IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 8, 2020, accepted December 13, 2020, date of publication December 16, 2020,
date of current version December 30, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3045304*

# Performance Analysis of Opportunistic Fog Based Radio Access Networks

**JOFINA JIJIN, (Member, IEEE), BOON-CHONG SEET, (Senior Member, IEEE), AND PETER HAN JOO CHONG, (Senior Member, IEEE)**
Department of Electrical and Electronic Engineering, Auckland University of Technology, Auckland 1010, New Zealand
Corresponding author: Boon-Chong Seet (boon-chong.seet@aut.ac.nz)

**ABSTRACT** The advent of the Internet-of-Things (IoT) has led to a rapid growth in data generation. The amount of computation resources required to process the massive data generated by IoT devices, along with the new intelligent use and applications of IoT data such as smart city that can be computation intensive and delay sensitive, have caused an increase in demand for locally available resources at network edge for computation offloading. To address this issue, we have proposed the concept of opportunistic fog radio access network (OF-RAN), which extends the computation capacity of existing fog-RAN (F-RAN) by establishing virtual fog access points (v-FAP) opportunistically using resourceful user devices that participate as service nodes. In this article, we develop an analytical model to evaluate the offloading performance of three RAN architectures: the traditional cloud radio access network (C-RAN), the existing F-RAN, and the proposed OF-RAN. The performances are analyzed in terms of their energy consumption, completion delay, and failure rate, under the effect of varying scenarios.

**INDEX TERMS** Opportunistic fog, radio access network, energy analysis, delay analysis, failure analysis.

## I. INTRODUCTION

The advancements and convergence in wireless, computing, sensor and actuation technologies have enabled a plethora of smart devices collectively known as the Internet-of-Things (IoT) [1]. These devices are deployed ubiquitously and in large numbers for a diverse range of applications, leading to massive data generation and an exponential growth in demand for transmission and computation resources. In order to meet these challenges, network architectures such as cloud radio access network (C-RAN) [2] and fog radio access network (F-RAN) [3] have been introduced. Although C-RAN has immense computation resources in the cloud, it suffers from a number of drawbacks, such as heavy workload at the centralized baseband unit (BBU) pool, stringent backhaul capacity constraint, and difficulty in catering to delay sensitive applications [4]. F-RAN, on the other hand, deploys fog access points (FAPs) at network edge to provide cloud-like services to IoT devices. The FAPs can be deployed as new dedicated entities in an existing infrastructure, or on existing entities of an infrastructure such as a small cell base station augmented with fog functionality [5].

The associate editor coordinating the review of this manuscript and approving it for publication was Rashid Mehmood.

Recently, we have proposed the opportunistic fog radio access network (OF-RAN) [6], which is evolved from the concepts of F-RAN and opportunistic networks (oppnets). The latter are mission-oriented ad hoc networks setup to utilize opportunistically available local resources. Each oppnet grows from a 'seed' node, which recruits one or more available local 'helper' nodes to assist with a specific mission. In our proposed OF-RAN, the *seed node* and *service nodes* are equivalent to the FAP of F-RAN, and helper nodes of oppnet, respectively. A seed node recruits locally available resourceful user devices such as high-end smart phones as service nodes that function collectively as a virtual FAP (v-FAP) to serve a resource-limited client such as an IoT device [7], [8]. The resourceful user devices can be incentivized as in [9] to lease their resources (e.g. computing, storage, and energy resources) for serving resource-limited clients and be remunerated based on their performance (e.g. in terms of timeliness and reliability). The computation to be offloaded from a client to a v-FAP is referred as *service task*.

We consider a scenario shown in Fig. 1 where OF-RAN, F-RAN, and C-RAN co-exist in the access layer to serve a terminal layer composing of both resourceful and resource-limited user/IoT devices. Our proposed OF-RAN can play a complementary role to F-RAN and C-RAN by harnessing

