IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 September 2023, accepted 17 September 2023, date of publication 22 September 2023,
date of current version 27 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3318480

RESEARCH ARTICLE

QoS Aware Integrated Management Technique for 5G mmWave-Based Hetnets

L. MANJUNATH[1,2], N. PRABAKARAN[1], S. V. ASWIN KUMER[1], E. MOHAN[3], BALAJI NATARAJAN[4], G. SAMBASIVAM[5], (Member, IEEE), AND VAIBHAV BHUSHAN TYAGI[6]

[1]Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh 522302, India
[2]Department of ECE, CVR College of Engineering, Hyderabad, Telangana 501510, India
[3]Department of ECE, Saveetha School of Engineering, SIMATS, Chennai, Tamil Nadu 602105, India
[4]Department of Computer Science and Engineering, Sri Venkateshwaraa College of Engineering and Technology, Ariyur, Pondicherry 605102, India
[5]School of Computing and Data Science, Xiamen University Malaysia, Sepang 43900, Malaysia
[6]Faculty of Engineering, ISBAT University, Kampala, Uganda

Corresponding author: Vaibhav Bhushan Tyagi (tyagi.fict@isbatuniversity.com)

ABSTRACT One of the important ultimatums in enhancing the fifth generation (5G) network's capacity is the performance limit due to spectrum occupancy. The mmWave technology overlay on 5G heterogeneous network (HetNet) and caching the contents are proposed as the solutions to this problem. By reducing the backhaul links occupancy, increasing the access link utilization and caching, the impact of spectrum occupancy problem can be minimized. In our earlier work, differential Quality of Service (QoS) was provisioned by managing the cache and backhaul resources using machine learning techniques. In this work, an integrated solution combining content, cache and user management is proposed to maximize network utilization and QoS. The problem of effective utilization of the network at the same time ensuring the QoS for users are solved as a multi objective optimization problem with the aid of hybrid meta heuristics with complementary exploration and exploitation capability. The user association to base station is made adaptive to load and cache hit ratio at the base station. By increasing the content proximity to the users, the load of backhaul links is minimized. Through these integrated management strategies, the proposed solution is able to provide higher QoS compared to existing works in terms of reduction in packet drop by 6%, reduction in delay by 39%, increase in network throughput by 8% and a consistent cache hit ratio more than 85%.

INDEX TERMS MmWave backhaul, hetnet, caching, multi criteria optimization, hybrid meta-heuristics.

I. INTRODUCTION

Smartphone revolution and Internet of Things (IoT) have created unprecedented demand for mobile data traffic. The data traffic is increasing exponentially [1] and traditional cellular networks are no longer able to solve the unprecedented demand and service quality disruptions due to outbreak of mobile data services. Fifth generation (5G) mobile communication systems are designed to address these challenges in rapid outbreak of mobile data services [2]. 5G networks amalgamate various solutions like mmWave, massive Multiple input Multiple output (MIMO) and Heterogeneous networks (Hetnets) to intensify the network capacity [3]. But these techniques are based on co-existence of backhaul between base station (BS) and core network. The effectiveness of these techniques depends on how well the traffic on backhaul links are managed. The traffic on backhaul links must be reduced and congestion bottleneck must be avoided without degrading the quality of service for users [4]. The effectiveness of techniques to improve the utilization of 5G networks is measured in terms of average potential throughput (APT) [5]. The spectrum resources shared between access and backhauling influences the APT. Various studies have pointed out the influence of backhaul links utilization over APT ([6], [7]). Reducing the backhaul link congestion increases the

The associate editor coordinating the review of this manuscript and approving it for publication was Yogendra Kumar Prajapati.

APT [8]. Various solutions like Caching ([9], [10], [11]), MIMO cooperation [12], content coding [13], user association [14] etc. have been proposed to reduce the congestion on the backhaul links. The existing works can be categorized to user management, content management and cache management approaches.

Most of these approaches do not consider multiple factors like ensuring QoS for users, adaptation to network characteristics & application traffic dynamics and over all utilization of network resources etc in their management decisions. Due to these factors their performance gains are limited. Based on this observation, this work proposes a integrated solution combining user, content and cache management with consideration for multi factor optimization. The decision on association of user to the secondary base station (SBS), caching of content, proactive downloading of content etc. are made in way to optimize the multi criteria factors like delay, cache utilization, ensuring user QoS, reducing load on backhaul links etc. The decisions in integrated user, content and cache management with consideration for multi factor optimization is solved as multi criteria optimization problem. Novel contribution of this work is listed below:

(i) Firstly, an integrated user, content and cache management technique to maximize the backhaul link utilization with consideration for multi criteria factor optimization is framed. The user association to SBS is based on temporal access patterns and load of the base SBS. Contents to be cached are selected based on popularity and similarity to popular contents learnt through Latent Dirichlet Allocation. By this way, popular items are globally placed and load on backhaul links is shifted to SBS, avoiding congestion on backhaul links in the proposed solution. Due to this, the QoS of the network in terms of packet delivery ratio, latency etc. are improved. Secondly, the problem of deciding the items to be cached is solved as optimization problem. Solution to this problem is found using hybrid meta heuristics combining particle swarm optimization with Bat algorithm. Use of hybrid meta heuristics with complement exploration and exploitation capability solves the local minima problem is using single optimization algorithm. Due to optimal placement of items in cache, the cache hit ratio increases in the proposed solution and this reduces the load on backhaul links. As the result, throughput of the network is increased.

Organization of sections is as follows. The survey on existing works to maximize the backhaul link utilization is discussed in Section II. The proposed integrated user, content and cache management solution is presented in Section III. Section IV presents the results of the proposed solution and comparison to most recent works. Section V concludes the work and presents the future work scope.

II. SURVEY

The survey in conducted in three categories of user management, content management and cache management.

A. USER MANAGEMENT

Mesodiakaki et al. [15] proposed heuristic solution to associate user and calculate backhaul route with minimal energy consumption. At every time period the solution finds the optimal configuration based on current demands from the user and user locations. A tradeoff between energy consumption and satisfying flow demand requirements were made through heuristics algorithms. Mesodiakaki et al. [16] used reinforcement learning for user association. Based on the network congestion, bandwidth is split to Secondary Base Station. Users are associated to SBS in such a way to fairly balance the load across SBS. Feng and Mao [17] managed the allocation of backhaul link resources to users using deep reinforcement learning (DRL). The backhaul resources are split to blocks and allocation of blocks is done using DRL to minimize the congestion on backhaul. Authors did not consider the QoS guarantee for users in resource allocation. Klaine et al. [18] allocates user to cell using reinforcement learning. The users are allocated to determine best cell based on their requirements. The solution is designed to minimize user dissatisfaction as much as possible. Ma et al. [19] solved the problem of associating user requirements to resources by using optimization techniques. The objective was to maximize the fairness utility of the backhaul links. Formulating the optimization problem as mixed integer non linear programming, author used Lagrangian dual method to solve it. Author considered network utility fairness as the only criteria in user association decision. Lee et al. [20] proposed a user association scheme, by loading balancing the backhaul resources. The problem of load balancing is solved using branch-bound technique. Maximizing call blocking probability was the factor considered for user association in this work.

