Clustered and Distributed Caching Methods for F-RAN-Based mmWave Communications

Adel Aldalbahi 1,*, Mohammed A. Jasim 2, Nazli Siasi 3, Mounir Bouzguenda 1, Hesham Enshasy 1 and Rajamohamed Sumsudeen 1

1 Department of Electrical Engineering, King Faisal University, Al Ahsa 34221, Saudi Arabia; mbuzganda@kfu.edu.sa (M.B.); henshasy@kfu.edu.sa (H.E.); rsumsudeen@kfu.edu.sa (R.S.)
2 School of Engineering and Technology, University of Washington, Tacoma, WA 98402, USA; jasim@uw.edu
3 College of Computing and Digital Media, DePaul University, Chicago, IL 60614, USA; nsiasi@depaul.edu

* Correspondence: aaldalbahi@kfu.edu.sa; Tel.: +966-054-3388552

Abstract: Fog-radio access networks (F-RANs) alleviate fronthaul delays for cellular networks as compared to their cloud counterparts. This allows them to be suitable solutions for networks that demand low propagation delays. Namely, they are suitable for millimeter wave (mmWave) operations that suffer from short propagation distances and possess a poor scattering environment (low channel ranks). The F-RAN here is comprised of fog nodes that are collocated with radio remote heads (RRHs) to provide local processing capabilities for mobile station (MS) terminals. These terminals demand various network functions (NFs) that correspond to different service requests. Now, provisioning these NFs on the fog nodes also yields service delays due to the requirement for service migration from the cloud, i.e., offloading to the fog nodes. One solution to reduce this service delay is to provide cached copies of popular NFs in advance. Hence, it is critical to study function popularity and allow for content caching at the F-RAN. This is further a necessity given the limited resources at the fog nodes, thus requiring efficient resource management to enhance network capacity at reduced power and cost penalty. This paper proposes novel solutions that allocate popular NFs on the fog nodes to accelerate services for the terminals, namely, the clustered and distributed caching methods. The two methods are analyzed and compared against the baseline uncached provisioning schemes in terms of service delay, energy consumption, and cost.

Keywords: clustered caching; distributed caching; fog computing; millimeter wave; network function

1. Introduction

Millimeter wave (mmWave) bands support wide bandwidth transmission without the need for sophisticated channelization techniques such as multi-carrier and career aggregation. This is attributed to the wide contiguous bandwidth chunks at these bands. This has allowed for mmWave to be considered in the New Radio (NR) standard of 5G standard, in the Frequency Range (FR) 2. A key limitation in mmWave bands here is the high path and penetration losses, along with atmospheric attenuation, oxygen absorption, and sensitivity to blockage. This makes mmWave links susceptible to significant degradation in signal quality.

The long separation distances between mobile station (MS) terminals and access points (radio remote heads) add complexity to the design of beamforming and access networks. For beamforming, new designs are required that deem low power consumption and multi-analog beamformers at the MS terminals and digital or hybrid beamformers at the BS for spatial multiplexing and multi-user connectivity. For access networks, it is essential to reduce fronthaul traffic to minimize end-to-end delays for the geographically distributed terminals. Solutions here include cloud-radio access networks (C-RANs) that allow for a centralized baseband unit (BBU) pool that processes various network functions (NFs). These NFs are migrated from the radio remote heads (RRHs) that feature limited resources to the...
C-RAN through fronthaul links, i.e., to benefit from the abundant resources at the cloud core. The C-RAN (BBU) here facilitates centralized scheduling and allocation at the RRHs. One limitation to this centralization structure is the need to collect information from the distributed RRHs, which increases signaling over the fronthaul links. This in turn prolongs the propagation delay for service requests, in addition to an increase in power consumption and cost for the network.

An alternative structure is the deployment of F-RAN at the edge of the network, which is more suitable for mmWave operations, as it reduces the separation distances between MS terminals and RRHs. The closer proximity to the MS terminals also reduces power consumption, which promotes the demands power efficiency for mmWave communications. Therefore, it is vital to reduce the fronthaul traffic and deploy NFs at the edge to provide low latency services. One solution here is caching the popular NFs from incoming traffic at fog nodes that are collocated with the RRH. This forms an F-RAN that offers existing copies of the NFs at less processing and storage demands. This eliminates online offloading between the cloud and fog, and instead allows for offline prefetching at the caching content with the popular NFs. As a result, network operators benefit from a reduction in the transmitted power and the fronthaul cost. Furthermore, caching at the fog nodes promotes network densification of mmWave small cells, given the efficient resource utilization that allows for higher network capacity. Overall, the goal is to enable fog nodes with new functionalities, such as virtualization and efficient caching for robust mmWave backhaul networks, that allow for flexible and agile operations.

There are only a few studies on caching in mmWave F-RAN that focus on optimization and federated and reinforcement learning to predict content popularity in a time-varying fashion. These studies lack a comprehensive analysis in terms of the network operation, such as cost and energy consumption. Further, there is a pressing need to consider the limited resources offered at the fog node for users that demand ultra-low latency levels. Furthermore, the 5G MiEdge project in [1] addresses the need of combining mobile edge computing (MEC) and mmWave systems, thus facilitating a system termed as mmWave edge cloud (MiEdge). The project aims to develop a 5G ecosystem that combines MiEdge, liquid RAN C-plane, and user/application-centric orchestration. This is based on the development of a new cellular network control plane in which user’s context information is collected and processed to forecast traffic requests, thus enabling a proactive resource allocation. This paper aligns with the objectives of the project in [1] by combining mmWave communications with edge/fog computing to facilitate 5G ecosystem.

