FACILITATING DIFFERENTIAL QoS ADAPTIVE TO USER, CONTENT AND NETWORK DYNAMICS IN mmWAVE BACK HAULING BASED 5G HETEROGENEOUS NETWORKS

L. Manjunath[1,2] and N. Prabakaran[3]
[1]Department of ECE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India
[2]Department of ECE, CVR College of Engineering, Hyderabad, Telangana, India
[3]Department of ECE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India
E-Mail: prabakaran@kluniversity.in

ABSTRACT

MmWave overlay 5G heterogeneous network (HetNet) proposed to boost the capacity of 5G networks has limited performance due to spectrum resource occupancy by backhaul links. Caching contents is a solution to solve the spectrum occupancy problem. Caching reduces the load on backhaul links and improves the utilization of access links. Most of the existing caching strategies use content popularity as the only factor and don't consider the priority of mobile nodes and need to provide differential QoS to users based on user characteristics and network dynamics. This work proposes a machine learning based strategy to manage the cache and backhaul resource allocation with the goal of providing differential QoS to user based on its priority and the current network dynamics at same time without degrading the average potential throughput. Cache management is adapted automatically to network dynamics, user characteristics and content characteristics using reinforcement learning.

Keywords: differential QoS, mm wave backhaul, hetnet, caching, fuzzy caching decision.

1. INTRODUCTION

Rapid proliferation of mobile devices and internet of things (IoT) devices are increasing mobile traffic exponentially bringing stress on mobile network resources and affecting communication bandwidth [1]. Millimeter wave (mmWave) based access and backhaul integration heterogeneous cellular networks (mABHetNets) has been envisioned in 5G dense cellular network to address the bandwidth crunch and supporting growing traffic demand ([2]-[4]). Low power small base stations (SBS) covering small cells are overload along with high power macro base station (BS) in mABHetNets. Mmwave backhaul links connect SBS to BS. Higher rate services are provided to users by both MBSs and SBSs through wireless access links. In addition MBSs maintain the backhaul capacity of SBSs through wireless backhaul link. Average potential throughput (APT) is an important performance metric in 5G dense cellular networks [5]. APT is highly dependent on the spectrum resources are shared between access and backhaul link as both use the same mmWave spectrum resource. The mmWave spectrum resources partitioning between access and backhaul links impacts the APT ([6]-[7]). Study in [8] found that almost 50% of mmWave spectrum is used in backhaul link and higher concentration of resources in backhaul restricts the APT. Caching contents at base station is a promising solution to this problem ([9] to [11]). Caching is based on the observation that few popular file account for most of the backhaul load. By proactively caching the popular contents at SBSs, the cached content can be delivered directly to user using the access link, thereby reducing the load on the backhaul links. Various caching strategies to reduce the load on backhaul and improve throughput links have been experimented ([12]-[15]). Most of current caching methods are based on content popularity and does not consider the network dynamics and user characteristics.

Without consideration for network dynamics and user characteristics, the QoS provided to users is same and this is not expected in an environment with users/devices of various priorities, mobility patterns and dynamic networks. This work proposes a solution to this problem by adapting the cache management and backhaul/access resource provisioning in such a way differential QoS is provided to the users at same time without degrading the APT. User characteristics in content access is modeled as machine learning based prediction problem. Based on the user content association prediction, network dynamics, user priorities a reinforcement learning based cache management is proposed. Following are the contributions of this work:

a) A novel machine learning based modeling of association between the user and the content.

b) A novel reinforcement learning based adaptive cache management with objective of differential QoS to user and maximization of APT. The reinforcement learning considers user characteristics, content characteristics, user dynamics in terms of mobility and network dynamics.

c) Scheduling of backhaul and access links based on the cache contents.

The paper is organized as follows. Section II presents the survey of existing cache management solutions and their research gaps. Section III presents the proposed solution and details of the novel contributions in this work. Section IV presents the results and comparison to state of art existing works. Section V presents the concluding remarks and scope for future work.

2. RELATED WORK

Chiang *et al* [10] proposed a content caching solution with the goal of reducing the backhaul energy consumption. The popular contents are cached in close proximity to access sites so that energy consumption for access is reduced. Authors proposed a greedy strategy where each transmission pointscache the contents greedily till its storage limit is reached. TP’s also cooperatives to reduce the inter cell interferences to reduce the energy consumption. The work focused only on reducing the energy consumption and did not consider the QoS and priority of users. Pantisano *et al* [11] proposed a novel cache aware approach to select the user equipment (UE) to be serviced based on content availability in cache and backhaul limitations. Small base station (SBS) calculates a preference based on content availability in terms of estimated time to deliver and send to the UE requesting content. If the UE accepts, matching is made and content is delivered to UE. If UE does not accept the preference provided by SBS, SBS searches for the next UE who can accept the preferences. The approach did not consider the priority of the user and danger of UE starvation. Tao *et al* [12] proposed a mixed integer non linear programming based caching solution with aim of reducing the sum of backhaul cost and transmit power cost. User requesting same content are grouped and BS with cached contents are allocated to groups based on the optimization of backhaul cost and transmit power cost. The solution cannot be used in realistic scenarios where the arrival time of same requests is diverse. Liu *et al* [13] analyzed the energy efficiency gain due to caching contents at base station. The authors indentified three important factors influencing the energy gain. Energy gain is modeled in terms of interference level and backhaul capacity. User QoS requirements were not considered in energy gain analysis. Gabry *et al* [14] proposed an optimization solution for content placement at edges with goal of minimization of backhaul rate and energy consumption. A convex optimization solution is proposed trading off between the backhaul rate and energy consumption. The contents are encoded using maximum-distance separable (MDS) codes. Due to MDS property, the UE can reconstruct the original data with minimal number of packets. Due to reduction in transmitted packets, energy consumption due to transmission is reduced. But the approach is not scalable as the cache size requirements shoots up. Llorca *et al* [15] used integer linear programming to decide the caching decisions based on global user request and network resources. The decision is made to maximize the energy gains. The work did not consider the network heterogeneity. Emara *et al* [16] proposed an optimization framework for file caching to maximize the hit probability. Authors inferred that ching strategies must be adapted to network parameters. Adaptation was done for multicast and unicast systems with Zipf distribution and popularity based caching respectively. But authors did not consider the user characteristics and network dynamics in caching decisions. Ahangary *et al* [17] proposed a hierarchical method for cache content placement. The method decides the content to cache based on its popularity and virality. Zipf distribution model is used for finding popularity of content. Popularity variation over time is calculated as virality. From popularity and virality, priority is calculated. Kabir *et al* [18] proposed a machine learning approach to analyze the behavior of the user and learn user profile. Based on the user profile, the contents to be cached is decided and placed at SBS. But the work did not consider the caching based on user profile from multi user perspective. ElBamby *et al* [19] proposed a joint user clustering and caching scheme for wireless small cell networks. Users are grouped into clusters based on social similarities. Each user group is associated with a SBS which caches the contents specific to that group. Though the solution performs better compared to random caching in terms of cache hit, the parameters like delay, energy are higher with increase in user mobility. Somesula *et al* [20] proposed a fuzzy logic based caching algorithm. The decision to cache content is made based on multiple factors like deadline, benefit and content request prediction. Content request prediction is made using echo state networks. Once the content is decided to be cached, the optimal place to cache content is solved as optimization problem using integer linear programming. Delay was the parameters considered for optimization. The work treated all the users at same priority without any differentiation for tariff differences between the users. Atiqur *et al* [21] proposed a caching algorithm to maximize the Quality of Experience (QoE) for the users. QoE was defined in terms of waiting time for video playbacks. The content to cache is decided based on satisfaction rate of the users. Peng *et al* [22] proposed an algorithm to increase the cache hit and reduce the content transmission delay. It is based on associating a value to cache content called file cache value. File cache value is calculated based on file size, popularity and time of requests. Lei *et al* [23] used deep learning for cache optimization. Deep learning is used find the SBS to user association mapping such that overall energy consumption for content delivery is reduced in Hetnet networks. Compared to other optimization algorithms like integer linear programming or convex optimization deep learning based optimization has lower computational complexity and able to provide 90% near optimal solution. But this work did not consider QoS optimization and user prioritization. Mohammed *et al* [24] used fuzzy logic decision making for caching the contents in SBS. File caching decision is made based on content popularity and user grouping. The fuzzy clustering decision proposed in this work is able to improve the cache hit but user prioritization and network dynamics are not considered for scheduling the traffic in SBS. Gupta *et al* [25] used collaborative filtering for caching decisions with the goal of increasing the cache hit ratio and reducing the delay in content delivery. The cache has two portions. One portion is for the popular contents. Another portion is for the content which is predicted to be accessed in future. Collaborative filtering is used to predict the contents which are going to accessed in future based on their similarity to recent popular contents. The approach viewed all users in same category and did not consider user prioritization in caching decisions. Mishra *et al* [26] proposed an algorithm called Rank-Directed Sparse Bayesian Learning (RD-SBL) for estimating the popularity of the content. Once the popularity of the content is determined, the top K contents for each user are selected and decided to be cached. But user prioritization was not considered in selection of top K contents. Chen *et al* [27] proposed a explicit caching technique applying heuristics algorithms. Caching is solved as joint problem of content recommendation, caching and delivery with goal of maximization of the user quality of experience. A file value for content is calculated over a period of time and content with higher file value is decided to be cached.