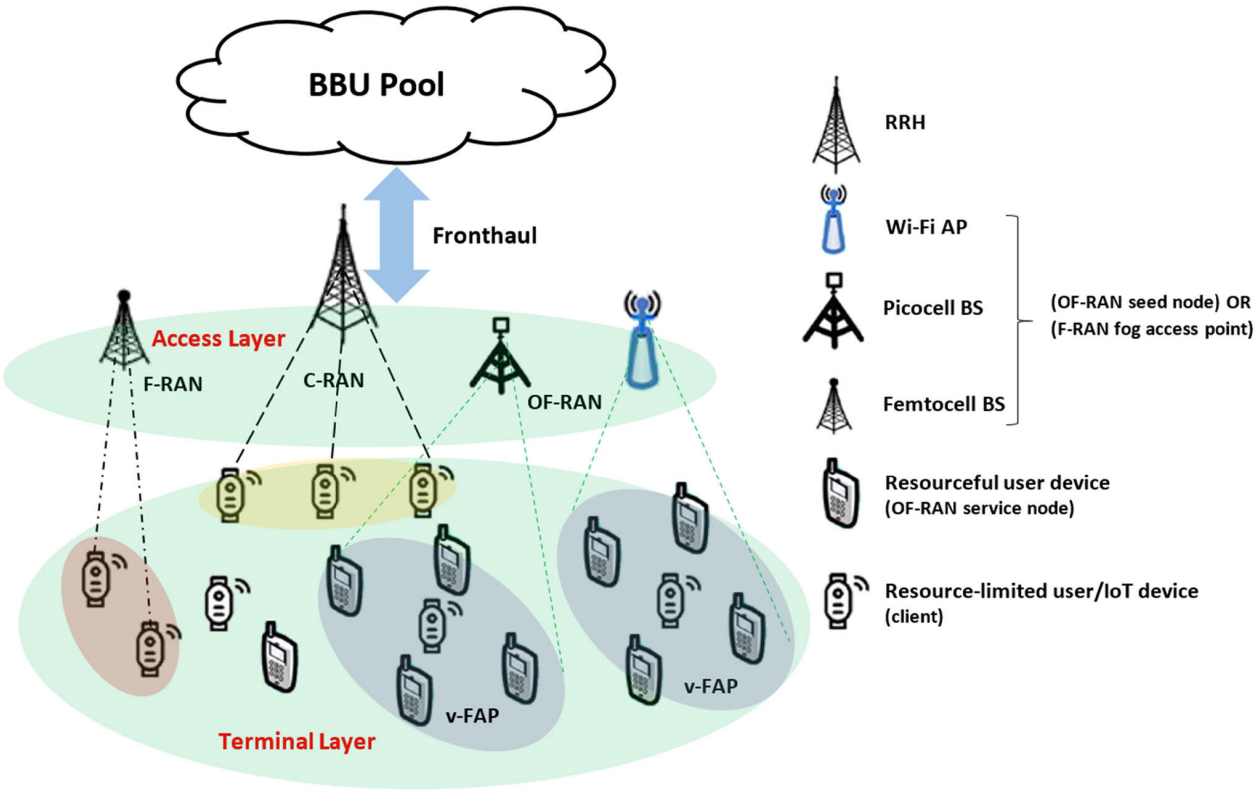


**FIGURE 1.** System architecture of co-existing RANs.

resourceful terminal devices to deal with the computation workloads from a large number of offloading clients simultaneously in a time- and energy-efficient manner. A resource-limited client can offload its task in three ways: (i) offload to C-RAN by transmitting the task to the BBU in the cloud through a remote radio head (RRH); (ii) offload to F-RAN by transmitting the task to the FAP; (iii) offload to OF-RAN by transmitting the task to the v-FAP. In F-RAN or OF-RAN offloading where multiple FAPs or v-FAPs are available, the RRH could assist the client in selecting the most appropriate FAP or v-FAP for offloading. To provide insights into the complementary nature of these RAN architectures, we develop an analytical model to evaluate their performances in terms of the energy consumption, completion delay, and failure rate, under various offloading scenarios.

The rest of the article is organized as follows. Section II reviews the related works. Section III presents the system model. Section IV develops the analytical models for the three RAN architectures under consideration. Sections V and VI discuss the simulation environment, and the results, respectively. Finally, Section VII concludes the article with some directions for future work.

## II. RELATED WORKS

This section reviews recent related works with a focus on the performance analysis of fog-based RANs. The authors in [10] proposed a power model to determine the power consumption and energy efficiency of F-RAN. They also evaluated its latency and compared the results with C-RAN. It was found that F-RAN incurred lower latency, but consumed more power, leading to a lower energy efficiency. However, in the latency analysis, the authors considered the processing time as a constant factor, and did not consider the impact of task complexity on both latency and power consumption.

