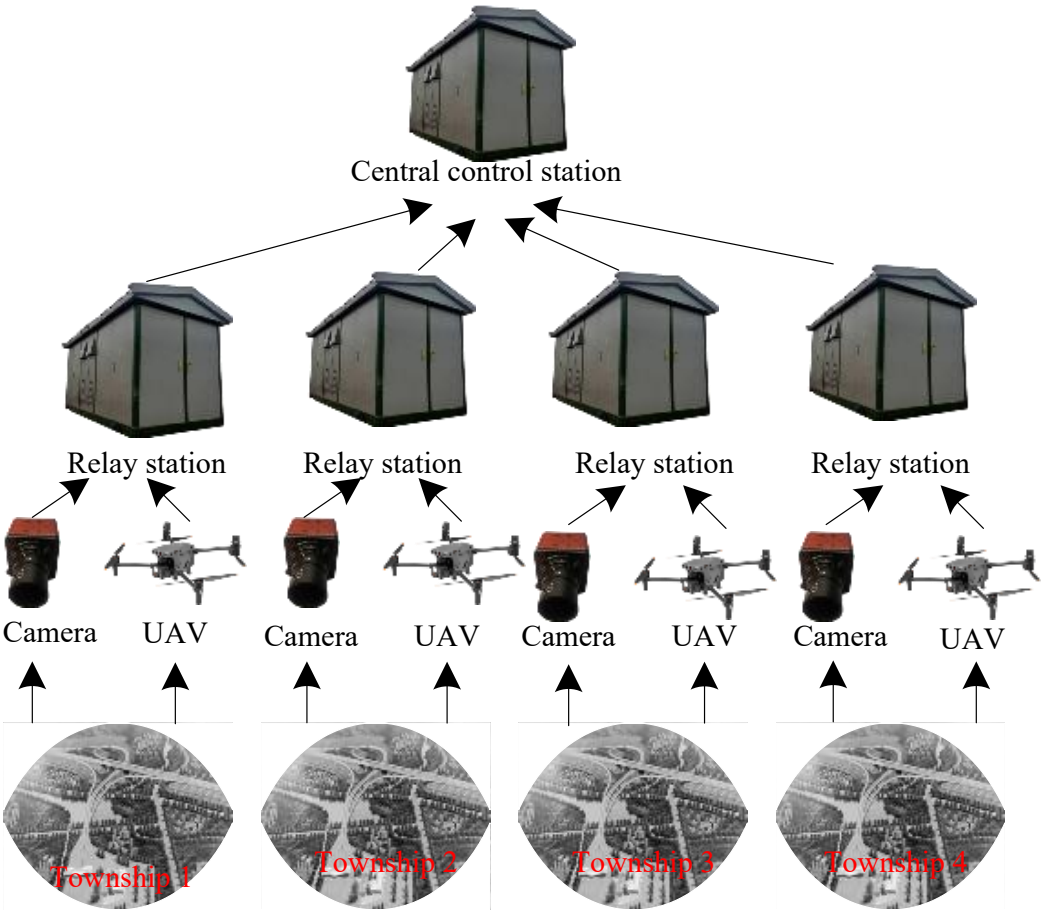

Fig. 4. Optimization topology of intelligent scheduling of cache resources.

Table I shows that high-definition cameras and AR/VR devices exhibit multi-mode characteristics of high bit rate and low latency in video streaming and immersive data, providing real data for algorithm differentiation of services and differentiated scheduling. The CDN node server has a cache capacity of 10TB, providing resources for multi-mode content collaborative placement and dynamic replacement, and supporting intelligent algorithm allocation optimization. The network switch has a 256Gbps backplane bandwidth and a 96Mpps packet forwarding rate, ensuring real-time and reliable data transmission between nodes, helping algorithms quickly respond to load changes and optimize paths. The simulation environment built with these devices fully meets the technical requirements of the algorithm's perception-decision-scheduling closed-loop in three dimensions—data generation, cache resources, and network transmission—ensuring that the algorithm can be successfully implemented in multimodal, highly dynamic scenarios based on real device parameters. In addition, the weight coefficients in the proposed algorithm were determined via grid search combined with empirical tuning, and kept fixed across all experiments to ensure result comparability. Specifically, the transmission demand weight coefficients were set to 0.25, 0.30, 0.25, and 0.20, respectively; the cache value coefficients were set to 0.4, 0.3, and 0.3, respectively; the reward function weights were set to 0.35, 0.30, 0.20, and 0.15, respectively; and the replacement weight coefficients were set to 0.4, 0.35, and 0.25, respectively. These parameter assignments cover the four categories of formulas, including the reward function, transmission demand weights, cache value, and replacement weights, and all experiments adopt this fixed set of values.