From the survey on existing works, most of the cache management solutions focused on optimization of individual parameters like energy, delay etc and maximizing the cache hit. Not much works considered user prioritization and providing differential QoS during access of contents by the users. Though some works explored proactive caching, they did not consider timing the proactive caching based on network dynamics. Backhaul resource allocation in combination with caching was not considered to maximize the performance gain. Most of the work were based on individual SBS and did not consider cooperation with cache at BS. The proposed solution in this work is framed to address these research gaps.

3. PROPOSED SOLUTION

The system model of the proposed solution is given in Figure-1. The Base station (BS) serves the requests of user equipment's (UE). In addition, small base station (SBS) deployed in coverage area of BS act as relay. Though coverage area of SBS is generally overlapping and one UE can also be served by multiple SBS, this work restricts a UE be served by one SBS. There is cache at both SBS and BS. SBS caches popular contents, so that it can served to users from cache avoiding the load on backhaul connecting SBS to BS whenever possible. SBS also forms a mesh to provide connectivity to BS for SBS out of coverage of BS. BS also caches contents which are popular across multiple SBS or moved out from SBS to BS due to cache size restrictions. BS allocates channels in timeslots to the SBS over backhaul links. SBS too allocated channels in timeslots the UE in its coverage area. The design requirements of the proposed solution over the system model are as follows:

a) UE's have priority levels depending on the user using it. UE with higher priority must be provided with higher QoS compared to UE with lower priority. At same time UE's with lower priority must not be starved.

b) Contents to be cached at SBS must be based on both current popularity and predicted popularity.

c) Bandwidth provided by BS to SBS must be adaptive to user dynamics and network dynamics to ensure differential QoS based on user priority.

The proposed solution for facilitating differential QoS is built on four strategies of: BS resource management, cache management, SBS resource management and geographic load balanced routing. Each of these strategies are detailed in below subsections.

A. BS Resource Management

BS allocates time slots to SBS for content download. The slots must be allocated to SBS in such a way to provide differential QoS for UE based on their priority but at same time, the low priority UE must not be starved. This work proposes a multi queue time slot scheme to accomplice differential QoS without starving the low priority UE.

There are K queues $\{Q_1, Q_2, \ldots. Q_K\}$ at BS corresponding to each of K priority level. The requests from the UE's are placed in their corresponding priority level. The weights are initially allocated to each of queue based on their priority level say $\{W_1, W_2, \ldots. W_K\}$ such that $W_1 > W_2 > \ldots. W_K$

In the total slot N, the slots are allocated to each queue based on their ratio of weight as

$$S(Q_k) = \frac{W_k \times N}{\sum_{i=1}^{K} W_i} \quad (1)$$

The weights are increased in proportion to number of requests in the queue as

$$\mathrm{W_i} = |\mathrm{Q_i}| \times (\mathrm{K} - \mathrm{i}) \quad (2)$$

By this way the time slots are split in proportion of weights to UE for content download.

B. Cache Management

Cache is managed at both BS and SBS. BS cache is used for managing overflow of cache at SBS. Fuzzy logic is applied to decide if a particular item needs to be cached: locally (cache at SBS), globally (cache at BS) or skip from caching. The decision to cache is made based on following factors.

Table-1. Caching decision factors.