Along these lines, this paper proposes a novel resource allocation and node placement in mmWave F-RAN for content caching of popular NFs at the edge of the fog nodes. The goal is to allocate the best fog node that allows for the largest content caching, along with reduced delay and cost. The work includes complete modeling and analysis that accounts for the physical and network layers. This includes the network model, user beamforming model for mmWave transmission, request model, content popularity, and cache model.

This paper is organized as follows. Prior work on content caching in mmWave F-RAN is presented in Section 2. The network architecture and model are then introduced in Section 3. This is followed by the proposed caching methods in Section 4, along with conclusions in Section 5.

2. Related Work

Few studies look at the caching problem in F-RAN for mmWave operations. The work in [2] proposes a low complexity subchannel assignment and power control mechanism for mmWave F-RAN that includes caching, user experience constraints, interference suppression, and energy efficiency. The power problem is formulated as an optimization model, and the alternative direction of the multipliers method is leveraged. However, the work here investigates the energy efficiency only and lacks analysis of network delay and cost. Further, the work in [3] accounts for user mobility in transmission scheduling for caching at the fog nodes. The optimal scheduling problem is formulated as a stochastic
nonlinear mixed-integer programming, and multi-hop relaying caching is proposed. The work leverages device-to-device (D2D) communication in the hopping process to enable simultaneous transmission in the scheduling of caching at the edge nodes. The analysis is limited to the amount of cached data under different hop numbers, locations, time slots, and transmission power.

Another set of caching schemes is proposed for F-RAN, albeit not in the context of mmWave NR operations. For example, the authors in [4] propose a cooperative coded caching method by deploying deep reinforcement learning to search for the optimal content coded caching approach. The search here is applied to every request by the controller in a high-power node in terms of the deep Q network model, i.e., to enhance the probability of successful transmission. Furthermore, in [5], an online caching approach is introduced that considers the time-varying nature of the popular content with a focus on long-term normalized delivery delay, i.e., based on the temporal dependency of the coding times aggregated during multiple time slots in the high signal-to-noise (SNR) ratio regime. The authors in [6] deploy a cooperative multi-point (CoMP) transmission approach at fog nodes with cache modules that serve clustered users. They aim to optimize the minimum weighted signal-to-interference noise ratio (SINR) among all clusters while considering cluster fairness and load balancing in the backhaul. Namely, the work aims to jointly optimize the clustering formulation for increased fairness and multicast beamforming, with consideration for power consumption. The authors in [7] assume that fog access points for a distributed cache cluster that cooperatively serve MS requests in efforts to reduce fronthaul traffic. Here, cache placement is optimized by concatenating an MDS code with a repetition code, i.e., to save energy and bandwidth. Namely, repeating the same packet of some F-APs allows for multicasting over the fronthaul link, which saves energy and bandwidth.

In [8], fog computing is used as an intermediate communication interface between the underlying and global tiers of Information-centric Networks (ICN), where content is processed and stored at the fog nodes. The goal here is to reduce the total cached content using content labeling and sharing. The goal here is to reduce the content request delay and enhance the cache hit rate. Another cooperative caching approach in [9] applies graph theory to maximize the offloaded traffic. Specifically, the work formulates the clustering problem while considering cooperative caching and local content popularity. Thereafter, a graph-based approach is proposed to solve the problem.

Reference [10] also aims to enhance the cache hit rate in F-RAN by again formulating the edge caching problem as an optimization model to determine the optimal policy for a content popularity prediction algorithm. The algorithm considers the content features and user preferences, as well as an offline user preference learning algorithm based on the online gradient descent and follows the regularized leader method. The goal is to estimate upcoming content popularity at reduced complexity and track the content with spatial and temporal dynamics in time. In [11], online caching based on time-varying content is analyzed to characterize the long-term normalized delivery time, which captures the temporal dependence of the coding latencies accrued across multiple time slots in the SNR regime. Further, the work studies online caching and delivery schemes for serial and pipelined transmission modes across fronthaul and edge network entities.

In [12], the cooperative caching problem is formulated as a probability-triggered combinatorial multi-armed bandit problem. Then, an enhanced multi-agent reinforcement learning algorithm is developed to solve the problem. The solution combines user preference and content popularity prediction based on a real dataset to enhance the cache hit rate. Federated learning is also applied in [13] to develop a content popularity prediction framework for D2D communication in F-RAN in an effort to maximize the cache hit ratio. Note that the work considers individual privacy in the content popularity prediction model. Further note that it only utilizes the user’s local model in the training process.

Furthermore, the work in [14] presents an F-RAN model based on a femto-base station cluster (FBSC) structure to coordinately serve users on the mmWave bands. However, it still relies on microwave bands between fog stations and central units. Overall, the
work focuses on resource allocation problem for caching placement and lacks analysis of the caching mechanism including caching delay, network cost, and energy consumption. It rather optimizes the hybrid precoder to reduce the transmission delay as a function of the BS transmit power for different metrics such as the volume of cached content. A joint caching and recommendation policy is proposed in [15] for F-RAN by considering a dynamic request model of various times. The caching problem is formulated as an optimization model and reinforcement learning is leveraged to maximize the net profit of each F-AP. Furthermore, a double deep Q-network is used to allocate optimal caching policy with a content recommendation at reduced complexity. Again, the work here lacks the context of mmWave systems, i.e., incorporating beamforming and channel models. Finally, a content caching strategy for F-RAN is developed in [16] using a federated deep reinforcement learning algorithm to enhance the caching performance in terms of content request delay and cache hit rate. Here, the model learns content popularity at multiple cooperative F-Aps, and then a deep Q-network is applied to learn the request content data in each F-AP. However, this work again lacks the context of mmWave operations and the analysis on network cost and energy consumption.