In [11], a cooperative algorithm is proposed to enable cooperation between multiple F-RANs to provide low-latency computing services. The cooperation is coordinated by a master fog node that allocates computation tasks to each of the other cooperating fog nodes, considering their available resources. However, the authors have only evaluated the service latency, and did not consider other aspects such as energy expenditure. It is also unclear how the cooperative F-RAN may perform against other architectures such as a hybrid cloud-fog RAN. Similarly, the authors in [12] only focused on the latency issue, and proposed to minimize the latency of offloading users through a joint optimization of the communication and computation resources.

In [13], the authors analyzed the performance of an F-RAN under different caching strategies and transmission modes. The former refers to different ways of utilizing mobile devices, RRHs, and FAPs to store and deliver popular content to the clients. The latter defines the ways by which the client can access the content, such as from a fog access point (FAP mode), from other user devices via relaying (relay mode), or remotely from the cloud (C-RAN mode). A testbed was also implemented in [14] to demonstrate their

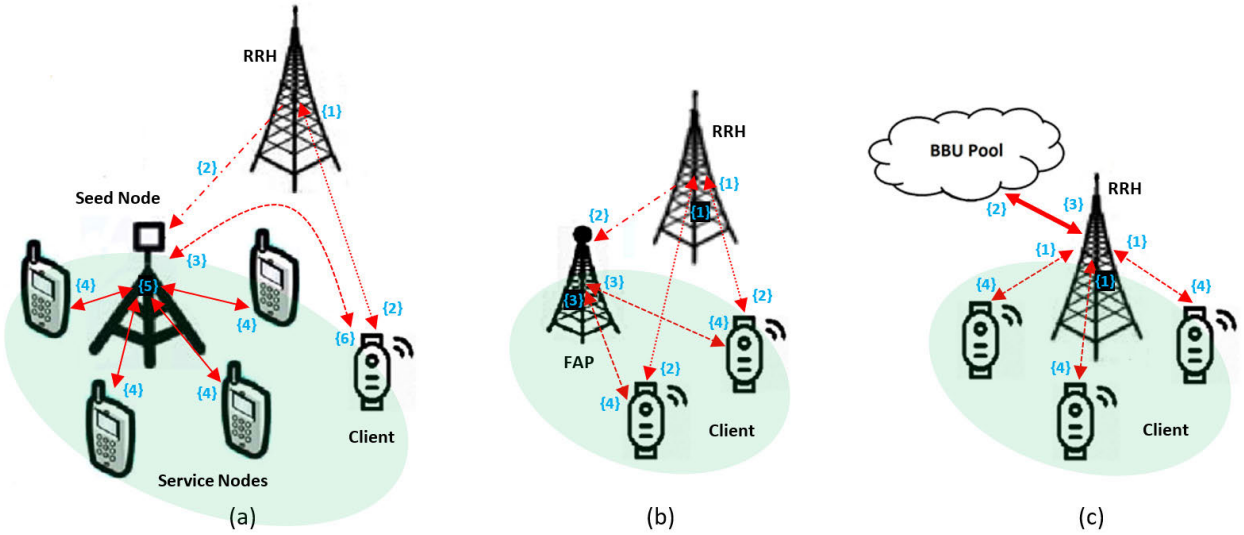


**FIGURE 2.** Network model for: (a) OF-RAN; (b) F-RAN; and (c) C-RAN.

F-RAN for video content acquisition. In these works, however, the authors have only focused on evaluating the caching but not the offloading performance.

In [15], the performance of F-RAN under an opportunistic computation offloading strategy is studied. The strategy utilizes a probabilistic computation offloading model, which determines the likelihood of the wireless channel in use to support the transmission rate required for offloading, leading to three possible processing modes: local mode in which client processes the task by itself; fog mode in which client offloads the task to a FAP; and cloud mode in which client offloads the task to cloud computing center via a FAP. However, the authors have only analyzed the average delay performance, and did not consider factors such as the delay sensitivity of the task in their offloading strategy.

Similarly focused on offloading strategy, but additionally concerned about jamming and interferences from nearby radio devices during offloading, the authors in [16] proposed reinforcement learning based schemes that jointly optimizes the selection of edge device for offloading, offloading rate, and transmit power so that the computational latency and energy consumption are minimized while the offloading signal-to-noise-plus-interference ratio is maximized. The authors analyzed the computation complexity of the proposed schemes and showed they can reduce computational latency and energy in the presence of jamming and interferences.