B. CONTENT MANAGEMENT

Fadlallah and Tulino [21] proposed caching-aided coded multicasting with the aim of improving bandwidth efficiency. The content is split to chunks and different parts of chunks are at different nodes. On request, the missing chunks are collected from other nodes and assembled. This scheme maximizes the storage utilization of cache and reduces the backhaul traffic. The fractional coding considered in this work is random and does not give importance to portions of content which needs to get prioritized for storage. Mao et al. [22] reduced latency and increased reliability using linear network coding. Authors proposed two schemes: rate proportional traffic splitting scheme and adaptive coded forwarded scheme. Both these schemes performed well in multi hop scenario, but the computational complexity of the scheme is higher at user end. Carreira et al. [23] proposed a scalable video coding architecture with high bit saving rate over backhaul links. The scheme is applicable for only video frames and does not consider audio contents. Though the scheme has provisions for adapting to network conditions, it did not consider quality degradation during higher congestion in network.

TABLE 1. Comparison of literature survey.

Author	Work Summary	Proposed Solution in Our Research
Mesodiakaki et al [15]	Addressed the joint problem of user association and backhaul routing with two objectives of providing high spectrum efficiency and minimizing routing power consumption	Improvising QoS and cache management is not considered when compared to proposed solution
Sande et al [16]	Reinforcement learning was used to associate user to base station based on load	The proposed solution associated not only based on load but also based on history of cache hit ratio
Feng et al [17]	Reinforcement learning was used to manage the backhaul resources and minimize the congestion on backhaul links.	Compared to proposed solution, this approach lacked QoS improvement and Cache management.
Somesula et al [26]	Caching decision was based on deadline and cost benefit using fuzzy logic	The proposed solution considered five factors as caching decision as compared to only two considered in the existing work
Atiqur et al [27]	Caching decision was based only on user satisfaction rate	The proposed solution considered network factors too in addition to user satisfaction rate
Chiang et al [24]	Caching decision was based only on reducing energy consumption	The proposed solution considered user factors and network factors too in addition to energy consumption in the caching decision
Tao et al [29]	Optimization based caching solution was proposed to reduce the transmission power and backhaul cost	The proposed solution reduced the backhaul usage and indirectly reduced the backhaul cost. In addition, it also improved the network utilization and user QoS.
Wang et al [30]	Authors considered joint design and optimization of the caching and user association policy to minimize the average download delay.	The proposed solution also considered reducing packet loss, increasing throughput in addition to reducing the delay. Caching decision is made on five different network and user factors in proposed solution compared to one factor in existing work
Chatzieleftheriou et al [31]	Authors considered joint user association and content caching. The caching decision was optimized to maximize the total cache hit ratio	The proposed solution considered five different factors in caching decision in addition to maximizing the total cache hit ratio.
Li et al [32]	Authors optimized content caching and user association to minimize the content download latency. Content caching is based only on popularity.	The proposed solution considered five different user, network factors in caching decision. Content caching considered both content similarity and popularity.

Torre et al. [25] proposed a novel content distribution framework for 5G network with objective of reducing energy consumption and latency. The contents are coded using Random Linear Network Coding (RLNC) before distribution. The scheme was not tested for the realistic scenario of contents with different popularity in access requests from different nodes.

C. CACHE MANAGEMENT

A fuzzy logic based caching strategy was proposed by Somesula et al. [26]. Caching decision was based on deadline and cost benefit. Integer linear programming was used to decide the best place to cache the content with delay as optimization criteria. Authors also predicted the content request using echo state network and use the prediction information to cache the contents. User differentiation based on priority was not considered in this work. Caching was proposed as a solution to improve the Quality of experience in wireless network in Atiqur et al. [27]. User satisfaction rate improvement is used as deciding factor in cache management. Satisfaction rate is calculated based on waiting time for user requests. Caching is proposed as a solution to reduce the energy consumption of backhaul links by Chiang and Liao [24]. Energy consumption is reduced by placing contents close to request location. Placement of content is controlled using a greedy strategy. Authors also reduced the energy consumption due to inter cell interferences using cooperative download. QoS improvement and user prioritization were not considered in this work. Pantisano et al. [28] used the content availability information in cache to decide the bandwidth allocation for user. Estimated time to deliver the content is calculated based on content availability and user preference. Bandwidth is allocated proportional to estimated time to deliver the content. User starvation is a problem in this approach. Optimization based caching solution was proposed by Tao et al. [29]. Mixed integer nonlinear programming was used to reduce the transmission power and backhaul cost. User requests are processed by grouping the users based on requested content. The solution cannot be applied in scenarios of diverse request frequency. Optimization algorithm was used to decide cache placement in this work. Backhaul rate and energy consumption rate minimization were the optimization criteria considered in this work. The authors also encoded the contents using Maximum Distance Separable (MDS) code, so that the content can be reconstructed with minimal packets at receiver end. By effectively reducing the number of packets transmitted, energy consumption was lessened at backhaul links, but this approach is not scalable. The comparison summary of the survey is presented in Table 1.

From the survey, most approaches do not consider multiple factors for optimization. Integration of content, cache, and user management approaches with consideration for multi factor optimization is not existent in literature. This integration can bring higher performance gains, and this motivates our research work.

III. PROPOSED CROSS LAYER OPTIMIZATION

The proposed integrated solution involves joint management of user, content, and cache at SBS. The proposed system is for the network model given in Figure 1.

The User Equipment (UE) are served by the SBS in their coverage area. SBS are connected to Base station (BS) via the mmWave backhaul links. Content requests from UE are directed to SBS which first checks the availability of the content in its local cache. If the content is found in the local cache, it is served to UE by SBS using the wireless link between UE and SBS, thereby avoiding load on backhaul links. If the content is not available in the SBS cache, request is directed to BS which downloads the content from internet and provides to SBS from where it is sent to the UE. The architecture of the proposed integrated management solution is given in Figure 2. The core of the work is content selection and content placement to cache. This process is done considering multiple QoS factors and user preferences. Content to be placed in cache is selected by BS using the content selection module. The content selection is based on popularity of the content and its similarity to other popular content. In addition, BS can receive predicted popular contents from the recommendation servers and proactively upload to cache during the idle times of backhaul links. This improves the backhaul link utilization. The Content selection happens periodically. From the contents selected, the placement of contents to cache of SBS is solved as multi factor optimization problem using hybrid-meta heuristics in the content placement decision module.

This decision facilitates the probability for user to get the requested content from cache and delay in downloading the content from SBS to be minimized. The contents decided to be placed in the cache are split to blocks by content coding module and distributed across the SBS. Each of the modules are detailed below in the following subsections.