Factors	Details
Popularity (A1)	Popularity in terms of access times
Predicted popularity (A2)	Predicted popularity of a content based on its correlation to past popular contents
Access frequency (A3)	Time interval between access of the content
Size of the content (A4)	File size of the content
Download delay (A5)	The time to download the content over backhaul from BS

The popularity of content (f) given as A1 is defined as:

$$A1\,(f) = \frac{p_c(f)}{\sum_{f=1}^{F} p_c(f)} \tag{3}$$

Where $p_c(f)$ is defined in term of content type C to which f belong as:

$$p_c(f) = \begin{cases} P_b(f)\,, & if\ f \in C \\ 0 & \end{cases} \tag{4}$$

Where

$$P_b(f) = \frac{1}{U}\sum_{u=1}^{U} p(u).\,p(C|u) \tag{5}$$

Where $p(u)$ is the probability user placed the request for content f and $p(c|u)$ is the probability that user placed request for content belonging to content type C.For new content which is downloaded for first time, the popularity is predicted in relation to other popular contents in the cache. Say there is an existing popular content X and a new content whose popularity is not known Y, then based on the meta content of X and Y, Latent Dirichlet Allocation (LDA) topic modeling approach proposed in [30] is applied to get the topic vector scores for each of X and Y. From the topic vector scores, the similarity between the content X and Y is calculated as

$$Sim(X,Y) = \frac{\sum_{j=1}^{N} TP(X)_i * TP(X)_j}{\sqrt{\sum_{j=1}^{N} TP(X)_j^{\,2}}\sqrt{\sum_{j=1}^{N} TP(X)_j^{\,2}}} \tag{6}$$

From the similarity score of each of related popular contents, predicted probability is calculated maximum value of popularity value of the related content

$$A2\,(Y) = max \prod_{\forall X} Sim(X,Y) \tag{7}$$

The access frequency (A3) is modeled in terms of exponential moving average of past delay in access to that content. Say delay to last accessed time is T, the access frequency at current time $A3_t$ is calculated as

$$A3_t = \alpha.T + (1-\alpha)A3_{t-1} \tag{8}$$

Download delay (A5) is calculated as the average of past download delay for the same content. It is given as:

$$A5_t = \frac{\sum_{i=1}^{b} A5_i}{n} \tag{9}$$

The uncertainty in selecting among three decisions of (i) caching locally at SBS (L), (ii) caching at BS (G) and (iii) skip caching (N) is handled using Fuzzy inference system [28]. The contents with higher popularity, higher predicted popularity must have maximum chance of caching. But at same time, if the access frequency of content is low, then content can be positioned at BS cache instead of SBS cache. The size of the content is selected an important factor in caching decision as the large file size must be given more priority then a small file size as the download time will be more. Download delay is also important factor impacting the decision of caching as the content can be downloaded instead of caching when the delay expected for content download by user is higher than download delay.

Differing from the traditional way of defining fuzzy rule base, this work adopts mining approach to map the input attributes (A1 to A5) to three caching decisions (L, G, N). A test run is conducted to collect the <A1, A2… A5> parameters from different UE/contents and a training dataset is created with each of the <A1, A2,…A5> rows labeled with one of three caching decision (L,G,N) by domain expert with goal of maximizing the cache hit and minimizing the delay.

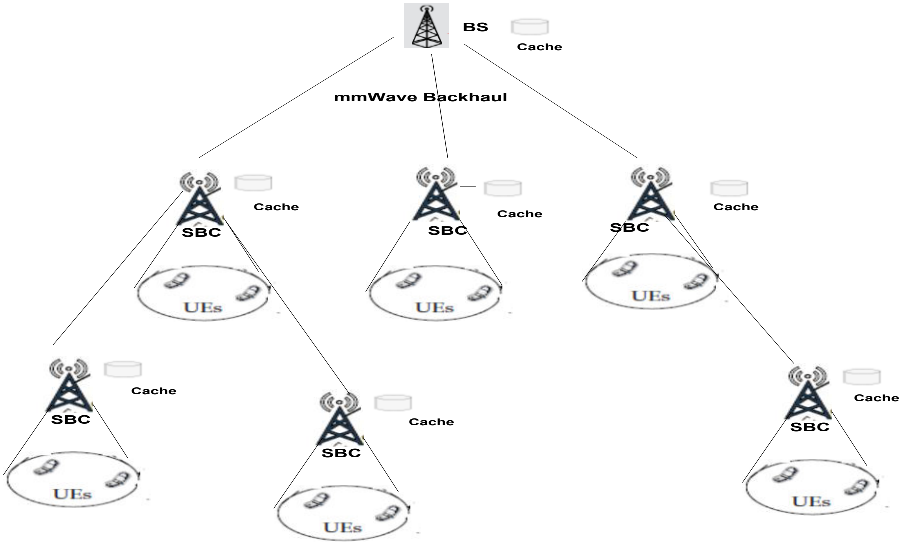

Figure-1. System model.

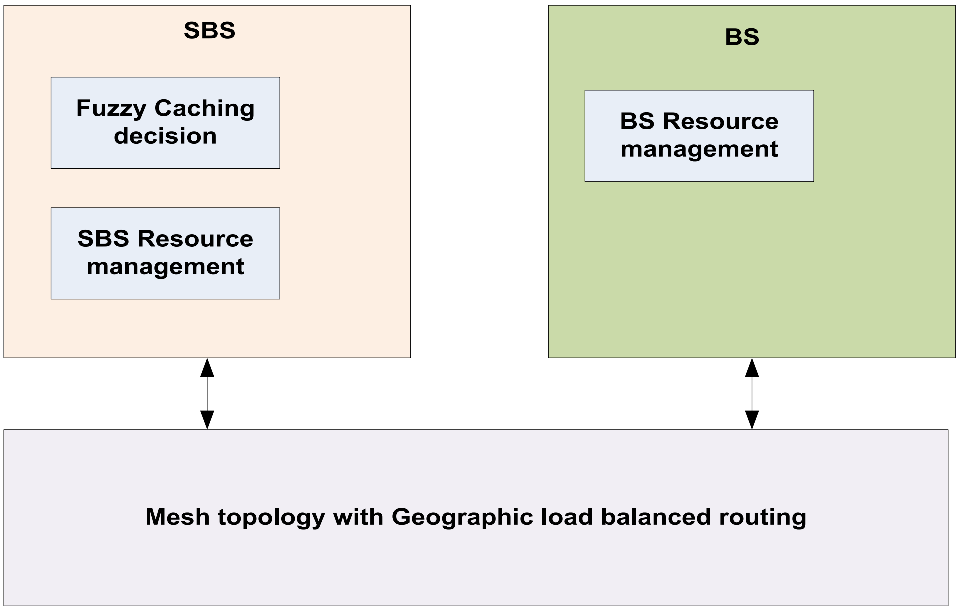

Figure-2. Proposed architecture.