The deep reinforcement learning has also been applied to realize network slicing in F-RAN in [17], where computing, caching and radio resources are orchestrated to meet the performance requirements of two different types of services: hot-spot and vehicle- to-infrastructure (V2I). To address the challenges of high cost of data offloading and model training for implementing network intelligence at the edge, an evolved architecture of F-RAN is proposed in [18], which employs federated learning (a.k.a. collaborative learning) to realize intelligent signal processing and network management with less communication overhead and greater efficiency than existing centralized learning paradigms.

In our recent work [19], we addressed the task-to-node assignment in OF-RAN as a multi-objective optimization problem. The goal was to optimally assign the computation task from an offloading client to the service nodes of a v-FAP with the objectives of minimizing the energy and latency of v-FAP while maximizing the fairness among service nodes. The impact of various parameters on the optimality of the assignment was evaluated. However, this work has neither analyzed the scalability of the OF-RAN architecture under increasing task complexity, nor compared the performances between OF-RAN and existing architectures such as F-RAN and C-RAN.

## III. SYSTEM MODEL

### A. NETWORK MODEL

Fig. 2 shows the considered network model for our proposed OF-RAN as well as existing F-RAN and C-RAN architectures. In the OF-RAN, a resource-limited client offloads its task by first sending a request {1} including the task requirements to its associated RRH, which in turn notifies the client {2} to offload its task to an available seed node within its neighborhood. The client then sends its task {3} to this seed node for processing. Upon receiving, the seed node firstly determines the optimal task-to-node assignment (TNA) based on the performance objectives [19]. The optimal TNA defines a set of suitably selected service nodes for the v-FAP, and appropriately sized sub-tasks to be assigned to each service node. Based on this assignment, the seed node sends the sub-tasks to the service nodes for processing {4}, collates the processed sub-tasks from the service nodes {5}, and forwards them to the client {6}. To emulate a real-world scenario, we consider both service nodes and service tasks to be heterogeneous, each having a different computation capacity, and complexity, respectively.

We further consider that the service nodes are registered RAN users that can be trusted to assist the resource-limited clients when called upon. This trust can be facilitated by a blockchain-enabled OF-RAN architecture [20] in which the smart contract is used to implement an algorithm for distributed formation and management of v-FAPs among trustless user devices acting as service nodes.

In the F-RAN, a client similarly offloads its task by first sending a request {1} to its associated RRH, which acts as a F-RAN controller [21] in charge of receiving offloading requests and distributing them to the FAPs. The RRH then notifies the client {2} to offload its task to an available FAP within its neighborhood. The client then sends its task {3} to this FAP for processing. On completion, the FAP forwards the processed task to the client {4}. However, unlike in OF-RAN where each v-FAP only serves a single client, the FAP in F-RAN may serve multiple clients at a time.

On the other hand, a client in C-RAN offloads its task {1} to the associated RRH, which in turn sends it to BBU pool in the cloud {2} for processing. The RRH receives the processed task {3} from BBU pool, and then forwards it to the client {4}. The wireless access links between RRH, clients, FAPs, seed nodes, and service nodes are considered to be using millimeter waves (mmWave), while the wired fronthaul link between the RRH and BBU pool in the cloud is using an optical fiber.

### B. PATH LOSS MODEL

The path loss model calculates the power loss of a signal as it travels through space. For the wireless access links in this article, the close-in (CI) free-space reference distance model [22] proposed for 5G systems is used to calculate the path loss. Compared to other path loss models such as 3GPP's alpha-beta-gamma (ABG) model, the CI model offers computation simplicity yet better accuracy in path loss prediction across a wide range of frequencies and distances [23].