TABLE I. SPECIFIC MODELS AND PARAMETERS OF THE EQUIPMENT

Equipment	Parameters	Actual value
High-definition camera	Model	DS-2CD3T45G1-I5
	Resolution / Million Pixels	400
	Frame rate/fps	25
	Video encoding format	H.265/H.264
AR/VR Equipment	Model	Inspire 2+X7
	Video resolution	6016×3200
	Frame rate/fps	30
CDN node server	Model	FusionServer Pro 2288H V5
	Cache capacity/TB	10
Network switch	Model	S5720S-28P-LI-AC
	Backplane bandwidth/Gbps	256
	Support Agreement	HTTP, FTP, DNS
	Backplane bandwidth/Gbps	256
	Packet forwarding rate/Mpps	96

In a simulated metropolitan area network CDN comprising two nodes, A and B, the algorithm described in this study was used to perform intelligent scheduling optimization of multimodal cache resources. The resource transmission rates of nodes A and B in the metropolitan area network CDN were measured over a 0~8-second time window, and the per-unit overhead was calculated. The results are shown in Fig. 5.

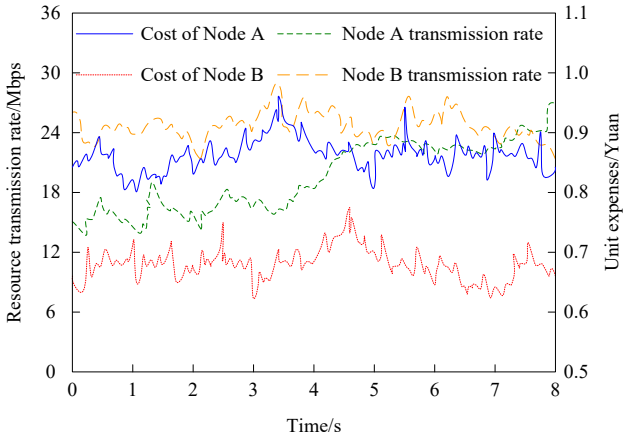

Fig. 5. Intelligent scheduling optimization results of cache resources.

Fig. 5 shows that at the initial time of 0 seconds, the transmission rate of Node A fluctuates within the range of approximately 12 Mbps to 27 Mbps, with unit cost fluctuating between approximately 0.78 and 1.0 yuan, while the transmission rate of Node B fluctuates within the range of approximately 18 Mbps to 30 Mbps, with unit cost fluctuating between approximately 0.6 and 0.8 yuan. Based on real-time state awareness, the algorithm intelligently schedules part of the load from Node A to Node B. Consequently, between 1 and 8 seconds, Node A's overall cost shows a downward trend, and the unit cost steadily increases to 0.96. Although node B has slightly increased the cost of undertaking additional tasks, its transmission rate has been optimized to 30Mbps. This indicates that the algorithm is not simply load sharing, but rather achieves joint optimization of system resource utilization and allocation performance through intelligent decision-making, while reducing the overall cost of key nodes and synergistically improving the transmission efficiency of two nodes.

To simulate the 24-hour operation of multi-mode cache resource intelligent scheduling in metropolitan area networks, the algorithm proposed in this study, the multi-resource scheduling algorithm [4], the Lagrangian relaxation-based decomposition algorithm [6], and MDP combined with relaxation-truncation heuristic algorithm [7] were used for scheduling, and the download time of each algorithm resource was calculated. The results are shown in Fig. 6.

Fig. 6 shows that the proposed algorithm exhibits excellent performance and stability in long-term scheduling. Within 24 hours, its resource download time remained stable in the low range of 4.4-4.9 seconds with minimal fluctuations. The overall time consumption of the comparative method is higher, mostly between 4.7-5.8 seconds, and there are significant peaks in the load algorithm at 18:00 and the network distance algorithm at 21:00. This indicates that single-dimensional algorithms are difficult to cope with the dynamic changes in multi-mode service loads in metropolitan area networks. The algorithm in this article integrates multi-dimensional state perception, such as node cache, bandwidth, and computing power, to achieve

accurate prediction and intelligent decision-making. It can dynamically optimize content placement and distribution paths, provide lower and more stable download latency in complex and time-varying environments, and effectively improve user experience and system efficiency.