The notations used in the equations are presented in Table 2.

A. CONTENT SELECTION

The contents with higher popularity are selected for placing in the cache. The content popularity (f) is calculated as:

$$PC\,(f) = \frac{p_c\,(f)}{\sum_{f=1}^{F} p_c\,(f)} \tag{1}$$

$p_c(f)$ is calculated as:

$$p_c\,(f) = \begin{cases} P_b\,(f)\,, & if\ f \in C \\ 0, & otherwise \end{cases} \tag{2}$$

where, C is the type to which content f belongs and

$$P_b\,(f) = \frac{1}{U} \sum_{u=1}^{U} p\,(u)\,.p(C|u) \tag{3}$$

The probability of user placing the request for content f ($P_b(f)$) is given as p (u) and p(c|u) is the probability of content belonging to content type C. Content popularity for a new

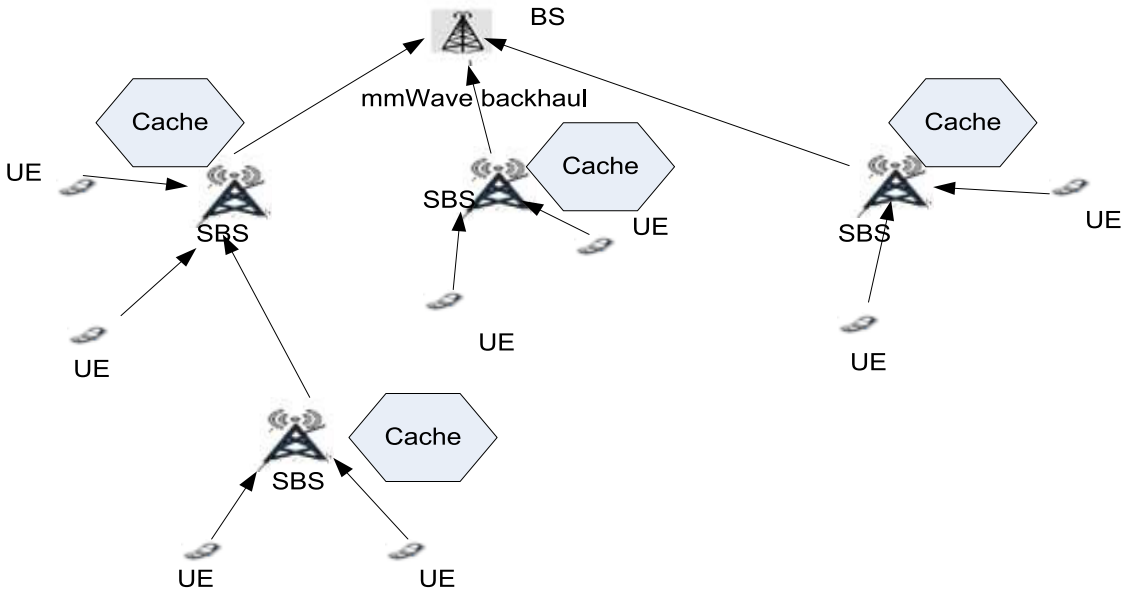

FIGURE 1. Network model with SBS connecting to BS via mmWave backhaul at top level and far off SBS are connected to SBS close to BS.

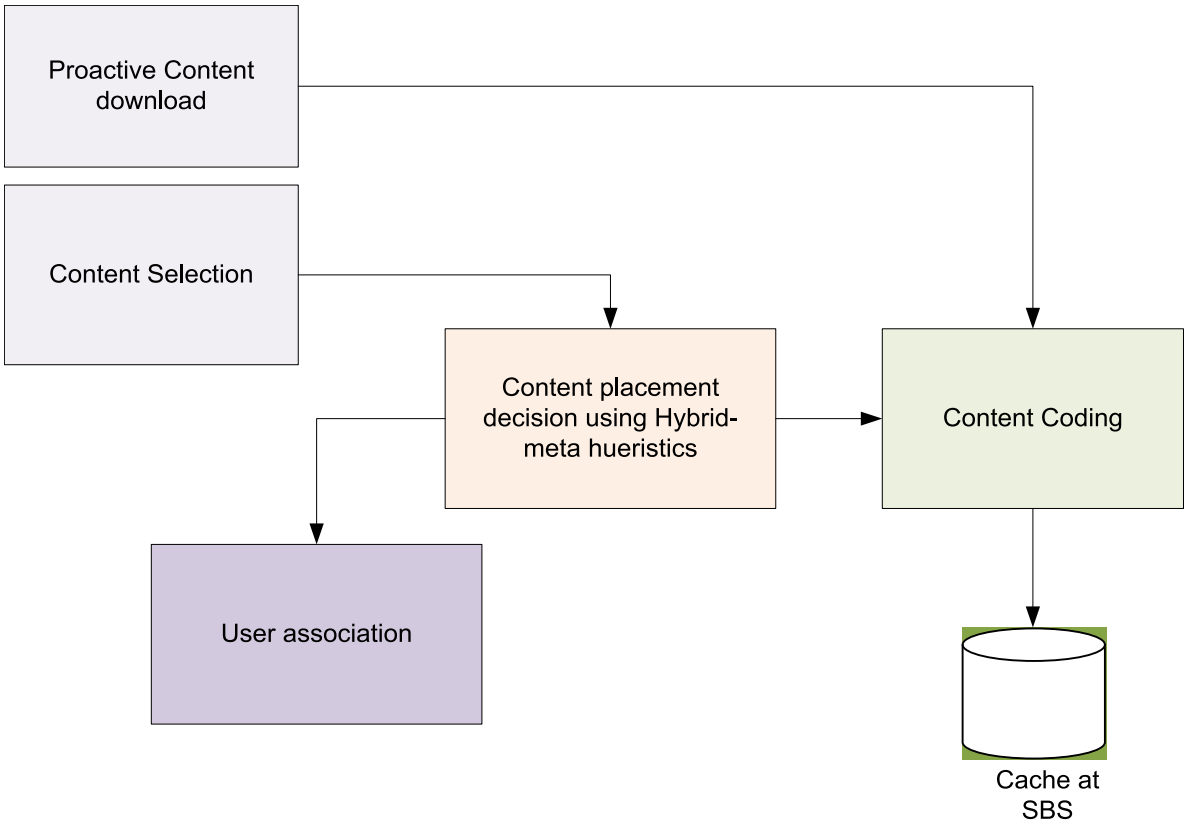

FIGURE 2. Proposed integrated management architecture combining content, user and cache management.

content is predicted based on its relation to contents already in cache. Meta data of content is used in modelling the relation between the contents. Latent Dirichlet Allocation is used for topic modelling from the meta data. The same concept is used in this work to get the topic model of the contents. Say, there are two contents A and B, the similarity between the content is found by first finding their corresponding topic model (TP (A) and TP(B)) and finding the similarity between

their topic model as:

$$Sim\,(X, Y) = \frac{\sum_{j=1}^{N} TP(A)_i * TP(B)_j}{\sqrt{\sum_{j=1}^{N} TP(A)_j^2}\sqrt{\sum_{j=1}^{N} TP(B)_j^2}} \quad (4)$$

The popularity of the maximum similarity item is decided as the popularity of the new content. The output of the content selection is K popular contents.