Various outcomes have been proposed from the 5G MiEdge project in [1]. First, stochastic optimization and matching theory are leveraged in [17] to propose an online computation offloading method in MEC, while considering offloading requests, channel conditions, user mobility, and computation queues. However, the work lacks the integration of mmWave system and its requirements, and it lacks network function virtualization and content popularity analysis. Service migration is investigated due to user mobility in [18] that considers latency penalty. It proposes a method to resume the established service at a new mobile edge host to maintain service continuity. It aims to overcome service disruption and resource consumption in the backhaul links. However, it suffers from increased signaling when notifying the MS terminal about a new optimal edge host during path configuration. This adds a complexity at the MS to transmit service parameters. Proactive computation caching methods are presented in Reference [19], which considers task popularity, size, and cached resources that can vary from the incoming resource demands. However, the methods lack service delay analysis and implementation of network function virtualization. Reference [20] integrates MEC and mmWave communications to enable a high-level network architecture with an application- and user-centric orchestration that collects various parameters using a liquid RAN C-plane such as user position, network load, and data popularity. The work in [21] provides an overview and strengths, weaknesses, opportunities, and threats (SWOT) analysis of the integration of mmWave and MEC from a business and economic aspects of 5G systems. However, the work is only limited to SWOT analysis, without provisioning caching schemes or a technical framework. In [22], a prefetching algorithm is proposed for mmWave communications in MEC. It develops mobility and traffic models and reduces system latency and enhances user data rates. Finally, the work in [23] addresses computation offloading mechanisms that consider blockage effects in mmWave links in cloud-RAN. However, it lacks service request models, network function virtualization, and operating constraints of fog nodes.

Overall, existing schemes on caching in F-RAN focus on content prediction to enhance the hit ratio, while considering limited user request specifications such as computation and delay bound. Further, the studies lack analysis of the resource constraints at the edge of the network without accounting for network cost in the caching process as compared to online prefetching.

3. Multi-Layer Network Architecture

A multi-layer architecture is utilized here that is comprised of MS terminals, fog nodes in cluster distribution, and the cloud core, as depicted in Figure 1.
Figure 1. F-RAN Architecture for mmWave NR Operations.

MS Terminals (Layer I): User terminals that demand various services of different delay and capacity specifications. Terminals can be mobile stations, sensors, vehicles, desktops, laptops, etc., which are distributed across the fog nodes in each cluster, i.e., each cluster is comprised of multiple fog nodes.

Primary Nodes in Fog-RAN (Layer II): This layer consists of distributed, homogenous fog nodes that are collocated with RRHs at the proximity of MS terminals in Layer I and cloud-RAN in Layer III. They are equipped with beamforming architectures to provide high bandwidth links with the MSs. It is the gateway at the edge of the network that receives traffic requests and thus provides services at stringent delay bounds and limited resources.

Secondary Nodes in Fog-RAN (Layer II): Another set of fog nodes that possess higher resources as compared to the primary nodes, albeit less resources than the cloud core. Every secondary node manages a cluster of primary nodes through direct links. This intermediate structure combines the benefits of Layers II and IV, i.e., higher resources at the expense of a slight increase fronthaul delays.

Cloud-RAN (Layer IV): This layer is comprised of widely dispersed cloud nodes that possess abundant resources. It acts as the network BBU and contains the NFs that are offloaded to the fog nodes via wireless fronthaul links that operate on microwave sub 6 GHz.

4. System Model
4.1. Network Model

Consider a set of RRHs distributed over a geographical area, and each RRH is collocated with a fog node \( n_i \), \( n_i \in N \), where \( N \) is the total number of fog nodes in the network. A fog node acts as the processing unit that delivers various NF without relaying to the cloud core that acts as the BBU. Functions include control, communication, storage, and management, etc. The processing of various NFs at the edge of the network alleviates latency in the backhaul of the network, as opposed to traversing traffic to the core, which incurs aggregated delays. The local processing also saves in resources that can be used instead to enhance the network capacity, i.e., radio resources and processing units at the cloud core. Further, each Fog node \( n_i \) in the F-RAN possesses processing capacity and memory expressed by \( q_{\text{prc}}(n_i) \) and \( q_{\text{ee}}(n_i) \), respectively. Note that \( q_{\text{prc}}(n_i) \leq Q_{\text{prc}}(n_i) \) and
where the variables $a$, $A$, $\chi$, and $\gamma$ are the maximum capacity and memory, respectively. Additionally, incurred processing delay at node $n_i$ is denoted by $D_{prc}(n_i)$. Further, consider that the fog nodes are interconnected via wireless links, thus forming a set $E = \{e\}$ of links. The instantaneous available link bandwidth is denoted by $b(e)$, $b(e) \leq B(e)$, where $B(e)$ represents the maximum available bandwidth. Here, $D_{prp}(e)$ accounts for the propagation delay at the link.

4.2. Service Request Model from MS

A request $r$ that generates from the MS $m$ features various NF types $t \in T$ as per the specific service, where $T$ is the total number of NF types. Each NF $f_i$ requires computation and memory resources that are denoted by $q_{prc}(f_i)$ and $q_{me}(f_i)$, respectively. This accumulates to $Q_r$ in total processing and memory requirements for the request as per Equation (1). Moreover, the variable $b_r$ accounts for the link resources.

$$Q_r = \sum_{t \in F_r} q_{prc}(f_i) + \sum_{t \in F_r} q_{me}(f_i)$$  \hspace{1cm} (1)

Along with this, an incoming request to the fog nodes in the F-RAN has different requirements in terms of delay bound, lifetime, resources, and service types. The request is received by a fog node that is termed as the source node, $src$. Thereafter, it is traversed to the destination node $dst$, which can be the same MS, another MS, or a fog node. The intermediate nodes that interconnect the $src$ and $dst$ nodes must have sufficient resources to route the data and the NFs. Along with this, each request $r$ can be modeled by a 6-tuple $r = <src, dst, F_r, Q_r, b_r, \delta_r>$. The variable $F_r$ defines the set of NFs $F_r = \{f_i\}$. The variable $\delta_r$ defines the delay bound for the request.