Fuzzy C mean clustering algorithm is used for clustering the dataset to three clusters (L, G, N).
The cluster center is given as

$$\mathrm{D} = \{\, \mathrm{D}_{\mathrm{e,q}} \,, \mathrm{e} = 1{,}2 \ldots 5 \text{ and } \mathrm{q} = 1{,}2{,}3\} \qquad (10)$$

Where $D_{e,q}$ is the q^{th} feature of the e^{th}cluster. Gaussian function [29] is used to model the proximity of the q^{th} feature in r^{th} data as

$$\mathrm{G}\left(\mathrm{f}_{\mathrm{r,q}}, \mathrm{D}_{\mathrm{e,q}}, \sigma_{\mathrm{e,q}}\right) = \mathrm{e}^{\frac{(\mathrm{f}_{\mathrm{r,q}} - \mathrm{D}_{\mathrm{e,q}})^2}{\sigma_{\mathrm{e,p}}{}^2}} \qquad (11)$$

Where

$$\sigma_{\mathrm{e,q}} = \frac{1}{\mathrm{N_e}} \sum_{\mathrm{r}=1}^{\mathrm{N_e}} (\mathrm{f}_{\mathrm{r,q}} - \mathrm{D}_{\mathrm{e,q}})^2 \qquad (12)$$

Feature closeness is calculated as product of Gaussian function as

$$\Psi_{\mathrm{r,e}} = \prod_{\mathrm{q}=1}^{\mathrm{P}} \mathrm{G}\left(\mathrm{f}_{\mathrm{r,q}}, \mathrm{D}_{\mathrm{e,q}}, \sigma_{\mathrm{e,q}}\right) \qquad (13)$$

In terms of linear regression of weighted feature values, the output label of cluster is found as

$$\Phi_{\mathrm{r,e}} = \mathrm{W}_{\mathrm{e,0}} + \sum_{\mathrm{q}=1}^{\mathrm{P}} \mathrm{W}_{\mathrm{e,q,f_{r,q}}} \qquad (14)$$

With W representing the weights. The final cluster label is found as sum of weighted membership function. It is given as

$$\overline{\mathrm{N}}(\mathrm{r}) = \sum_{\mathrm{e}=1}^{\mathrm{P}} \Psi_{\mathrm{r,e}} \Phi_{\mathrm{r,e}} \qquad (15)$$

The linear regression is fit with suitable weights till the error is minimized. Error is calculated as

$$\mathrm{E} = \sum_{\mathrm{r}=1}^{\mathrm{N}} ||\overline{\mathrm{N}}(\mathrm{r}) - \mathrm{N}(\mathrm{r})||^2 \tag{16}$$

The Gaussian parameters $D_{e,q}, \sigma_{e,q}$ and the regression coefficients $W_{e,p}$ are tuned to reduce the error. Gradient decent method is used for fine tuning.

$$\mathrm{D}_{\mathrm{e,q}}(\mathrm{t}+1) = \mathrm{D}_{\mathrm{e,q}}(\mathrm{t}) + \eta_{\mathrm{C}} \frac{\partial \mathrm{E}}{\partial \mathrm{D}_{\mathrm{e,q}}} \tag{17}$$

$$\sigma_{\mathrm{e,q}}(\mathrm{t}+1) = \sigma_{\mathrm{e,q}}(\mathrm{t}) + \eta_{\sigma} \frac{\partial \mathrm{E}}{\partial \sigma_{\mathrm{e,q}}} \tag{18}$$

$$\mathrm{W}_{\mathrm{e,q}}(\mathrm{t}+1) = \mathrm{W}_{\mathrm{e,q}}(\mathrm{t}) + \eta_{\mathrm{W}} \frac{\partial \mathrm{E}}{\partial \mathrm{W}_{\mathrm{e,q}}} \tag{19}$$

Where $\eta_C, \eta_\sigma, \eta_W$ are the learning parameters in gradient descent method? For each class of L, G, N fuzzy Gaussian membership functions are found.

For every new content downloaded and periodically for content already in cache, the features A1 to A5 are extracted and Fuzzy Gaussian membership function is executed for all three classes of L,G and N. The decision is matched to the class which has highest value for the membership function.

The pseudo code for caching decision is given below:

Algorithm: Decide Caching
Input: Content x
Output: L, G, N

```
1. Extract features A1-A5 for content x as in Table-1.
2. Calculate $\Phi_{r,e}$ for each class of L, G, N as in Eq.14
3. res= Label of Max value of ($\Phi_{r,e}$) for L, G, N
4. return res
```

C. SBS Resource Management

SBS allocates time slot channels for UE and transmits contents in those time slots. In most cases, equal numbers of slots are allocated for all demanding users. This work adapts the slot scheduling to provide differential QoS for users based on their priority. The estimated traffic demand (FD) is calculated at start of every time frame for all the UE requesting for contents. It is calculated as

$$FD = \min(MA_i + D_i, T \times \Delta) \tag{20}$$

Where the exponential moving average of outgoing traffic to UE is given as MA_i and the traffic volume in buffer is given as $D_i$and the current transmission rate is given as T. $MA_i$is calculated as:

$$MA_i = \begin{cases} \alpha.T + (1-\alpha)MA_{i-1} \; if \; T \neq 0 \\ (1-\alpha)MA_{i-1}, otherwise \end{cases} \tag{21}$$

From the estimated traffic demand and priority of UE, the number of slots to be allocated to each UE is calculated as

$$ns_x = \frac{FD_x}{\sum_{x=1}^{n} FD_x} * \frac{P_x}{\sum_{x=1}^{n} P_x} \tag{22}$$

Where P is the priority of UE. By allocating slots to UE in proportion to their priority and also considering their past allocated traffic demand, differential QoS is provided to the users.

The pseudo code for slot allocation is given below:

Algorithm: AllocateSlot
Input: UE vector $<\mathrm{UE}_1, \mathrm{UE}_2, \ldots . \mathrm{UE}_n>$
Output: Slots for each UE

```
1. Totdemand=0;
1. for each iin UE vector
FD$_i$= Estimate traffic demand for UE I using Eq.21
FD += FD$_i$
2. for each I in UE vector
Slot[i] = calculate slot using Equation 22
3. return Slot[i]
```

D. Routing

SBS at one hop away from BS can reach directly through mmWave backhaul either in Line of Sight (LOS) or Non Line of Sight (NLOS) mechanism. But for SBS not in coverage of BS reach BS though multi hop manner. The path between the far off SBS to BS is established using geographic load balanced routing. For each SBS a load score is calculated in terms of percentage of spectrum occupancy.

Every periodically, SBS initiates the process of routing path establishment with BS. SBS sends a hello message broadcast to nearby SBS in its communication range. The SBS receiving the hello message broadcast replies with hello response containing its current load score. The SBS receiving the hello response select K SBS with shortest distance to BS. From the response of K selected SBS, the one with least value of load score is selected as the next hop for routing. This process is repeated at each SBS till the path to BS is established.

4. RESULTS

The performance of the proposed solution is simulated with following simulation configuration.

Table-2. Simulation configuration.

Parameter	Value
No of UE	50 to 250
No of SBS	10
Simulation area	3000m × 3000m
Simulation time	20 minutes
Content request rate	1 request per 10 sec
Content size	500 KB
Cache size	5000 to 10,000 KB
Energy consumption	0.5×10^{-8} joules/bit
SBS coverage area radius	200 m
BS coverage area radius	1500 m
Location of BS	Center of area
User priority level	1 to 3, 1 being highest

The performance is measured in terms of following metrics: packet delivery ratio, delay, throughput, energy consumption, cache hit ratio and acceleration ratio. Cache hit ratio is measured as the probability of finding the content in cache out of total number of content requests. Acceleration ratio is measured as ratio of saved delay to original delay (downloaded through BS). The performance of the proposed solution is compared against backhaul optimized Hetnetproposed by Li *et al* [31], dynamic meshed backhaul proposed by Tran *et al* [32] and fuzzy logic based mobile edge network proposed by Somesula *et al* [20].