B. CONTENT PLACEMENT

The K contents selected must be placed in M SBS based on multi factor optimization subject to constraints on cache size. The cache size at SBS is limited and it must be populated with contents in such a way that the load on backhaul link is minimized as much as possible. UE must be associated to SBS based on following requirements:

(i) The Content fetching probability from the cache of SBS is maximized.
(ii) The Delay in downloading the content through the backhaul link is minimal.
(iii) The User QoS must be maximized.

The contents to be delivered from BS to SBS can be coded to maximize the utilization of cache. Instead of storing the entire content at cache of one SBS, it can be split to blocks and stored in nearby SBS with minimal path loss. In summary, the placement of contents in the cache at SBS is based on multiple factors as listed below:

(i) Minimization of delay (D) at backhaul links
(ii) Maximization of cache utilization (Cu)
(iii) Maximization of user QoS (Q)
(iv) Minimization of load on backhaul links (L)
(v) Maximize the cache hit ratio (R)
(vi) Minimization of path loss between SBS (PL)

The delay at backhaul links for transmitting the files is given as:

$$D = \frac{1}{W log_2(SNR)} \quad (5)$$

where W is the bandwidth and signal to noise ratio (SNR) is calculated as:

$$SNR = \frac{Pt^{-\alpha}}{\sigma_N^2} \quad (6)$$

where Pt is the transmission power of BS, α is the path loss component, σ_N^2 is the Gaussian white noise power.

Cache utilization is measured as the ratio of number of unique items over all the SBS (N_U) to the total items that can be cached (N_{CT}). It is given as:

$$Cu = \frac{N_U}{N_{CT}} \quad (7)$$

User QoS is measured in terms of projection of average delay that will be experienced by users in downloading the content. The delay varies from the maximum value (d_{max}) to minimum value (d_{min}) for the unit of content. The value is d_{max}, when the content is directly downloaded from BS and it is d_{min}, when the content is downloaded from cache. Based on this, QoS is formulated as:

$$Q = \sum_{i=1}^{n} \frac{S_i}{f(d_{max}, d_{min})} \quad (8)$$

where n is the total number of contents and S_i is size of the content i.

$$f\,(d_{max}, d_{min}) = \begin{cases} d_{max}, & conent\ is\ not\ in\ cache \\ d_{min}, & content\ is\ in\ cache \end{cases} \quad (9)$$

The Load on the backhaul link is measured as the ratio of the bandwidth needed to download contents (W_C) to the bandwidth available at backhaul links (W_v), given as:

$$L = \frac{W_C}{W_v} \quad (10)$$

Cache hit ratio is formulated as ratio of number of contents found in cache (N_F) to total requests made (N_T), given as:

$$R = \frac{N_F}{N_T} \quad (11)$$

The path loss can be formulated as:

$$PL = C_L d^{-\alpha} \quad (12)$$

where C_L is loss intercept, d is average distance between SBS and its blockage, α is path loss exponent. The value of α_L ranges from 1.9 to 2.5 for mm-Wave links.

The caching decision must be evaluated against multi factors. The fitness function on multiple factors is formulated as:

$$F = w_1\frac{1}{D} + w_2 C + w_3 Q + w_4\frac{1}{L} + w_5 R + w_6\frac{1}{PL} \quad (13)$$

TABLE 2. Basic notations used in equations.

PC (f)	Popularity of content f
p(u)	Probability of user requesting for content
p(c|u)	Probability of conent u belonging to type C.
TP(X)	Topic vector score of content X
W	Bandwidth
Pt	Transmission power of BS
α	Path loss component
σ_N^2	Gaussian white noise power
f_{min}, f_{max}	Minimum and maximum frequency of Bat
v_i^t	Velocity of ith bat at time t.
x_i^t	Position of ith bat at time t
$X_i(t)$	Position of ith particle at time t
$V_i(t)$	Velocity of ith particle at time t
S_{best}	Locally best solution found by a particle
g_{best}	Globally best solution found by a particle
C_L	loss intercept
Cu	Cache utilization
d	average distance between SBS and its blockage

w_1 to w_6 are the weights allocated to each of the factors and the value is allocated in such a way that:

$$\sum_{i=1}^{6} w_i = 1 \tag{14}$$

The Modeling of the placement of the contents in SBS cache as a multi objective optimization problem, is commonly used technique. But this work proposes to use hybrid-meta heuristics to solve the problem. Compared to single optimization, hybrid meta heuristics better solves the local minima problem. Among the multiple combinations, this work adopts Particle Swarm Optimization (PSO) with Bat algorithm.

PSO is a bio inspired algorithm simulating the social behavior of swarms. Individual organism in the swarm shares its discovery with others, so that collectively everyone gets close the best hunt. This principle is emulated in the PSO algorithm to search for the optimal solution in the solution space. Compared to other optimization algorithms, PSO is simple, flexible, and versatile. Each organism moves randomly with different velocities and updates their positions. In PSO, each candidate solution is referred as particle. The particle updates its position((X_i) with a velocity(V_i) based on both local best (p_{best}) and global best (g_{best}) particle positions. Each particle updates it positions at time $t+1$ based position at time t and velocity calculated at time $t+1$ as in equation 15.

$$\begin{aligned} X_i\,(t+1) &= X_i\,(t) + V_i(t+1) \\ V_i\,(t+1) &= wV_i\,(t) + c_1 r_1\,(S_{besti}\,(t) - X_i\,(t)) \\ &\quad + c_2 r_2\,(g_{besti}\,(t) - X_i\,(t)) \end{aligned} \tag{15}$$

In equation 15, the configuration values c_1 and c_2 control the acceleration (number of position changes in the solution), r_1 and r_2 are the random values controlling the degree of acceleration. Number of particles (configurable) starts with random solution in the solution space. At each iteration, fitness value is calculated for each of the particle or solution they represent. The particle with maximum value of fitness function is selected as globally best. Other particle updates their position based on globally best and locally best solution among its closest neighbors. This process is repeated till the maximum iteration is reached or when there is no further change in the solution. The problem in this optimization method is that sometimes the result can converge to locally minimal solution. The proposed work solves this problem by using a hybrid optimization method, where the PSO is combined with Bat algorithm, so that hybrid method has better exploration and exploitation capability.