The path loss $PL_{u,v}$ in decibel (dB) of a link from node $u$ to node $v$ is given by (1), where $f$ is the signal frequency, $d_0$ is the close-in free space reference distance in meters, $c$ is the speed of light in meters per second, $\alpha$ is the path loss exponent, $d_{u,v}$ is the distance between node $u$ and node $v$ in meters, and $X_\sigma$ is the shadowing component in dB described by a zero-mean Gaussian random variable with standard deviation $\sigma$.

$$PL_{u,v}\,(dB) = 20\,log_{10}\left(\frac{4\pi f d_o}{c}\right) + 10\alpha\,log_{10}\left(\frac{d_{u,v}}{d_o}\right) + X_\sigma \quad (1)$$

The corresponding received signal power $P_v^{rx}$ in decibel-milliwatts (dBm), and data rate $R_{u,v}$ in bits per second (bps), based on the determined path loss, are given by (2), and (3), respectively, where $P_u^{tx}$ is the transmitted signal power in dBm of node $u$, $b$ is the channel bandwidth, and $P_v^{no}$ is the average noise power in dBm at the receiver node $v$.

$$P_v^{rx}\,(dBm) = P_u^{tx}\,(dBm) - PL_{u,v}\,(dB) \quad (2)$$

$$R_{u,v} = b\,log_2\left(1 + \frac{P_v^{rx}}{P_v^{no}}\right) \quad (3)$$

## IV. ANALYTICAL MODEL

This section presents the analytical model for evaluating the offloading performance of the OF-RAN, F-RAN, and C-RAN as illustrated in Fig. 2. In this model, expressions are obtained for three system level performance metrics, namely total delay, total energy consumption, and offloading failure.

### A. DELAY

In OF-RAN, the total delay $D_{OF}^{total}$ incurred while offloading a task from the client to a v-FAP comprising of one seed node and $N$ service nodes, is constituted of transmission, propagation, and processing delays. As shown in (4a), the transmission delay includes the time for sending a request of size $\partial$ from client to RRH, a notification of size $\varphi$ from RRH to client, a task of size $T$ (or $T'$ after processing) between client and seed node, and $N$ sub-tasks each of size $M_n$ (or $M_n'$ after processing) between seed node and $N$ service nodes, where $n$ is the index of a service node and $\sum_{n=1}^{N} M_n = T$.

The propagation delay between a transmitting node $u$ and receiving node $v$ is given by the ratio of their distance $d_{u,v}$ and the speed of light $c$. The processing delay incurred by a service node $Sv_{(n)}$ for a sub-task of size $M_n$ is given by the ratio of the number of floating-point operations (FLOPs) required by the sub-task (depending on $M_n$ in bits and task complexity $\gamma$ in FLOPs per bit) and the computation capacity $C_{Sv}$ of the service node in FLOPs per second (FLOPS). Without loss of generality, we assume all service nodes have the same computation capacity $C_{Sv}$, which can be found using (4b) where $\delta_{Sv}$ is the service node's performance in FLOPs per cycle per core, $\beta_{Sv}$ is the number of cores, and $\zeta_{Sv}$ is the processor frequency in hertz (or cycles per second).

For a seed node with $N$ or more antennas, it can transmit all $N$ sub-tasks at the same time using one antenna for each service node. Thus, all $N$ sub-tasks can be processed in parallel by the service nodes. Likewise, the seed node can simultaneously receive the processed sub-tasks from all $N$ service nodes. Hence, the delay between the seed node and service nodes, which include the time for transmission, propagation, and processing of all $N$ sub-tasks, is the maximum of all pair-wise delays between the seed node and each service node.

$$\begin{aligned} D_{OF}^{total} = {} & \left(\frac{\partial}{R_{C,R}} + \frac{d_{C,R}}{c}\right) + \left(\frac{\varphi}{R_{R,C}} + \frac{d_{R,C}}{c}\right) \\ & + \left(\frac{T}{R_{C,S}} + \frac{d_{C,S}}{c}\right) + \max_{n=1\ldots N}\left\{\left(\frac{M_n}{R_{S,Sv(n)}} + \frac{d_{S,Sv(n)}}{c}\right)\right. \\ & \left. + \frac{\gamma M_n}{C_{Sv}} + \left(\frac{M_n'}{R_{Sv(n),S}} + \frac{d_{Sv(n),S}}{c}\right)\right\} + \left(\frac{T'}{R_{S,C}} + \frac{d_{S,C}}{c}\right) \end{aligned} \quad (4a)$$

$$C_{Sv} = \beta_{Sv}\delta_{Sv}\zeta_{Sv} \quad (4b)$$

In F-RAN, the total delay $D_{FR}^{total}$ incurred while offloading a task from the client to a FAP is similarly derived as shown in (5a). The processing delay incurred by a FAP for a task of size $T$ is simply given by the ratio of the number of FLOPs required by the task (depending on $T$ and task complexity $\gamma$)

and the computation capacity $C_F$ of the FAP in FLOPS. Like OF-RAN, the $C_F$ can be found using (5b) where $\delta_F$, $\beta_F$, and $\zeta_F$ refers to the FAP's number of FLOPs per cycle per core, the number of cores, and processor frequency, respectively.