The packet delivery ratio is measured by varying the number of UE's and the result is given in Table-3.

Table-3. Packet delivery ratio.

No of UE's	Proposed	Li *et al* [31]	Tran *et al* [32]	Somesula *et al* [20]
50	93	87	82	81
100	92	86	81.5	79.5
150	91	85.2	79.12	78
200	90	84	78	76.9
250	89	83.2	76.8	75.1
Average	91	85.08	79.48	78.1

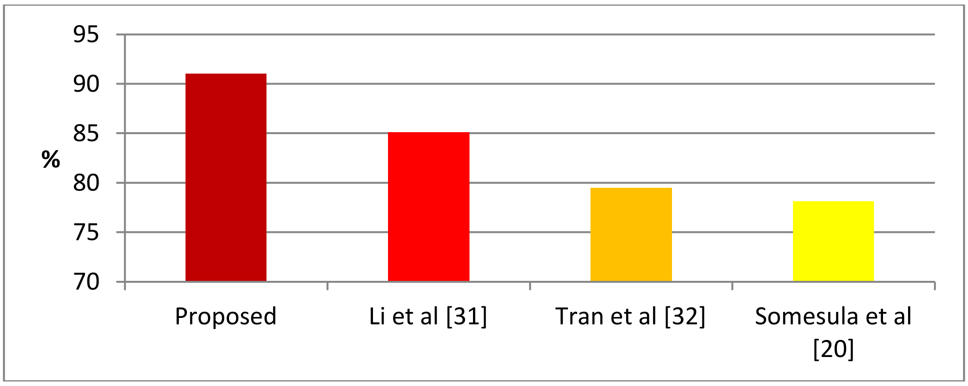

Figure-3. Comparison of packet delivery ratio.

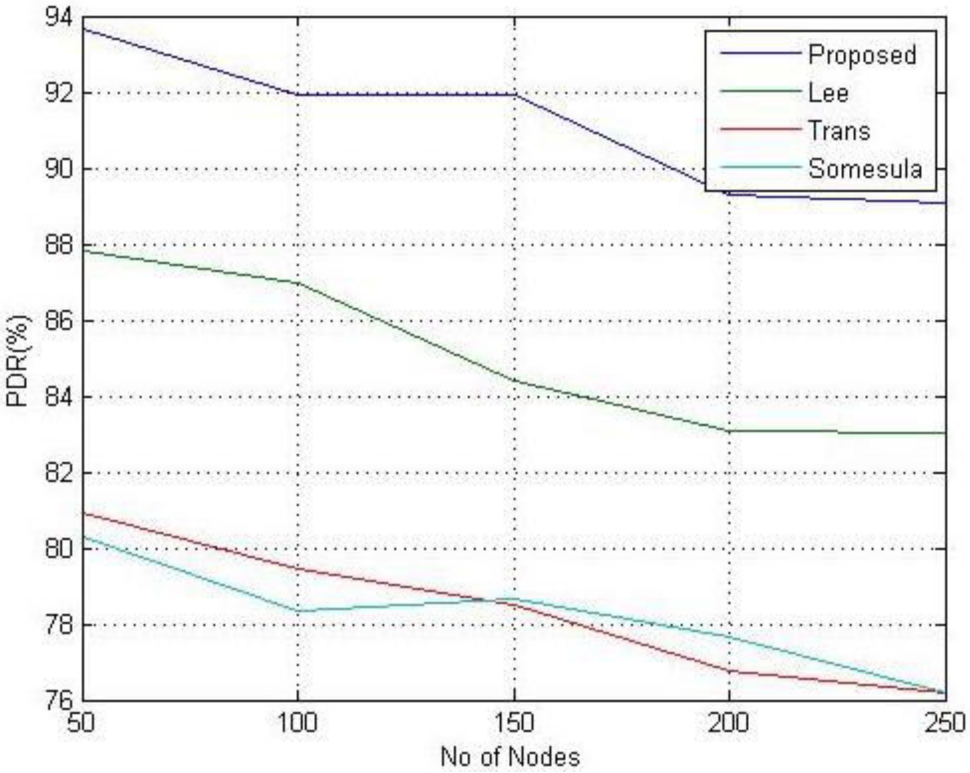

Figure-4. Packet delivery ratio VS No. of nodes (UE).

The packet delivery ratio is on average 6% more compared to Li *et al,* 11% more compared to Tran et al and 12% more compared to Somesula *et al*. The packet delivery ratio has increased in proposed solution due to use of geographic load balanced routing to create link between SBS. This avoids congestion proactively and hence the packet delivery ratio increased.

The results for average packet delay for various number of UE is given in Table-4.

Table-4. Comparison of delay.

No of UE's (Nodes)	Delay (milli sec)			
	Proposed	Li *et al* [31]	Tran *et al* [32]	Somesula *et al* [20]
50	15	18	21	24
100	14	16	18	22
150	13	15	17	21
200	12	15	16	20
250	12	14	15	18
Average	13.2	15.6	17.4	21

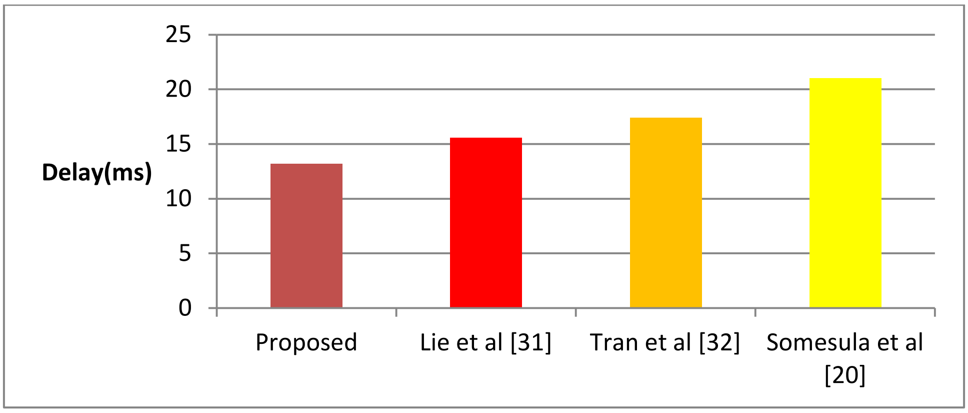

Figure-5. Comparison of average delay.