Bat algorithm is based on the natural behavior of bats in hunting for their foods. Bat hunt for their food using sonar. They start in their search for prey with a random velocity v_i from position x_i at sonar frequency of f and loudness A_o. They adjust the frequency of emitted pulses based on their proximity to the pray. Bat algorithms use this principle to move to the best solution in solution space in iterations. In each iteration, the position, velocity, and frequency are updated based on maximization of fitness function value. The fitness function is designed to model the proximity to the best solution. In bat algorithm, multiple bats are started with each at different positions. Position is a random solution in solution space. Each bat updates its position at time t based on its past position and velocity calculated at time t as

$$x_i^t = x_i^{t-1} + v_i^t \tag{16}$$

The velocity at time t is calculated as

$$v_i^t = v_i^{t-1} + \left(x_i^{t-1} - x_*\right) f_i \tag{17}$$

The frequency for updating the velocity is calculated as

$$f_i = f_{min} + (f_{max} - f_{min})\,\beta \tag{18}$$

β is the random variable and frequency ranges from minimum (f_{min}) to maximum (f_{max}) value. Exploration of pray in search space is facilitated with frequency adaptation.

The goal of hybrid optimization is to solve the local minima problem by combining optimization algorithms with better exploration and exploitation capability. In this work PSO with good exploration capability is combined with Bat algorithm having good exploitation capability to solve the local minimal problem. Bat algorithm starts with initial random solution and results in optimal solutions. PSO starts with the best solution found by Bat and results in final optimal solution.

The solution in our case is the content to be cached in each of the SBS. A M list (SBS) with K array items (K contents) is the solution space. K array is filled with 1 when particular content is cached at the SBS and it is 0 when the content is not cached at SBS. Say the cache size is L then only L items in K sized array can be 1. Initial solution is found using Bat algorithm which is M list of K binary array. Bat's final solution is given to PSO as the initial solution. The result of PSO is the optimal solution space having M list with K binary vector. The contents are placed to M SBS based on K binary vector decision Both PSO and Bat algorithm tries to maximize the fitness function F defined in Equation 13.

C. USER ASSOCIATION

Users are assigned with a preference score for all the SBS within its communication range. The preference score (*Score*) is based on content preference (*CP*) and load of the SBS (*Load*). The weighted sum of content preference and load is calculated as preference score.

$$Score\,(SBS) = m_1 \times CP(SBS) + m_2 \times Load(SBS) \tag{19}$$

m_1 and m_2 are the weights assigned to *CP* and *Load*. The value can be assigned from 0 to 1 in such a way that,

$$m_1 + m_2 = 1 \tag{20}$$

Content preference is the average cache hit ratio on a particular SBS. It is calculated periodically at each SBS and averaged. The load is calculated in terms of number of current users associated with the SBS. The SBS in communication range of user with highest preference score is selected for association of user.

D. CONTENT CODING

Keeping the entire content in SBS, reduces the cache utilization. To solve this problem, the contents are split to blocks and some of the blocks are distributed to neighboring SBS cache. When the user requests for the contents, the blocks are downloaded from neighboring SBS through SBS-SBS communication links. This avoids the need for keeping redundant contents in neighboring SBS.

IV. RESULTS

The proposed solution was simulated in MATLAB with configuration as given in Table 3. The topology given in Figure 1 is simulated using discrete event simulation model.

The content requests from UE are directed to SBS and satisfied at SBS if content is at SBS cache. Otherwise the requests are redirected to BS for content download. The requests and responses from each entity of UE, SBS and BS is realized as events. At the end of simulation time, performance parameters are calculated.

TABLE 3. Setup for simulation.

Parameter	Value
Users	50 ,100,150,200,250
Number of SBS	10
Area of simulation	3000 m × 3000 m
Simulation time	20 minutes
Content request rate from UE	1 request per 10 sec
Average Content size	500 KB
Cache size at SBS	5000 to 10,000 KB
Energy consumption per bit transmitted at UE,SBS,BS	0.5×10^{-8} joules/bit
Coverage radius of SBS	200 m
Coverage radius of BS	1500 m
Location of BS in simuation area	Center of area
UE priority level	1 to 3, 1 being highest
Performance comparison parmaeters	Packet drop rate (PLR), Delay, Cache hit ratio (CHR) , Throughput

Average value of packet drop rate, delivery ratio, delay and cache hit ratio are found through simulation and compared to most recent existing works. Packets drop rate (P) is measured as:

$$PLR = \frac{Number\ of\ packets\ dropped}{Number\ of\ packets\ from\ UE\ to\ BS} \quad (21)$$

Delay is measured as time taken for packet to transit from UE to BS. Cache hit ratio (CHR) is measured as:

$$CHR = \frac{Number\ of\ times\ content\ placed\ in\ cache}{Number\ of\ times\ content\ requested} \quad (22)$$

The effectiveness of proposed work is compared against fuzzy logic-based cache management solution proposed by Somesula et al. [26], content management-based solution proposed by Torre et al. [25] and user management based solution proposed by Zhang et al. [14].

The average packet loss rate is measured at the backhaul links as ratio of packet dropped to total packet scheduled for delivery over the backhaul links. The average packet loss rate over the backhaul links is an indicator of congestion of the links and lower the value, better the backhaul link is utilized. Table 4 provides the measurement of average value of packet drop rate across the solutions for varied number of UE's.

TABLE 4. Comparison of average packet loss.

No. of UE's	Proposed	Zhang et al [14]	Torre et al [25]	Somesula et al [26]
50	6	13	18	19
100	7	14	19	21
150	8	15	21	22
200	9	16	22	23
250	10	17	23	24
Average	8	15	20.6	21.8

Compared to the proposed work, Zhang et al. [14], Torre et al. [25], Somesula et al. [26] has 7%, 11.4%,12.2% higher packet drop rate. The packet drop rate at backhaul links has reduced due to integrated management of user, content and cache management in proposed solution compared to separate management of user, content, and cache. Figure 3 presents the results for packet delivery ratio (PDR) (100-P) over the simulation time. Compared to existing works, proposed solution has atleast 6% higher packet delivery ratio. The increase in packet delivery ratio in proposed work is due to selection of more stable path and control of congestion in those stable paths in the proposed solution The packet delivery ratio is measured for varying content request rate and the result is given in Table 5.

The average packet delivery ratio drops with increase in the content request rate but even at higher content request rate, the PDR is higher in proposed solution compared to existing works. On average, the PDR is atleast 9% higher compared to existing works with increase in content request rate. The PDR has increased in proposed solution due to availability of contents at close proximity.

The average delay for delivery of packet with content payload is measured at the backhaul links for different number of UE' and results are presented in Table 6.

Compared to Zhang et al, Torre et al and Somesula et al the proposed solution has 39%, 55% and 87.5% lower delay.