4.3. Beamformer Architecture at the MS

Different beamforming models exist for the MS terminal in the context of mmWave bands such as the analog beamforming models in [24–26] comprised of various array geometries. A request $r$ generated from MS $m \in M$ propagates through a wireless link. Each MS is equipped with RF chain $\psi_{MS}$ that is connected to a uniform linear array (ULA) for directional transmission with the F-RAN, where it radiates a single beam. The chain is composed of $A_{MS}$ antennas that are equally spaced at $d_{ant} = \lambda/2$, where $\lambda$ is the wavelength, $\lambda = \frac{c}{\chi}$, where $c$ is the speed of light and $\chi$ is the carrier frequency. Note that antennas are parallely fed by phase shifters to provide continuous beam scanning. Overall, this formulates an analog beamformer that radiates a primary beam vector $v(\Theta_{0}^{MS})$ that points towards $\Theta_{0}^{MS}$ direction. The vector for this direction is part of the vector matrix, $V_{MS}$ at the analog stage of the chain, i.e., $v = V_{an}$, where $V_{an}$ is the analog precoder. It is determined by the far-field array factor for the ULA, written as [27,28],

$$AF_{MS} = \frac{1}{A_{MS}} \sum_{i=1}^{A_{MS}} a \exp(j(A_{MS} - 1)(kd_{ant} \cos \Theta_{0}^{MS} + \beta))$$  \hspace{1cm} (2)

where the variables $a$, $k$, and $fi$ denote the amplitude of the antenna at the MS, wave number, i.e., $k = 2\beta/\gamma$, and the progressive phase shift between the elements at the MS, respectively.

4.4. Beamformer Architecture at the BS

Each mmWave F-RAN is equipped with a digital beamformer to allow for communication with multiple MSs and provide spatial multiplexing for an increased link capacity. Hence, the RRH are equipped with a UCA composed of $\psi_{RH}$ antennas that are equally spaced along circular ring. In contrast to the MS design, each antenna here is connected to one RF chain. Note that the total number of antennas is equal to the number of RF chains, i.e., $A_{RH} = \psi_{RH}$. Additionally, the overall radiated pattern from $r_{BS}$ represents a
beamforming vector, $p_{RH}$, in the beamforming matrix, $P_{RH} = p_{bb}p_{an}$, where $p_{bb}$ and $p_{an}$ represent the baseband and analog beamforming stages, respectively.

4.5. Downlink Model

Consider that the set of MSs that operate in time-division duplexing (TDD) mode with the F-RANs, with reciprocal channel state information (CSI) knowledge. The downlink received signal model in the RF domain $y_{an}$ at the MS is modeled as [29],

$$y_{an} = V_{RH}Hz + w,$$

where the variables $H$, $z$, and $w$ represent the complex channel, control signal, and additive white Gaussian noise (AWGN), respectively, where $w \sim N(0, \sigma_w^2)$, with variance $\sigma_w^2$. Here, the variable $V_{AP}$ denotes the beamforming matrix at the F-RAN. Furthermore, the received signal $y_{bb}$ at the MS subsequent to the combiner section, $C_{MS}$, is expressed as,

$$y_{bb} = P_{tr}C_{MS}V_{RH}Hz + C_{MS}^Hw,$$

where the variables $P_{tr}$ and $C_{MS}$ denote the transmitted signal power and the MS combiner, respectively, and where $C_{MS}$ includes baseband and analog combiners, such that $C_{MS} = C_{bb}C_{an}$. Furthermore, the instantaneously received signal $y_{inst}$ at the MS due to $p$ beamforming and $c$ combining vectors, where $v \in V_{RH}$ and $c \in C_{MS}$, is written as,

$$y_{inst} = P_{tr}^HvHz + c^Hw.$$

In general, the poor scattering propagation nature at mmWave bands imposes the use of geometric channel models, written as [30,31],

$$H = \sqrt{A_{RH}A_{MS}} \Gamma_{bl} \sum_{l=1}^{L} h_lV_{RH}C_{MS}^H,$$

where the variable $\Gamma_{bl}$ and $h_l$ in order denote the blockage path loss and the complex gain of the $l$-th path for $L$ number of paths captured in $K$ clusters. Note that path gains here are also assumed to be Rician-fading, i.e., $h_l \sim R(0, \zeta)$, where $\zeta$ represents the ratio between power in the first path and other paths that possess reduced power levels. Moreover, the variables $V_{RH}$ and $C_{MS}$ represent the response vectors that capture the channel at the F-RAN and MS, respectively. Therefore, the overall response vector for the beamformer at the MS is given by the RF precoding matrix and is computed by the periodic RF precoder at the F-RAN and MS, respectively. Therefore, the overall response vector for the beamformer at the MS is given by the RF precoding matrix and is computed by the periodic array factor for the UCA at $\phi_s^MS$ azimuth and $\phi_s^MS$ elevation pointing directions for each section (likewise for the F-RAN).