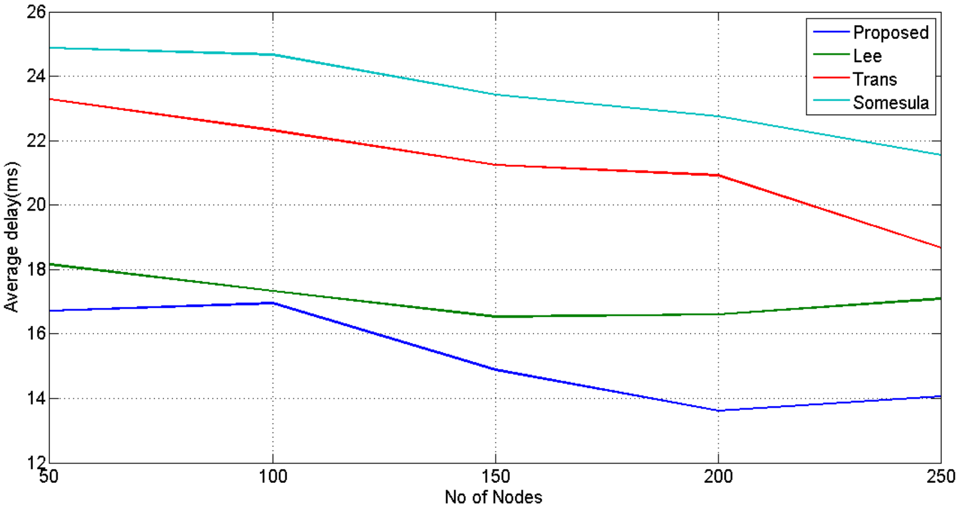

Figure-6. Average delay VS No. of Nodes (UE).

The average delay is 18.18% lower compared to Li *et al,* 31.81% lower compared to Tran et al and 59% lower compared to Somesul *et al*. The delay has reduced in proposed solution due to two important factors, reduced distance to BS due to geographic routing and increased cache hit due to fuzzy caching decision.

By varying the number of UE, the average throughput at SBS and the result is given in Table-5.

Table-5. Comparison of throughput.

No of UE's (Nodes)	Proposed	Li *et al* [31]	Tran *et al* [32]	Somesula *et al* [20]
50	92	84	83	81
100	140	136	129	125
150	181	173	167	160
200	224	208	201	195
250	267	251	242	238
Average	180.8	170.4	164.4	159.9

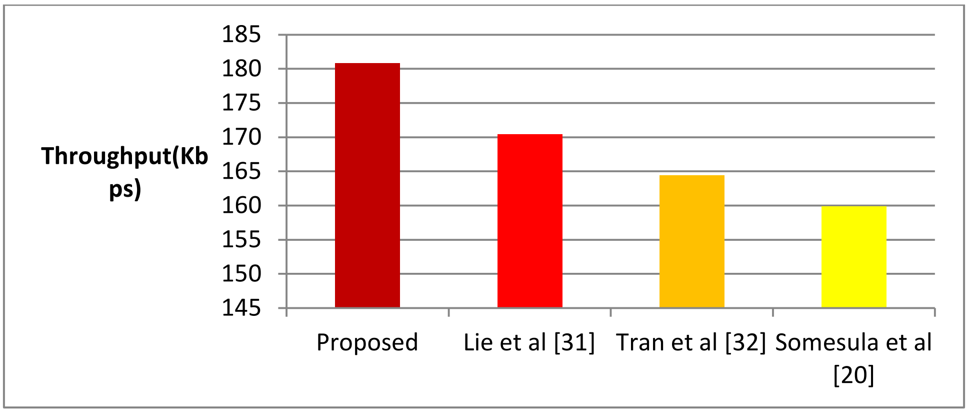

Figure-7. Comparison of average throughput.

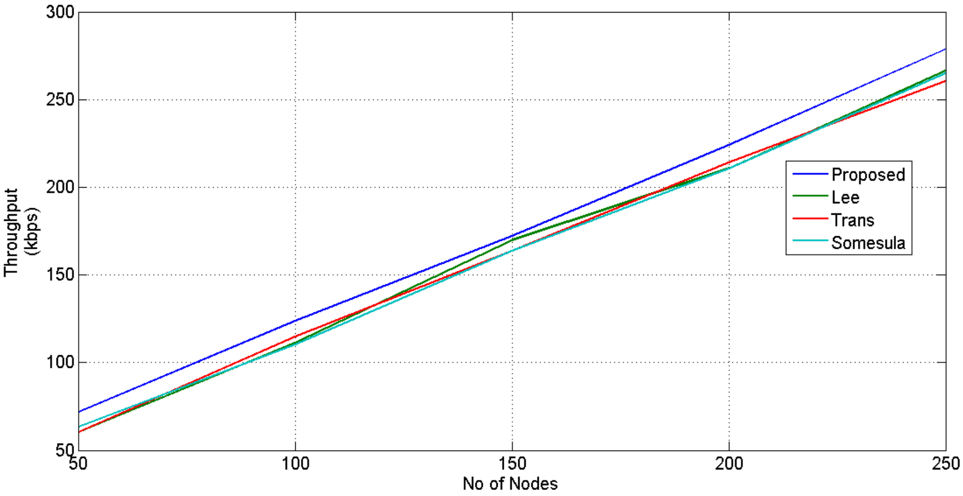

Figure-8. Throughput vs no. of nodes (UE).

The average throughput in proposed solution is 5% more compared to Li *et al*, 9% more compared to Tran *et al* and 11.5% more compared to Somesula *et al*. The throughput has increased in proposed solution due to efficient use of resources at BS and SBS.

The energy consumption is measured by varying UE's and the result is given in Table-6.

Table-6. Comparison of energy consumption.

No of UE's	Proposed	Li *et al* [31]	Tran *et al* [32]	Somesula *et al* [20]
50	72	80	82	85
100	85	92	93	97
150	89	97	100	106
200	94	104	108	116
250	97	116	118	127
Average	87.4	97.8	100.2	106.2

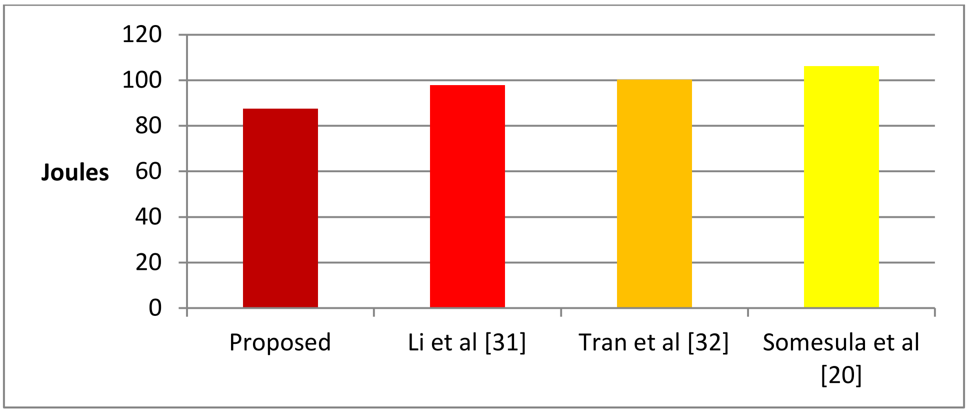

Figure-9. Comparison of energy consumption.