The delay has reduced in the proposed solution by large factor due to three reasons (i) association of UE to optimal SBC where load is minimal (ii) perfecting of content, so that

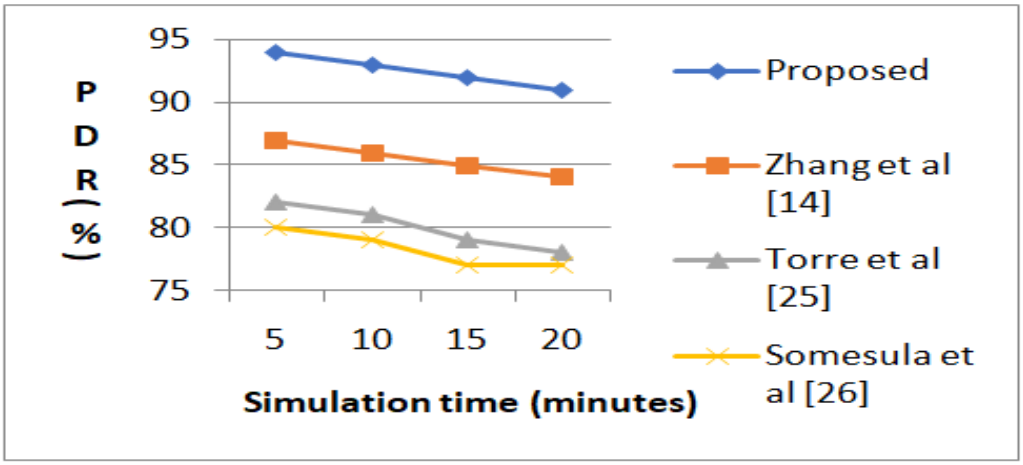

FIGURE 3. Comparison of PDR over the simulation time.

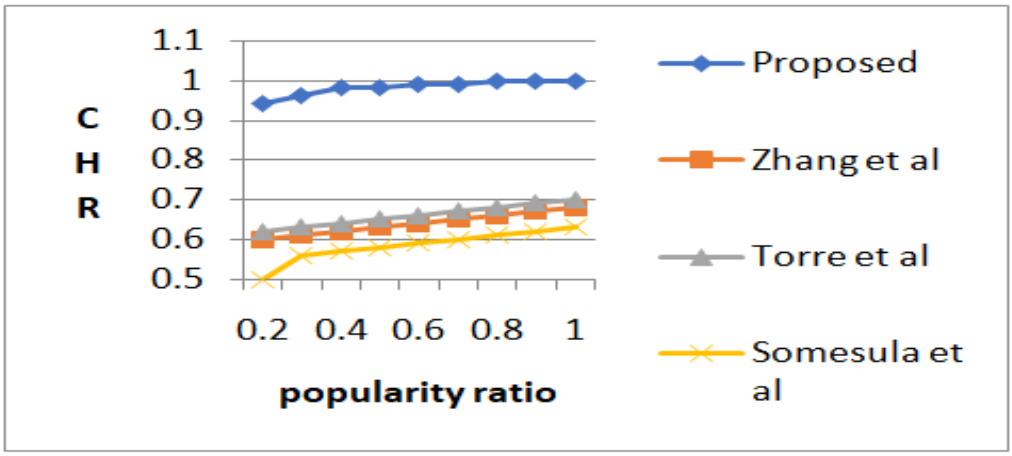

FIGURE 4. Comparison of cache hit ratio (CHR) varying popularity ratio.

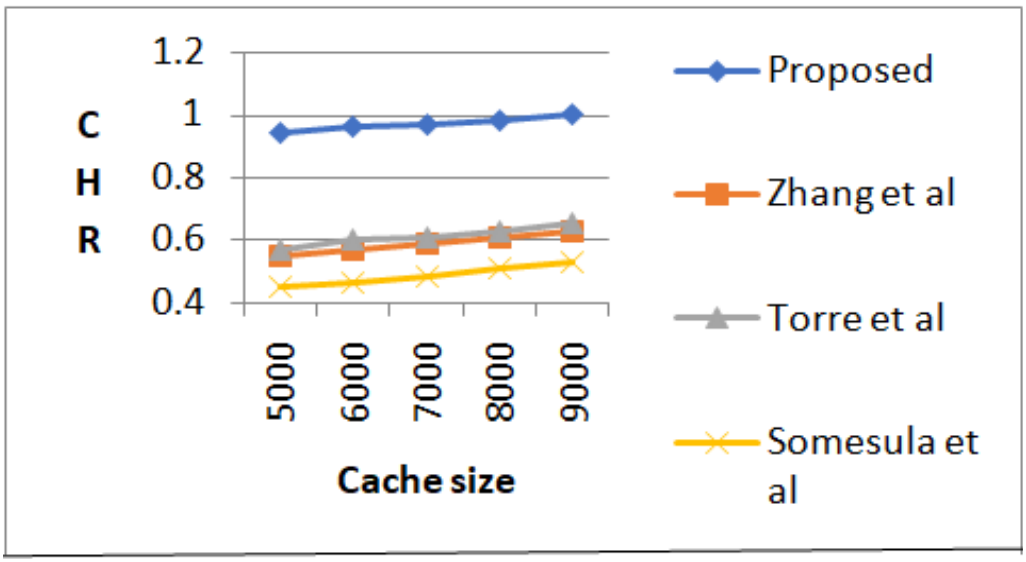

FIGURE 5. Comparison of cache hit ratio (CHR) varying the cache size.

TABLE 5. Comparison of PDR.

Content request rate	Proposed	Zhang et al [14]	Torre et al [25]	Somesula et al [26]
1	94	87	82	80
5	89	80	77	76
10	86	75	73	72
Average	89.6	80.6	77.3	76

content can be found with maximal probability at SBC and (iii) reducing the transferable bit size due to content coding.

TABLE 6. Comparison of average delay.

No. of UE's	Proposed	Zhang et al [14]	Torre et al [25]	Somesula et al [26]
50	13	18.1	20.9	24.1
100	12	16.2	18.1	22.2
150	11	15.1	17.2	21.1
200	10	15.2	16.1	20.2
250	10	14.2	15.2	18.1
Average	11.2	15.76	17.5	21.14

The average throughput is measured at backhaul links varying the different number of UE's and the result is given in Table 7.

TABLE 7. Throughput measurements.

No. of UE's	Proposed	Zhang et al [14]	Torre et al [25]	Somesula et al [26]
50	95	84.1	82.9	80.9
100	147	136.2	129.1	123.8
150	189	173.1	166.4	159.6
200	228	208.2	200.9	194.8
250	272	251.1	242.2	237.6
Average	186.2	170.54	164.3	159.34

Compared to Zhang et al, Torre et al and Somesula et al, the proposed work has 8%,11% and 14% higher throughput. Higher throughput in proposed solution is due to significant reduction in delay over backhaul links and scheduling of predictive downloads in the idle times of backhaul. In all three existing works, predictive download of contents during idle time of backhaul has not been considered. For various number of UE, energy consumed in measured and the result is presented in Table 8.

Compared to Zhang et al, Torret et al, Somesula et al, the proposed solution has 9%, 12%,18% lower energy consumption. The energy consumption has reduced in proposed

TABLE 8. Comparison on energy consumed.