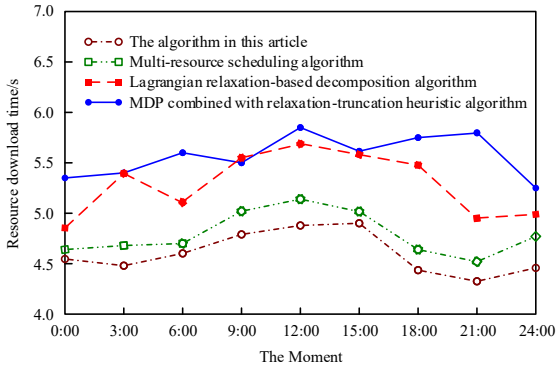

Fig. 6. Intelligent scheduling of multi-modal cache resources for optimizing resource download time.

From Fig. 7, it can be seen that the reward value rapidly increases in the first 500 steps, indicating that the agent has initially learned effective scheduling strategies. The reward fluctuation for 500-1800 steps decreases and steadily increases. After 2000 steps, the reward value converged to about 0.92, indicating that the model has stably converged. Throughout the entire training process, there was no occurrence of gradient explosion or overestimation of Q-values, which verifies the stability and applicability of DQN in this problem.

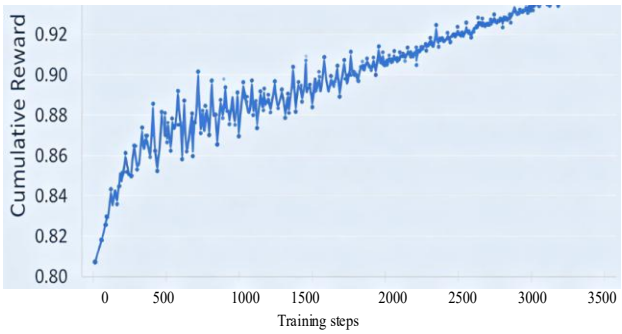

Fig. 7. Average reward convergence curve during DQN training process.

In simulation experiments, the number of user requests within the metropolitan area network was set to 100, 200, 300, 400, and 500. Three types of nodes, A, B, and C, were used for collaborative cache resource scheduling. The proposed algorithm was applied to optimize the intelligent scheduling of multimodal cache resources. The cache usage and optimized overhead of nodes A, B, and C were analyzed, and the effectiveness of the optimized intelligent cache resource scheduling was evaluated using data transmission security metrics. The results are shown in Table II. Data transmission security is defined as the proportion of successfully transmitted packets that have not been tampered with or lost to the total number of packets sent. This indicator comprehensively reflects the scheduling strategy's ability to ensure data integrity and confidentiality in path selection and congestion control.

TABLE II. OPTIMIZATION EFFECT OF INTELLIGENT SCHEDULING OF CACHE RESOURCES

The number of user requests	Indicators	Actual value
100	Node A cache/GB	320.56
	Node B cache/GB	330.78
	Node C cache/GB	348.66
	Total Expenses / Yuan	125.34
	Data transmission security/%	95.23
200	Node A cache/GB	310.23
	Node B cache/GB	345.67
	Node C cache/GB	344.1
	Total Expenses / Yuan	230.67
	Data transmission security/%	93.56
300	Node A cache/GB	305.89
	Node B cache/GB	350.32
	Node C cache/GB	343.79
	Total Expenses / Yuan	345.89
	Data transmission security/%	92.12
400	Node A cache/GB	300.45
	Node B cache/GB	355.67
	Node C cache/GB	343.88
	Total Expenses / Yuan	460.12
	Data transmission security/%	98.78
500	Node A cache/GB	295.78
	Node B cache/GB	360.45
	Node C cache/GB	343.77
	Total Expenses / Yuan	575.34
	Data transmission security/%	99.45

Table II shows that as the number of user requests increases from 100 to 500, the algorithm dynamically adjusts the cache load between nodes based on state awareness: Node A's cache load decreases from 320.56GB to 295.78GB, Node B increases from 330.78GB to 360.45GB, and Node C remains relatively stable. This indicates that the algorithm can intelligently redistribute loads based on real-time conditions, prevent single-point overload, and achieve multi-node adaptive load balancing. The total cost increases approximately linearly with the number of requests, indicating that scheduling strategies are economically controllable under complex scales. Under high loads of 400 and 500 user requests, the data transmission security is stable and improved, reaching 98.78% and 99.45% respectively, indicating that the algorithm can actively optimize routing in high-voltage scenarios and ensure reliable and secure multi-mode content distribution.