4.6. Cache Model

Assume that the set of locally installed NFs at the cache of the fog nodes in the F-RAN is denoted by $S = \{s\}$, i.e., $S \subset F_r$ with a processing and memory resource of $q_{pre}(s)$ and $q_{me}(s)$, respectively. Here, the cache status of a NF is represented by a binary variable,

$$y_{ij} = \begin{cases} 1, & \text{if } f_i \text{ is cached on } n_j, \\ 0, & \text{otherwise}. \end{cases}$$

Here, the set of cached NFs is predetermined by content popularity and most requested NFs as per the study in [32] as depicted in Figure 2 (retrieved from the study in [32]). It shows the most popular NFs in the network, where $f_1$ and $f_2$ account for call-in and call-out, $f_3$ and $f_4$ account for sms-in and sms-out, and $f_5$ represents internet-access traffic, respectively. Note that the set of popular traffic varies according to the geographic areas and nature of the traffic. Here, the same algorithm can be applied to other types of NFs and content popularity as well. Along with this, the cached NFs demanded by the MS terminal
$m$ can be directly retrieved from the local fog node, whereas uncached content needs to be traversed to the cloud core for processing, i.e., conveyed to the cloud BBU via the fronthaul links.

![Content Popularity for Network Functions](image)

**Figure 2.** Content Popularity for Network Functions [32].

5. Caching Placement in mmWave F-RAN

The received traffic from various MS terminals at the fog nodes differs in terms of the processing capacity, lifetime, and delay bound. Along with this, the goal is to place popular NFs at the edge of the network at the F-RAN without offloading to the cloud BBU. However, given the limited available resources of F-RAN, efficient cache placement methods are proposed on the fog nodes.

Two schemes are studied to distribute cached content across the fog nodes. Namely, clustered caching and distributed caching methods. In the earlier groups, all the popular NFs across a single fog node in Layer III, given the abundant resources across the secondary fog nodes. Meanwhile, the distributed caching method allocates a single popular NF across every node in each cluster. The interconnection between the cluster nodes yields in the total cached NFs, where each host fog is directly connected to an adjacent node that caches another popular NFs. Consider the details of the caching methods.

**Clustered Caching Method:** Consider an incoming request $r$ that generates from an MS terminal $m$, requesting $F$ set of NFs. This request is received as the src node, which is the closest fog node to the terminal. The traffic traverses to Layer III, which contains the cached NFs. If any of the NFs is not supported at the node, then mapping is conducted across the direct neighbor in Layer II, which is in the direction of the path toward the src node. Consider the set of popular NFs as $S = |s|$ for each batch requests $R = |r|$. This set is group mapped on the secondary fog node as,

$$\sum_{s \in S} n_j = 1, \forall n_j \in N, \forall r \in R.$$

(8)

Now the path for request $r$ to the host secondary node is established to minimize the function $F = \min (D(src, n_j))$,

$$Path(r) = \left\{ n_i, e|\min (D(n_i, n_j)) \right\}, \forall r \in R$$

(9)

where $D(., .)$ is the path delay between any two nodes across Layer II and III.

Here, a set of shortest paths result based on the end-to-end delay, which are sorted in ascending order, after which the least-delay path is selected. If any of the requested NFs in
\( F_r \) does not exist in \( S \), then mapping is performed across the closest primary fog node in Layer II, which is termed as the host primary node \( n_i \), i.e.,

\[
\pi_i \leftarrow n_i | \text{min}D(n_i, n_j), \forall n_i \in Path(r) | d(n_i, n_j) = 1.
\]

(10)

where mapping is formulated as,

\[
F_{\text{map}} = F_r - S, \text{i.e., } \sum_{\forall f \in F_{\text{map}}} \pi_i = 1, \forall f \notin S.
\]

(11)

**Distributed Caching Method:** The popular NFs are cached separately across the fog nodes of each cluster. Then, incoming requests are routed to the sequence of interconnected nodes that host the entire cached NFs. The rationale behind this distribution mechanism is to avoid increased traffic directed toward a single node and link, i.e., in effort to avoid node/link congestion and failure. Accordingly, a single popular NF \( s \in S \) is cached on each host node \( n_i \) in Layer II,

\[
\sum_{\forall s \in S} \pi_i = 1, \forall n_i \in N
\]

(12)

The result is a set of interconnected nodes and links that form the cached path, \( Path(S) \), where all incoming requests are routed across it, \( Path(S) = \{\pi_i, e\} \). Along with this, an incoming request demands \( F_r \) NFs. If any \( f_i \exists \{\pi_i | \pi_i \in Path(S)\}, \forall f_i \in F_r \) then the method maps \( f_i \) on the first node in the path closest to the MS \( m \).

6. **Performance Evaluation**

The proposed methods are evaluated versus the conventional uncached approaches in terms of service delay, network cost, and energy consumption. The uncached approaches map network functions online based on the least path delays between terminal users and fog nodes in Layers II and III, i.e., without prior knowledge about content popularity as opposed to the proposed caching methods.

The complexity of the proposed methods depends on the scale of network size in terms of nodes and links modeled as \( O(|N| \log |E|) \). Further, the use of LSTM results in an additional complexity scaled by the number of network functions that need to be forecasted. A saliency for the LSTM algorithm here is that it features a reduced computational complexity versus other deep learning methods. It accepts random weight initializations and a variety of state information without the requirement of a prior setting of the input states [33]. The complexity per weight and time is modeled as backpropagation through time (BPTT), \( O(1) \) [34]. This is scaled by the number of network functions \( F_r \). Along with this, the overall run-time complexity is formulated as \( O(|N| \log |E| + |F_r|) \).