No. of UE's	Proposed	Zhang et al [14]	Torre et al [25]	Somesula et al [26]
50	75	79.8	82.4	85.2
100	86	91.6	92.6	97.1
150	91	96.8	100.2	105.6
200	96	103.5	107.8	115.6
250	99	116.1	117.4	126.4
Average	89.4	97.56	100.08	105.98

solution dueto reduction in congestion and higher probability of locating the contents in the cache.

By varying SBC cache size, Cache Hit Ratio (CHR) is measured, and the results are presented in Figure4. Compared to existing works, the CHR is higher in proposed solution at all cache size. On average, CHR is at least 10% higher in the proposed work.

The cache hit ratio has increased due to optimal selection of items to cache based on multi criteria optimization factors in the proposed solution.

By varying the popularity ratio, the cache hit ratio is measured and the results are given in Figure5. As the popularity ratio, increases, the choice of content to be cached becomes difficult. This can reduce the cache hit ratio if the items to cache are not distributed in such way without much redundancy.In the proposed solution, the items to cache are distributed minimizing the redundancy. Due to this cache hit ratio was higher and consistent in the proposed solution compared to existing works.

The proposed solution integrated cache management, content management and user management with joint consideration multi object QoS optimization. The aim in comparison was to demonstrate that the proposed integration performed better than individual strategies. In accordance with this aim, three recent solutions in three strategies of cache management, content management and user management were selected for comparison. For cache management strategy, Somesula et al. [26] was selected, for content management strategy Torre et al. [25] was selected, for user management strategy Zhang et al. [14] was selected. Somesula et al considered delay as only optimization parameter, Torre et al considered energy and latency as the optimization parameters and Zhang et al. considered delay as the only parameter for optimization. In comparison to these solutions, the proposed solution considered multiple parameters of delay, cache utilization, user QoS, load on backhaul links, cache hit ratio, path loss in optimization. Integration of all the three management strategy along with multi parameter optimization has improved the performance gain in the proposed solution. In addition to delay and packet drop reduction, throughput and cache hit has increased in the proposed solution. Though optimization strategies and integer linear programming are

used in existing works, they did not consider the problem of local minima during optimization Use of hybrid meta-heuristics combining two optimization techniques with exploration and exploitation has solved the local minima problem in proposed solution.

The computational complexity of hybrid metaheuristics algorithm for deciding the optimal content to cache is measured varying the cache size and the result is given in Table 9.

TABLE 9. Execution time of hybrid meta heuristics.

Cache size	Execution time (s)
5000	40
6000	52
7000	67
8000	94
9000	108

The execution time increases linearly with the increase in cache size. Though the execution time is higher, it is justified as the hybrid meta heuristics algorithm execution to decide items to be cached is not executed frequently and done once in a longer interval of time. One way to time it is to check if the cache hit ratio drops below a threshold or if the number of new arrival items is greater than a threshold. The method to time the execution of hybrid metaheuristics algorithm is considered for future research.

V. CONCLUSION

An integrated solution combining user, content and cache management was proposed in this work. The decision on items to be cached is made using hybrid meta-heuristics solution involving multiple optimization factors. Once the items are found to be cached in each SBS, the decision of user association is done to reduce the delay and congestion. Proactive downloading of content is done to increase the backhaul link utilization. The overall impact of the integrated solution is that it has 7% lower packet drop, 39% lower delay, 8% higher throughput and consistent cache hit above 85%. Testing the solution using different hybrid metaheuristics algorithms is in the scope of future work.

REFERENCES

[1] *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021 White Paper*, Cisco, San Jose, CA, USA, 2017.

[2] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 3rd Quart., 2016.

[3] X. Jia, M. Zhou, M. Xie, L. Yang, and H. Zhu, "Optimal design of secrecy massive MIMO amplify-and-forward relaying systems with double-resolution ADCs antenna array," *IEEE Access*, vol. 4, pp. 8757–8774, 2016.

[4] S. Biswas, T. Zhang, K. Singh, S. Vuppala, and T. Ratnarajah, "An analysis on caching placement for millimeter–micro-wave hybrid networks," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1645–1662, Feb. 2019.

[5] A. AlAmmouri, J. G. Andrews, and F. Baccelli, "A unified asymptotic analysis of area spectral efficiency in ultradense cellular networks," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 1236–1248, Feb. 2019.

[6] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Trans. Commun.*, vol. 61, no. 10, pp. 4391–4403, Oct. 2013.

[7] Z. Shi, Y. Wang, L. Huang, and T. Wang, "Dynamic resource allocation in mmWave unified access and backhaul network," in *Proc. PIMRC*, Hong Kong, Aug. 2015, pp. 2260–2264.

[8] C. Saha, M. Afshang, and H. S. Dhillon, "Bandwidth partitioning and downlink analysis in millimeter wave integrated access and backhaul for 5G," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8195–8210, Dec. 2018.

[9] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: Design aspects, challenges, and future directions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 22–28, Sep. 2016.

[10] P. S. Rao, K. Yedukondalu, and R. Ganesh, "FPGA implementation of digital 3-D image skeletonization algorithm for shape matching applications," *Int. J. Electron.*, vol. 108, no. 8, pp. 1326–1339, Jan. 2021.

[11] P. S. Rao and K. Yedukondalu, "Hardware implementation of digital image skeletonization algorithm using FPGA for computer vision applications," *J. Vis. Commun. Image Represent.*, vol. 59, pp. 140–149, Feb. 2019.

[12] A. Liu and V. Lau, "Cache-induced opportunistic MIMO cooperation: A new paradigm for future wireless content access networks," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2014, pp. 46–50.

[13] M. M. Sande, M. C. Hlophe, and B. T. Maharaj, "Instantaneous load-based user association in multi-hop IAB networks using reinforcement learning," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.

[14] T. Zhang, Y. Wang, W. Yi, Y. Liu, C. Feng, and A. Nallanathan, "Two time-scale caching placement and user association in dynamic cellular networks," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2561–2574, Apr. 2022.

[15] A. Mesodiakaki, E. Zola, R. Santos, and A. Kassler, "Optimal user association, backhaul routing and switching off in 5G heterogeneous networks with mesh millimeter wave backhaul links," *Ad Hoc Netw.*, vol. 78, pp. 99–114, Sep. 2018.

[16] A. Mesodiakaki, E. Zola, and A. Kassler, "User association in 5G heterogeneous networks with mesh millimeter wave backhaul links," in *Proc. IEEE 18th Int. Symp. World Wireless, Mobile Multimedia Netw.*, Jun. 2017, pp. 1–6.

[17] M. Feng and S. Mao, "Dealing with limited backhaul capacity in millimeter-wave systems: A deep reinforcement learning approach," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 50–55, Mar. 2019.

[18] P. V. Klaine, M. Jaber, R. D. Souza, and M. A. Imran, "Backhaul aware user-specific cell association using Q-learning," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3528–3541, Jul. 2019.