The energy consumption in proposed solution is on average 11.8% lower compared to Li *et al*, 14.64% lower compared to Tran *et al* and 21.51% lower compared to Somesula *et al*. The energy consumption reduction in proposed solution is due to reduced load on backhaul, use of multi-level cache with fuzzy caching decision for higher cache hit ratio.

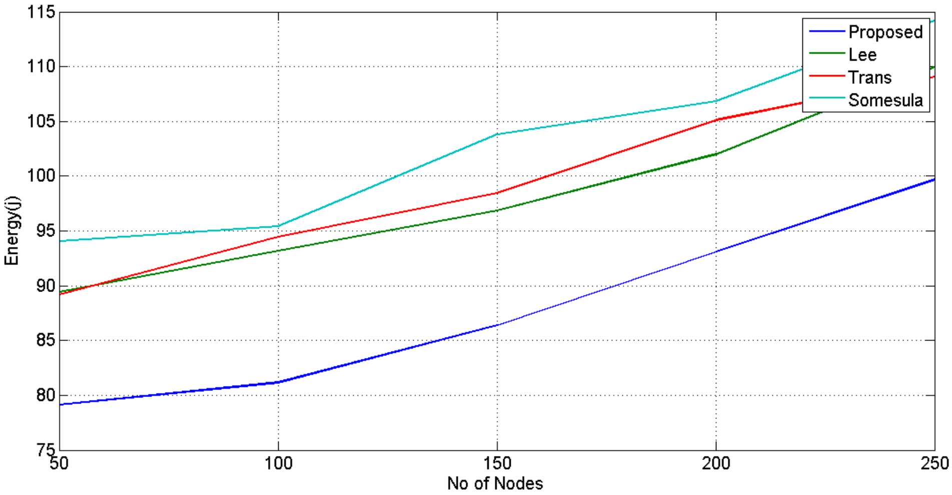

Figure-10. Energy consumption vs no. of nodes (UE).

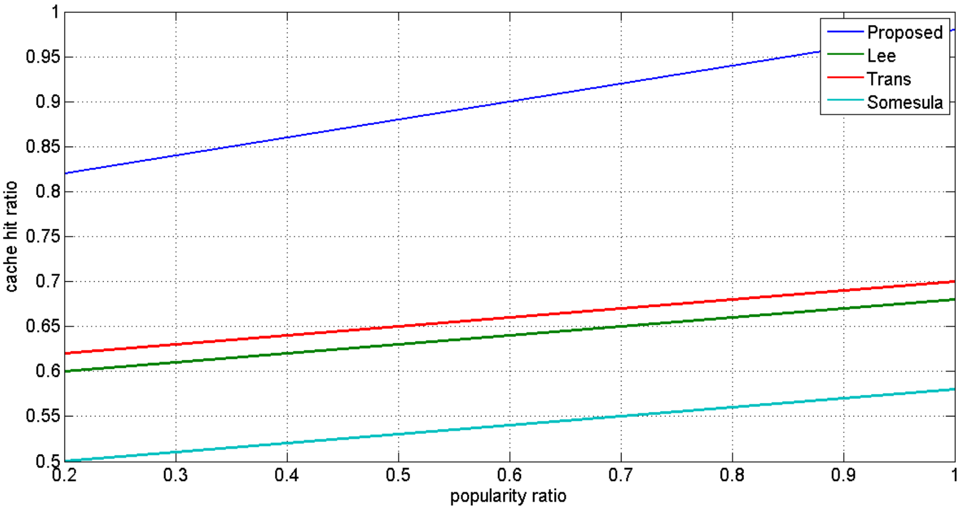

Figure-11. Cache hit ratio vs ratio of popular contents.

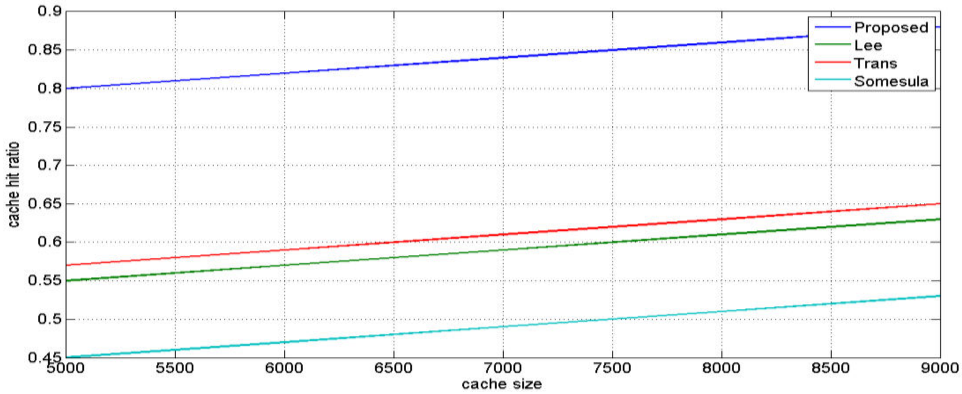

Figure-12. Cache hit ratio vs cache size.

The cache hit ratio is measured by varying the cache size from 5000 to 10,000 KB and the result is given in Figures 11 & 12. The cache hit ratio is higher in proposed solution due to multi criteria-based decision using fuzzy logic in proposed solution. The cache hit ratio is measured by varying the popularity ratio of total contents and the results are given in Figure-12. The hit increases with increase in the popularity ratio and is higher in proposed solution.

5. CONCLUSION AND FUTURE SCOPE

This work integrated four strategies of BS resource management, cache management; SBS resource management and geographic load balanced routing to provide differential QoS in mmWave backhauling based 5G heterogeneous networks. The proposed solution is adaptive to user and network dynamics. Using fuzzy clustering decision, the proposed solution increased a cache hit ratio and reduced the overall delay for content download. Through use of resource scheduling based on user prioritization, differential QoS was provided to users. The proposed solution increased the packet delivery ration by 6%, reduced the delay by 18% and reduced the energy consumption by 11% compared to existing works. The work can be extended with cache revocation and effective content coding schemes to reduce the network overhead.

REFERENCES

[1] C. Madapatha *et al*. 2020. On Integrated Access and Backhaul Networks: Current Status and Potentials. IEEE Open J. Commun. Soc. 1: 1374-1389.

[2] C. Dehos, J. L. Gonza´lez, A. D. Domenico, D. Kte´nas and L. Dussopt. 2014. Millimeter-wave Access and Backhauling: The Solution to the Exponential Data Traffic Increase in 5G Mobile Communications Systems? IEEE Commun. Mag. 52(9): 88-95.