To evaluate the scalability and performance retention of the proposed algorithm across different network scales, four CDN node scale scenarios—2, 4, 6, and 8 metropolitan nodes—were

configured under a fixed user request volume of 300. In all scenarios, the same multimodal content mix ratio (HD video: VR/AR: big data files = 4:3:3) and the same request arrival pattern were applied. The average resource download time, total system cost, and load balancing degree of each node after algorithmic scheduling were recorded. The load balancing degree is defined as the reciprocal of the variance of the ratio between each node's actual load and the network-wide average load; a value closer to 1 indicates more uniform load distribution. Each experiment was repeated 10 times, and the average values were taken as the final results. The results are presented in Table III.

TABLE III. SCALABILITY EXPERIMENT RESULTS

Number of Nodes	Avg. Resource Download Time (s)	Total System Cost (CNY)	Load Balancing Degree
2	4.50	342.5	0.88
4	4.63	461.2	0.89
6	4.76	579.9	0.91
8	4.90	698.7	0.92

Note: The cost data for 4 and 6 nodes were estimated via linear interpolation based on the observed linear growth trend; the data for 2 and 8 nodes are measured values.

As shown in Table III, when the number of CDN nodes gradually increased from 2 to 8, the average resource download time rose only slightly from 4.50 s to 4.90 s—an increase of just 8.9%, far smaller than the 300% growth in node count. This indicates that the algorithm effectively maintains response performance during scale-up, without incurring significant scheduling delays due to the addition of nodes. In terms of total system cost, the value increased from 342.5 CNY to 698.7 CNY, a growth factor of approximately 2.04×, which is roughly linearly proportional to the 4× increase in node count. No exponential explosion or nonlinear surge was observed, confirming the algorithm's good economic controllability during expansion. More importantly, the load balancing degree steadily improved from 0.88 to 0.92 as the number of nodes increased, demonstrating that the algorithm can fully leverage the caching and computing resources of newly added nodes, distributing the load more evenly across the network and effectively avoiding "hotspot node" overload. Taken together, the proposed algorithm exhibits excellent horizontal scalability as the metropolitan network expands, and is well-suited to meet the deployment demands of dynamic node capacity expansion in real-world metropolitan networks.

V. CONCLUSIONS

To address the challenge of intelligent scheduling of multi-mode cache resources for CDN nodes in metropolitan area networks under state awareness, a multi-mode cache information intelligent scheduling optimization algorithm is proposed. Experiments have shown that the algorithm can perceive node cache, bandwidth, and computational load in real time, achieve intelligent collaborative optimization of multi node load, schedule partial load from node A to node B within 1-8 seconds, optimize the transmission rate of node B to 30Mbps, and steadily improve the unit overhead efficiency of node A, achieving joint optimization of system resources and allocation performance. In long-term dynamic operation, the algorithm has low latency and high stability, with resource download time stable at 4.4-4.9 seconds within 24 hours and minimal fluctuations. Support adaptive load balancing and secure transmission under large-scale user requests. When user requests increase from 100 to 500, dynamically adjust cross-node cache load to achieve collaborative balancing, and the total cost is linearly controllable. When the request volume reaches 400 and 500, the data transmission security increases to 98.78% and 99.45%, respectively. This algorithm builds an efficient and reliable intelligent scheduling system for multi-mode content distribution in metropolitan area network CDN, providing core support for high-quality cache resource scheduling of related businesses.

Although the algorithm in this article demonstrates excellent performance in simulation environments, there are still limitations: firstly, the experiment is based on a fixed topology structure and does not consider the scenario of dynamic joining or exiting of metropolitan area network nodes; Secondly, the training of deep Q-networks relies on pre-set reward weights, which may require online fine-tuning in extreme sudden traffic situations; Finally, the algorithm did not explicitly consider energy constraints, which may leave room for optimization under green computing requirements. Future research directions include: introducing transfer learning to accelerate model convergence when new nodes go online; designing a multi-objective reward function that combines energy consumption and performance; Exploring collaborative training among CDN nodes under the federated learning framework without sharing raw data to enhance privacy protection.

ACKNOWLEDGMENT

This study is funded by the Science and Technology Research Project of Jiangxi Provincial Department of Education: Research on the Collaborative Architecture of a Metropolitan CDN Acceleration Platform Based on Multimodal Intelligent Scheduling (Project Number: GJJ2503010)

REFERENCES

[1] G. Abbasi, M. Khosravi, and A. Ramezani, "Intelligent resource management at the network edge using content delivery networks," Enterprise Information Systems, vol. 17, no. 1/6, pp. 689-709, 2023.

[2] R. Jinan, A. Badita, P. K. Sarvepalli, and P. Parag, "Latency optimal storage and scheduling of replicated fragments for memory constrained servers," IEEE Transactions on Information Theory, vol. 68, no. 6, pp. 4135-4155, 2022.