[19] H. Ma, H. Zhang, X. Wang, and J. Cheng, "Backhaul-aware user association and resource allocation for massive MIMO-enabled HetNets," *IEEE Commun. Lett.*, vol. 21, no. 12, pp. 2710–2713, Dec. 2017.

[20] Y. L. Lee, T. C. Chuah, A. A. El-Saleh, and J. Loo, "User association for backhaul load balancing with quality of service provisioning for heterogeneous networks," *IEEE Commun. Lett.*, vol. 22, no. 11, pp. 2338–2341, Nov. 2018.

[21] Y. Fadlallah, A. M. Tulino, D. Barone, G. Vettigli, J. Llorca, and J.-M. Gorce, "Coding for caching in 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 106–113, Feb. 2017.

[22] W. Mao, M. Narasimha, M. Simsek, and H. Nikopour, "Network coding for integrated access and backhaul wireless networks," in *Proc. 29th Wireless Opt. Commun. Conf. (WOCC)*, May 2020, pp. 1–6.

[23] J. Carreira, S. M. M. de Faria, L. M. N. Tavora, A. Navarro, and P. A. A. Assuncao, "Scalable coding of 360-degree video for streaming adaptation at 5G network edges," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Oct. 2020, pp. 1–5.

[24] Y.-H. Chiang and W. Liao, "ENCORE: An energy-aware multicell cooperation in heterogeneous networks with content caching," in *Proc. IEEE INFOCOM*, San Francisco, CA, USA, Apr. 2016, pp. 1–9.

[25] R. Torre, I. Leyva-Mayorga, S. Pandi, H. Salah, G. T. Nguyen, and F. H. P. Fitzek, "Implementation of network-coded cooperation for energy efficient content distribution in 5G mobile small cells," *IEEE Access*, vol. 8, pp. 185964–185980, 2020.

[26] M. K. Somesula, R. R. Rout, and D. V. L. N. Somayajulu, "Deadline-aware caching using echo state network integrated fuzzy logic for mobile edge networks," *Wireless Netw.*, vol. 27, no. 4, pp. 2409–2429, May 2021.

[27] R. Atiqur, A. Liton, and G. Wu, "Content caching strategy at small base station in 5G networks with mobile edge computing," *Int. J. Sci. Bus., IJSAB Int.*, vol. 4, no. 4, pp. 104–112, 2020.
[28] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Cache-aware user association in backhaul-constrained small cell networks," in *Proc. WiOpt*, Hammamet, Tunisia, May 2014, pp. 37–42.
[29] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.
[30] Y. Wang, X. Tao, X. Zhang, and G. Mao, "Joint caching placement and user association for minimizing user download delay," *IEEE Access*, vol. 4, pp. 8625–8633, 2016.
[31] L. E. Chatzieleftheriou, G. Darzanos, M. Karaliopoulos, and I. Koutsopoulos, "Joint user association, content caching and recommendations in wireless edge networks," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 46, no. 3, pp. 12–17, Jan. 2019.
[32] Y. Li, H. Ma, L. Wang, S. Mao, and G. Wang, "Optimized content caching and user association for edge computing in densely deployed heterogeneous networks," *IEEE Trans. Mobile Comput.*, vol. 21, no. 6, pp. 2130–2142, Jun. 2022.

L. MANJUNATH received the M.Tech. degree in computer network engineering from Visveswaraiah Technological University, Belgaum, Karnataka, in 2009. He is currently a Research Scholar with the Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation (Deemed to be University), Guntur, Andhra Pradesh, India. He is also a Faculty with the Department of Electronics and Communication Engineering, CVR College of Engineering, Vastunagar, Mangalpalli, Hyderabad. His research interests include interference management and backhauling techniques for mmWave 5G hetnets, ultra dense hetnets, and wireless sensor networks.

N. PRABAKARAN received the M.E. degree in applied electronics from Sathyabama Deemed University, Chennai, in 2007, and the Ph.D. degree from the Faculty of Engineering, Sathyabama Deemed University, in January 2015. He was with Sathyabama Deemed University, from June 2006 to July 2017, and the School of Electrical and Electronics Engineering, Department of Electronics and Telecommunications Engineering. He is currently an Associate Professor of electronics and communication engineering and a Skilling Professor with the Koneru Lakshmaiah Educational Foundation (Deemed to be University), Guntur, Andhra Pradesh. His research interests include wireless communications networks and the IoT systems.

S. V. ASWIN KUMER received the degree in electronics and communication engineering from the Pallavan College of Engineering, Kanchipuram, in April 2008, the master's degree from Embedded System Technology SRM University, Kanchipuram, in May 2012, and the Ph.D. degree in the implementation of image fusion using artificial neural network from SCSVMV (Deemed to be University), Enathur, in February 2019.

He is currently an Associate Professor with the Department of Electronics and Communication Engineering, KLEF (Deemed to be University), Guntur. He has more than 14 years of teaching experience. His research interests include digital communication and digital signal processing.

E. MOHAN received the M.E. degree in computer science engineering from Satyabhama University, the M.B.A. degree from Madras University, and the Ph.D. degree in computer science and engineering from Vinayaka Missions University. He has more than two decades experience in academic field. He is currently a Professor with the Saveetha School of Engineering, SIMATS, Chennai, Tamil Nadu, India. He titled three books and four scholars completed their Ph.D. under his guidance. Throughout his career, he is having good academic records of accomplishment and published many refereed journals in the reputed publications. His research interests include image processing, WSN, the IoT, ML, and datamining.

BALAJI NATARAJAN received the Ph.D. degree (full-time) in computer science and engineering from Pondicherry University, Pondicherry, India, in 2017. He is currently a Professor and the Head of the Department of Computer Science and Engineering, Sri Venkateshwaraa College of Engineering and Technology, Ariyur, Pondicherry. He has 15 years of teaching, research, and industry experience. He has published more than 50 research papers in various reputed international journals and conferences. His research interests include web services, service oriented architecture, evolutionary algorithms, artificial intelligence, and machine learning.

G. SAMBASIVAM (Member, IEEE) received the Ph.D. degree in computer science and engineering from Pondicherry University, Pondicherry, India. He was the Dean of the School of Information and Communication Technology, ISBAT University, Uganda. He is currently an Assistant Professor with the School of Computing and Data Science, Xiamen University Malaysia, Sepang, Malaysia. His research interests include artificial intelligence, machine learning, deep learning, graph neural networks, web service computing, and soft computing techniques.

VAIBHAV BHUSHAN TYAGI received the B.Tech. degree from UPTU, Lucknow, in 2007, the M.Tech. degree from IIT Roorkee, in 2011, and the Ph.D. degree, in 2015. He has more than 13 years of research and teaching experience around the globe. He is currently an Associate Professor (ECE) and the Dean of FICT, ISBAT University, Kampala, Uganda. He was worked in several administrative and academic positions in India, Ethiopia, and Uganda. His research interests include sensor applications in signal processing, signal modeling, artificial intelligence, and deep learning.