6.1. **Network Service Delay**

The service delay is determined by the total provisioning time required to allocate cached NFs on the nodes, time required to provide a copy of a cached NF, and the time required to map any uncached NFs, i.e., instantiation delay. This delay also includes the hopping time over all intermediate nodes and link propagation delay in the path from the src to the nodes that host the NFs. Given the stringent delay, it is vital here to provide services at the least delays. Figure 3 shows the service provisioning delay for the proposed caching methods for various numbers of incoming MS requests versus the conventional uncached methods. At low incoming requests, the network is still underutilized and most of the nodes are unoccupied. Therefore, the delay here is relatively low for the two methods, e.g., ranges of 2.5–3 ms and 3–3.5 ms for the clustered and distributed methods for the first 80 requests, respectively. In the clustered approach, incoming requests are routed to the secondary nodes that host all the cached NFs. This routing leads to hopping over primary nodes, which causes a slight delay (e.g., 3 ms at 80 requests) without providing NFs to them.
In the distributed method, the cached NFs are hosted at the primary nodes in Layer II, which requires each node to provide service to the request, thus yielding in increased delay, e.g., 4.4 ms at 80 requests. At high traffic volumes, some NFs are uncached, which results in mapping requirements of new NFs. Along with this, the distributed method maps these additional new NFs on the established caching path, which aggregates the delay further. Meanwhile, the clustered approach maps the new NFs on the direct neighbors of the secondary node, which yields in shorter paths, thus yielding shorter delays that approach 5.5 ms at 200 requests.

Meanwhile, the uncached methods suffer from noticeable delays attributed to the provisioning delays of nodes after the terminals initiate service requests. Here, both the distributed and clustered uncached methods search for the best node that yields in the least aggregate path delays to host the network functions. This is opposed to the proposing caching methods that allocate the NFs in advance, thus reducing the instantiation time of the functions running on the nodes. Here the distributed uncached method suffers the highest delay approaching 9.5 ms at 200 requests, i.e., attributed to the high number of nodes in the request path hosting the NFs. This delay is reduced in the clustered uncached method by aggregating the NFs on a single node, i.e., conditioned by available resources, thus reducing the total nodes in the path, which minimizes the link delays between the nodes.

6.2. Network Cost

The service cost across all the network is proportional to the aggregated delays across the paths, i.e., nodes and links in the request path occupy resources that add to the total cost. Namely, the cost for a single request \( r \) of \( \text{Path}_r \) is gauged as,

\[
\text{cost}(r) = \sum_{n_i \in \text{path}_r} \sum_{f_j \in \text{path}_r} \Lambda(n_i) \alpha_r + \sum_{e \in \text{path}_r} \sum_{v_k \in \text{path}_r} \Lambda(e) b_r, \tag{13}
\]

where the variable \( \Lambda(n_i) \) represents cost of the node per usage unit for the cached and uncached NFs requested by the request. Further, the variable \( \Lambda(e) \) denotes the cost of the usage unit for the link that is proportional to the link bandwidth demands \( b_r \).

For further analysis on cost, the work in [35] presents a comprehensive cost model to "minimize application deployment cost and maximize the cloud owner's revenue in terms
of the requested traffic, the backhaul capacity, and the number of MEC servers”. The model incorporates the leasing cost of the MEC and cloud resources by end-users considering latency requirements. It also includes the deployment cost of servers (e.g., software running cost), the number of users using the server, and the size of their traffic.

Figure 4 shows that the cost for the two proposed methods is approximately similar with a slight increase for the clustered approach. This is attributed to the high utilization of the secondary node in each cluster during the entire service time, along with the cost of the primary nodes and links that route the requests to the secondary nodes in Layer III. For example, the cost approaches 78 units at 80 requests. The distributed method requires multiple primary nodes to cache the various NFs, which also increases the cost. However, the reduction in cost here is due to the unoccupancy on the secondary nodes, particularly at low traffic volumes. However, at higher traffic, the distributed method is compelled to map uncached NFs on the secondary nodes to accommodate the high demand, in particular after 80 requests.

![Figure 4. Overall Network Cost for Various MS Requests.](image)

The uncached provisioning methods demand a higher number of nodes in the request path, given the redundant mapping of all incoming NFs, regardless of the traffic popularity; this is mostly shown in the clustered approach that favors nodes in the upper fog layer that features more available resources, at the detriment of increased cost. For example, the cost is increased by 10% for the clustered method as compared to the cached counterpart at 200 requests, and 35% increase for the distributed uncached method.

6.3. Energy Consumption in the Network

Energy consumption is a vital efficiency performance factor in network operations. Here, the total power consumption in the network is gauged during the caching and provisioning process for the NFs across the network nodes and links. It also includes the beamforming power consumption at the RRHs to communicate with the MS requests. The total power consumption during the caching and provisioning times across the fog BBU nodes, links and the beamforming architectures account for the energy consumption in the network. First, for the network infrastructure, the work here adopts the power consumption model in [36], which measures the power usage in the cloud for a certain processing duration. Along with this, the power consumption at a single BBU node is $z$, 

$$z(n_i) = \beta(n_i)z(n_i)_{\text{max}} + (1-\beta(n_i))\sigma(n_i)z(n_i)_{\text{max}}, \quad (14)$$
where $f_i(n_i)$ is the power consumption for the node in idle status, and $z(n_i)_{\text{max}}$ accounts for the maximum power consumption at the node. Furthermore, the variable $w(n_i)$ denotes the node saturation rate, which depends on the utilization rate. Along these lines, the total power consumption for all the nodes in the path for request $r$ is determined by,

$$z_{\text{cons}} = \sum_{n_i} z(n_i), n_i \in N.$$  \hspace{1cm} (15)

This is added to the power consumption due to the number of switches in the network and their power consumption as,

$$z_{\text{total}} = z_{\text{cons}} + X_z(x),$$ \hspace{1cm} (16)

where $z(x)$ is the power consumption for switch $x$ for $X$ total switches.