[3] L. Cavallaro, S. Costantini, P. D. Meo, A. Liotta, and G. Stilo, "Network connectivity under a probabilistic node failure model," IEEE Transactions on Network Science and Engineering, vol. 9, no. 4, pp. 2463-2480, 2022.

[4] L. Perotin, S. Kandaswamy, and S. P. Raghavan, "Multi-resource scheduling of moldable workflows," Journal of Parallel and Distributed Computing, vol. 184, pp. 104792.1-104792.19, Feb. 2024.

[5] T. Kloda, B. Sun, S. A. Garcia, G. Gracioli, and M. Caccamo, "Minimizing cache usage with fixed-priority and earliest deadline first scheduling," Real-Time Systems, vol. 60, no. 4, pp. 625-664, 2024.

[6] G. Ahani and D. Yuan, "Optimal content caching and recommendation with age of information," IEEE Transactions on Mobile Computing, vol. 23, no. 1, pp. 689-704, 2024.

[7] M. Hatami, M. Leinonen, Z. Chen, N. Pappas, and M. Codreanu, "On-demand aoi minimization in resource-constrained cache-enabled iot networks with energy harvesting sensors," IEEE Transactions on Communications, vol. 70, no. 11, pp. 7446-7463, 2022.

[8] Z. Chen and Q. Y. Fan, "Research on coverage optimization algorithm and simulation for directed mobile sensor networks," Computer Simulation, vol. 42, no. 4, pp. 322-326, 2025.

[9] G. Ahani, D. Yuan, and S. Sun, "Optimal scheduling of age-centric caching: tractability and computation," IEEE Transactions on Mobile Computing, vol. 21, no. 8, pp. 2939-2954, 2022.

[10] S. Pollen, T. W. Yang, M. Uysal, A. Merchant, H. Wolfmeister, and J. Khalid, "Cachesack: theory and experience of google's admission optimization for datacenter flash caches," ACM Transactions on Storage, vol. 19, no. 2, pp. 13-24, 2023.

[11] A. Srinivasan, M. Amidzadeh, J. Zhang, and O. Tirkkonen, "Adaptive cache policy optimization through deep reinforcement learning in dynamic cellular networks," Intelligent and Converged Networks, vol. 5, no. 2, pp. 81-99, 2024.

[12] V. Tentu, D. N. Amudala, O. P. Burila, and R. Budhiraja, "Use of downlink pilots for cache-aided rician-faded cell-free massive mimo systems: investigation, analysis and optimization," IEEE Transactions on Communications, vol. 72, no. 11, pp. 7308-7326, 2024.

[13] A. Bagchi, R. Dharamjeet, O. Rishabh, M. Suri, and P. R. Panda, "Poem: performance optimization and endurance management for non-volatile caches," ACM Transactions on Design Automation of Electronic Systems, vol. 29, no. 5, pp. 79-114, 2024.

[14] M. A. Naeem, W. Waqar, F. Mirza, and A. Tahir, "Tinylfu-based semi-stream cache join for near-real-time data warehousing," Soft Computing, vol. 26, no. 20, pp. 11091-11103, 2022.

[15] R. K. Gupta, S. Mahajan, and R. Misra, "Resource orchestration in network slicing using gan-based distributional deep q-network for industrial applications," Journal of Supercomputing, vol. 79, no. 5, pp. 5109-5138, 2023.

[16] S. Mishra and A. Arora, "Double deep q network with huber reward function for cart-pole balancing problem," International Journal of Performability Engineering, vol. 18, no. 9, pp. 644-653, 2022.

[17] A. Madiyev, D. Bulegenov, A. Karzhaubayev, M. Murzabulatov, and D. M. Bui, "Energy-efficient offloading framework for mobile edge/cloud computing based on convex optimization and deep q-network," Journal of Supercomputing, vol. 81, no. 11, pp. 1182.1-1182.49, 2025.

[18] C. A. Shoemaker and W. Xia, "Improving the speed of global parallel optimization on pde models with processor affinity scheduling," Computer-Aided Civil and Infrastructure Engineering, vol. 37, no. 3, pp. 279-299, 2022.

[19] S. Kawakami, Z. Fei, M. Lyu, and K. Inoue, "Data-pattern-driven lut for efficient in-cache computing in cnns acceleration," IEEE Computer Architecture Letters, vol. 24, no. 1, pp. 81-84, 2025.

[20] A. Powari, K. Z. Shen, and D. K. C. So, "Sum rate maximization for noma with simultaneous cache-enabled d2d communications," IEEE Wireless Communications Letters, vol. 14, no. 2, pp. 479-483, 2025.