[3] Polese, Michele, *et al*. 2019. Integrated Access and Backhaul in 5G mmWave Networks: Potentials and Challenges. arXiv preprint arXiv:1906.01099.

[4] NR. 2017. Study on Integrated Access and Backhaul, document 3GPP TR 38.874.

[5] A. Al Ammouri, J. G. Andrews and F. Baccelli. 2019. A Unified Asymptotic Analysis of Area Spectral Efficiency in Ultradense Cellular Networks. IEEE Trans. Inf. Theory. 65(2): 1236-1248.

[6] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas and A. Ghosh. 2013. Millimeter Wave Beamforming for Wireless Backhaul and Access in Small Cell Networks. IEEE Trans. Commun. 61(10): 4391-4403.

[7] Z. Shi, Y. Wang, L. Huang and T. Wang. 2015. Dynamic Resource Allocation in MmWave Unified Access and Backhaul Network. in Proc. PIMRC, Hong Kong. pp. 2260-2264.

[8] C. Saha, M. Afshang and H. S. Dhillon. 2018. Bandwidth Partitioning and Downlink Analysis in Millimeter Wave Integrated Access and Backhaul for 5G. IEEE Trans. Wireless Commun. 17(12): 8195-8210.

[9] D. Liu, B. Chen, C. Yang and A. F. Molisch. 2016. Caching at the Wireless Edge: Design Aspects, Challenges, and Future Directions. IEEE Commun. Mag. 54(9): 22-28.

[10] Y. Chiang and W. Liao. 216. ENCORE: An Energy-aware Multicell Cooperation in Heterogeneous Networks with Content Caching. In Proc. IEEE INFOCOM, San Francisco, CA. pp. 1-9.

[11] F. Pantisano, M. Bennis, W. Saad and M. Debbah. 2014. Cache-aware User Association in Backhaul-constrained Small Cell Networks. In Proc. WiOpt, Hammamet. pp. 37-42.

[12] M. Tao, E. Chen, H. Zhou and W. Yu. 2016. Content-centric Sparse Multicast Beamforming for Cache-enabled Cloud RAN. IEEE Trans. Wireless Commun. 15(9): 6118-6131.

[13] D. Liu and C. Yang. 2016. Energy Efficiency of Downlink Networks with Caching at Base Stations. IEEE J. Sel. Areas Communication. 34(4): 907-922.

[14] F. Gabry, V. Bioglio and I. Land. 2016. On Energy-Efficient Edge Caching in Heterogeneous Networks. IEEE J. Sel. Areas Communication. 34(12): 3288-3298.

[15] J. Llorca *et al*. 2013. Dynamic In-network caching for Energy Efficient Content Delivery. In Proc. IEEE INFOCOM. pp. 245-249.

[16] M. Emara, H. Elsawy, S. Sorour, S. Al-Ghadhban, M. Alouini and T. Y. Al-Naffouri. 2018. Optimal Caching in 5G Networks with Opportunistic Spectrum Access. in IEEE Transactions on Wireless Communications. 17(7): 4447-4461.

[17] S. Ahangary, H. Chitsaz, M. J. Sobouti, A. H. Mohajerzadeh, M. H. Yaghmaee and H. Ahmadi. 2020. Reactive caching of viral content in 5G networks. 2020 3rd International Conference on Advanced Communication Technologies and Networking (CommNet).

[18] Kabir Asif, Shahid Iqbal Muhammad, Jaffri Zain Ul Abidin, Rathore Shoujat, Kitindi Edvin and Chughtai Gohar Rehman. 2018. User Aware Edge Caching in 5G Wireless Networks. International Journal of Computer Network and Information Security.

[19] M.S. El Bamby, M. Bennis, W. Saad, M. 2014. Latva-AhoContent-aware user clustering and caching in wireless small cell networks2014 11th International Symposium on Wireless Communications Systems (ISWCS), IEE. pp. 945-949.

[20] Manoj Kumar Somesula, Rashmi Ranjan Rout and D. V. L. N. Somayajulu. 2021. Deadline-aware caching using echo state network integrated fuzzy logic for mobile edge networks. Wirel. Netw. 27, 4(May 2021): 2409-2429.

[21] Rahman Atiqur& Ali Md Liton&Guangfu Wu, 2020. "Content Caching Strategy at Small Base Station in 5G Networks with Mobile Edge Computing. International Journal of Science and Business, IJSAB International. 4(4): 104-112.

[22] Peng T., Wang H., Liang C., *et al*. 2021. Value-aware cache replacement in edge networks for Internet of Things. Trans Emerging Tel Tech. 32: e4261

[23] Lei L., You L., Dai G., Vu T. X., Yuan D., Chatzinotas S. 2017. A deep learning approach for optimizing content delivering in cache-enabled HetNet. In: IEEE. 2017: 449453.

[24] Mohammed L., Jaseemuddin M., Anpalagan A. 2018. Fuzzy soft-set based approach for femtocaching in wireless networks. In: IEEE.

[25] Gupta Divya, Rani Shalli, Ahmed Syed Hassan, Verma, Sahil, Ijaz Muhammad Fazal and Shafi Jana. 2021. Edge Caching Based on Collaborative Filtering for Heterogeneous ICN-IoT Applications. Sensors. 21. 10.3390/s21165491.

[26] Mishra S. K., Pandey P., Arya P., Jain A. 2018. Efficient proactive caching in storage constrained 5G small cells. In: IEEE. 291296.

[27] Chen L., Su Y., Luo W., Hong X. & Shi J. 2018. Explicit Content Caching at Mobile Edge Networks with Cross-Layer Sensing. Sensors (Basel, Switzerland). 18(4): 940.

[28] Iancu I. 2012. A mamdani type fuzzy logic controller. In Fuzzy logic-controls, concepts, theories and applications (InTech).

[29] Sitansu Kumar Das, Sanjoy Kumar Saha, Dipti Prasad Mukherjee Multiple Objects Segmentation with Fuzzy Rule-Base Trained Topology Adaptive Active Membrane ICVGIP 10, December 12-15, 2010, Chennai, India Copyright 2010 ACM 978-1-4503-0060-5/10/12.

[30] Ji, S., Li, X., Huang, Z. *et al*. 2021. Suicidal ideation and mental disorder detection with attentive relation networks. Neural Comput & Applic.

[31] Li Yilin, Luo Jian, Stirling-Gallacher Richard and Caire Giuseppe. 2019. Integrated Access and Backhaul Optimization for Millimeter Wave Heterogeneous Networks.

[32] Tran Gia Khanh, Santos Ricardo, Ogawa Hiroaki, Nakamura Makoto, Sakaguchi Kei and Kassler Andreas. 2018. Context-Based Dynamic Meshed Backhaul Construction for 5G Heterogeneous Networks. Journal of Sensor and Actuator Networks. 7. 43. 10.3390/jsan7040043.