Figure 5 shows the total energy consumption levels. Similar to the analysis for the delay and cost, the clustered method groups cached NFs on the secondary node, which requires higher power consumption as compared to the primary nodes. The power consumption here also includes the utilization rate of the beamforming architectures at the F-RAN. Overall, the clustered method suffers from higher energy consumption, i.e., 11% higher at 200 requests versus the distributed caching counterpart. The behavior of the uncached methods follows the trend of the cost analysis attributed to the increased number of nodes and links in the request path. This increased operational cost suffers from an increase in power and energy consumption as well.

![Figure 5. Energy Consumption in the Network.](image)

7. Conclusions

This paper proposes cache distribution methods for popular network functions at fog-access radio to reduce service provisioning and fronthaul delays. This setting makes it suitable for millimeter wave communications that demand reduced fronthaul delays to alleviate the effect of path and penetration losses. Results show that the clustered caching method minimizes the total time required for service provisioning as compared to the distributed method. Meanwhile, the latter achieves a slight reduction in network cost and energy. Overall, this supports a tradeoff based on network preference, i.e., the priority metric selected for operation. Future efforts will investigate the two methods in terms of online caching for real-time traffic.
Author Contributions: Conceptualization, A.A., M.B., H.E. and M.A.J.; methodology, N.S.; software, M.B.; validation, M.A.J., N.S. and R.S.; formal analysis, R.S.; investigation, M.B.; resources, H.E.; data curation, R.S.; writing—original draft preparation, A.A. and M.A.J.; writing—review and editing, H.E.; visualization, M.A.J.; supervision, A.A.; project administration, A.A.; funding acquisition, A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by Deanship of Scientific Research at King Faisal University, grant No. 1811025.

Conflicts of Interest: The authors declare no conflict of interest.

References
1. Koslowski, K.; Pilz, J.; Weiler, R.; Keusgen, W.; Haustein, T.; Tran, G.K.; Ogawa, H.; Nishiuchi, H.; Tomura, T.; Sakaguchi, K. 5G-MiEdge—Millimeter-wave Edge Cloud as an Enabler for 5G Ecosystem. *IEEE Trans. Commun.* **2021**, *69*, 3652–3663. [CrossRef]
2. Zhang, H.; Zhu, L.; Long, K.; Li, X. Energy efficient resource allocation in millimeter-wave-based fog radio access net-works. In Proceedings of the 2018 2nd URSI Atlantic Radio Science Meeting (AT-RASC), Gran Canaria, Spain, 28 May–1 June 2018.
3. Niu, Y.; Liu, Y.; Li, Y.; Zhong, Z.; Ai, B.; Hui, P. Mobility-Aware Caching Scheduling for Fog Computing in mmWave Band. *IEEE Access* **2018**, *6*, 69358–69370. [CrossRef]
4. Fan, Y.; Wang, X. Deep Reinforcement Learning for Cooperative Coded Caching Strategy in Fog Radio Access Network. In Proceedings of the 2020 IEEE 6th International Conference on Computer and Communications (ICCC), Chengdu, China, 11–14 December 2020.
5. Azimi, S.M.; Simeone, O.; Sengupta, A.; Tandon, R. Online edge caching in fog-aided wireless networks. In Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017.
6. Chen, D.; Kuehn, V. Weighted max-min fairness oriented load-balancing and clustering for multicast cache-enabled F-RAN. In Proceedings of the 2016 9th International Symposium on Turbo Codes and Iterative Information Processing (ISTC), Brest, France, 5–9 September 2016.
7. Mostafa, S.; Sung, C.W.; Xu, G.; Chan, T.H. Cooperative Caching for Ultra-Dense Fog-RANs: Information Optimality and Hypergraph Coloring. *IEEE Trans. Commun.* **2021**, *69*, 3652–3663. [CrossRef]
8. Wang, M.; Wu, J.; Li, G.; Li, J.; Li, Q. Fog computing based content-aware taxonomy for caching optimization in information-centric networks. In Proceedings of the 2017 IEEE Conference on Computer Communications Workshops (INFOCOM Workshops), Atlanta, GA, USA, 1–4 May 2017.
9. Cui, X.; Jiang, Y.; Chen, X.; Zhengy, F.; You, X. Graph-based cooperative caching in Fog-RAN. In Proceedings of the 2018 International Conference on Computing, Networking and Communications (ICNC), Maui, HI, USA, 5–8 March 2018.
10. Jiang, Y.; Ma, M.; Bennis, M.; Zheng, F.C.; You, X. User preference learning-based edge caching for fog radio access net-work. *IEEE Trans. Commun.* **2018**, *67*, 1268–1283. [CrossRef]
11. Azimi, S.M.; Simeone, O.; Sengupta, A.; Tandon, R. Online Edge Caching and Wireless Delivery in Fog-Aided Networks With Dynamic Content Popularity. *IEEE J. Sel. Areas Commun.* **2018**, *36*, 1189–1202. [CrossRef]
12. Jiang, F.; Zhang, X.; Sun, S. A D2D-enabled cooperative caching strategy for fog radio access networks. In Proceedings of the 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications, London, UK, 31 August–3 September 2020.
13. Jiang, F.; Cheng, W.; Gao, Y.; Sun, Y. Caching strategy based on content popularity prediction using federated learning for F-RAN. In Proceedings of the 2021 IEEE/CIC International Conference on Communications in China (ICCC Workshops), Xiamen, China, 28–30 July 2021.
14. Hao, W.; Sun, G.; Muta, O.; Zhang, J.; Yang, S. Coordinated Hybrid Precoding Design in Millimeter Wave Fog-RAN. *IEEE Syst. J.* **2019**, *14*, 673–676. [CrossRef]
15. Yan, J.; Jiang, Y.; Zheng, F.; Yu, F.R.; Gao, X.; You, X. Distributed Edge Caching with Content Recommendation in Fog-RANs Via Deep Reinforcement Learning. In Proceedings of the 2020 IEEE International Conference on Communications Workshops (ICC Workshops), Virtual Conference, 7 June 2020.
16. Zhang, M.; Jiang, Y.; Zheng, F.C.; Bennis, M.; You, X. Cooperative edge caching via federated deep reinforcement learning in fog-rans. In Proceedings of the 2021 IEEE International Conference on Communications Workshops (ICC Workshops), Montreal, QC, Canada, 14–23 June 2021.
17. Merluzzi, M.; Di Lorenzo, P.; Barbarossa. Dynamic joint resource allocation and user assignment in multi-access edge computing. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019.
18. Yunoki, K.; Shinbo, H. Carry-on state service handover between edge hosts for latency strict applications in mobile networks. In Proceedings of the 2018 21st International Symposium on Wireless Personal Multimedia Communications (WPMC), Chiang Rai, Thailand, 25–28 November 2018.
19. Di Pietro, N.; Calvanese, E. Proactive computation caching policies for 5G-and-beyond mobile edge cloud networks. In Proceedings of the 2018 26th European Signal Processing Conference (EUSIPCO), Rome, Italy, 3–7 September 2018.
20. Tran, G.K.; Nishiuchi, H.; Frascolla, V.; Takinami, K.; De Domenico, A.; Strinati, E.C. Architecture of mmWave edge cloud in 5G-MiEdge. In Proceedings of the 2018 IEEE International Conference on Communications Workshops (ICC Workshops), Kansas City, MO, USA, 20–24 May 2018.

21. Frascolla, V.; Engelsch, J.; Takinami, K.; Chiaraviglio, L.; Salsano, S.; Yunoki, K.; Barberis, S.; Palestini, V.; Sakaguchi, K.; Haustein, T.; et al. Millimeter-waves, MEC, and network softwarization as enablers of new 5G business opportunities. In Proceedings of the 2018 IEEE Wireless Communications and Networking Conference (WCNC), Barcelona, Spain, 31 March 31–3 April 2018.

22. Nishiuchi, H.; Tran, G.K.; Sakaguchi, K. Performance Evaluation of 5G mmWave Edge Cloud with Prefetching Algorithm-Invited Paper. In Proceedings of the 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), Porto, Portugal, 3–6 June 2018.

23. Barbarossa, S.; Ceci, E.; Merluzzi, M.; Calvanese-Strinati, E. Enabling effective mobile edge computing using millimeter wave links. In Proceedings of the 2017 IEEE International Conference on Communications Workshops (ICC Workshops), Paris, France, 21–25 May 2017.

24. Jaesim, A.; Aldalbahi, A.; Siasi, N.; Oliveira, D.; Ghani, N. Dual-Beam Analog Beamforming for mmWave Communications. In Proceedings of the 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 10–12 October 2019.

25. Jasim, M.A.; Siasi, N.; Aldalbahi, A.; Ghani, N. Soft self-handover scheme for mmWave communications. In Proceedings of the 2019 IEEE SoutheastCon, Huntsville, AL, USA, 11–14 April 2019.

26. Aldalbahi, A.; Siasi, N.; Ababneh, M.; Jasim, M. Grating Lobes for Enhanced Scattering Intensity in Millimeter Wave Sparse Channels. In Proceedings of the 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 7–9 January 2019.

27. Jasim, M.; Aldalbahi, A.; Shakhatreh, H. Beam aggregation for instantaneous link recovery in millimeter wave communications. In Proceedings of the 2018 14th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), Limassol, Cyprus, 15–17 October 2018.

28. Jasim, M.; Ghani, N. Adaptive initial beam search for sparse millimeter wave channels. In Proceedings of the 2017 26th Wireless and Optical Communication Conference (WOCC), Newark, NJ, USA, 7–8 April 2017.

29. Jasim, M.; Ghani, N. Generalized pattern search for beam discovery in millimeter wave systems. In Proceedings of the 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall), Toronto, ON, Canada, 24–27 September 2017.

30. Alkhateeb, A.; El Ayach, O.; Leus, G.; Heath, R.W. Channel Estimation and Hybrid Precoding for Millimeter Wave Cellular Systems. IEEE J. Sel. Top. Signal Process. 2014, 8, 831–846. [CrossRef]

31. Aldalbahi, A.; Shahabi, F.; Jasim, M. BRNN-LSTM for Initial Access in Millimeter Wave Communications. Electronics 2021, 10, 1505. [CrossRef]

32. Aldalbahi, A.; Jasim, M.A.; Shahabi, F.; Mazin, A.; Siasi, N.; Oliveira, D. Deep Learning for Primary Sector Prediction in FR2 New Radio Systems. IEEE Access 2021, 9, 157522–157539. [CrossRef]

33. Aldalbahi, A.; Shahabi, F.; Jasim, M. Instantaneous beam prediction scheme against link blockage in mmwave communications. Appl. Sci. 2021, 11, 5601. [CrossRef]

34. Sepp, H.; Schmidhuber, J. Long short-term memory. Neural Comput. 1997, 9, 1735–1780.

35. Nakazato, J.; Nakamura, M.; Yu, T.; Li, Z.; Maruta, K.; Tran, G.K.; Sakaguchi, K. Market Analysis of MEC-Assisted Beyond 5G Ecosystem. IEEE Access 2021, 9, 53996–54008. [CrossRef]

36. Méndez-Rial, R.; Rusu, C.; Alkhateeb, A.; González-Prelcic, N.; Heath, R.W., Jr. Channel estimation and hybrid combining for mmWave: Phase shifters or switches? In Proceedings of the Information Theory and Applications Workshop (ITA), San Diego, CA, USA, 1–6 February 2015.