$$D_{FR}^{total} = \left(\frac{\partial}{R_{C,R}} + \frac{d_{C,R}}{c}\right) + \left(\frac{\varphi}{R_{R,C}} + \frac{d_{R,C}}{c}\right) + \left(\frac{T}{R_{C,F}} + \frac{d_{C,F}}{c}\right) + \frac{\gamma T}{C_F} + \left(\frac{T'}{R_{F,C}} + \frac{d_{F,C}}{c}\right) \quad (5a)$$

$$C_F = \beta_F \delta_F \zeta_F \quad (5b)$$

In C-RAN, the total delay $D_{CR}^{total}$ incurred while offloading a task from the client to BBU is given by (6a), in which the transmission delay includes not only the time for sending a task of size $T$ (or $T'$ after processing) between the client and RRH, but also between the RRH and BBU via the optical fronthaul, where $c_{(op)}$ denotes propagation speed in the optical fiber. Like F-RAN, the processing delay incurred by a BBU for a task of size $T$ is given by the ratio of the FLOPs required by the task (depending on $T$ and task complexity $\gamma$) and the computation capacity $C_B$ of the BBU in FLOPS. Similarly, the $C_B$ can be found using (6b), where $\delta_B$, $\beta_B$, and $\zeta_B$ refers to the BBU's number of FLOPs per cycle per core, the number of cores, and processor frequency, respectively.

$$D_{CR}^{total} = \left(\frac{T}{R_{C,R}} + \frac{d_{C,R}}{c}\right) + \left(\frac{T}{R_{R,B}} + \frac{d_{R,B}}{c_{(op)}}\right) + \frac{\gamma T}{C_B} + \left(\frac{T'}{R_{B,R}} + \frac{d_{B,R}}{c_{(op)}}\right) + \left(\frac{T'}{R_{R,C}} + \frac{d_{R,C}}{c}\right) \quad (6a)$$

$$C_B = \beta_B \delta_B \zeta_B \quad (6b)$$

### B. ENERGY

In OF-RAN, the total energy $E_{OF}^{total}$ incurred while offloading a task from the client to a v-FAP is constituted of communication energy and processing energy, as shown in (7a). The communication energy includes the energy for sending a request of size $\partial$ from client to RRH, a notification of size $\varphi$ from RRH to client, a task of size $T$ (or $T'$ after processing) between client and seed node, and $N$ sub-tasks each of size $M_n$ (or $M_n'$ after processing) between seed node and $N$ service nodes, where $n$ is the index of a service node and $\sum_{n=1}^{N} M_n = T$.

The processing energy depends on the service node's energy efficiency $E_{Sv}$ in joules per cycle, the size of each sub-task $M_n$ in bits, and the OF-RAN computation intensity $\omega_{OF}$ in CPU cycles per bit. The $\omega_{OF}$ can be found using (7b) where $\gamma$ is the task complexity in FLOPs per bit, $\delta_{Sv}$ is the service node's performance in FLOPs per cycle per core, and $\beta_{Sv}$ is the number of cores.

$$E_{OF}^{total} = \frac{\partial P_C^{tx}}{R_{C,R}} + \frac{\varphi P_R^{tx}}{R_{R,C}} + \frac{T P_C^{tx}}{R_{C,S}} + \sum_{n=1}^{N} \frac{M_n P_S^{tx}}{R_{S,Sv(n)}} + \sum_{n=1}^{N} \omega_{OF} M_n E_{Sv} + \sum_{n=1}^{N} \frac{M_n' P_{Sv(n)}^{tx}}{R_{Sv(n),S}} + \frac{T' P_S^{tx}}{R_{S,C}} \quad (7a)$$

$$\omega_{OF} = \frac{\gamma}{\beta_{Sv} \delta_{Sv}} \quad (7b)$$

In F-RAN, the total energy $E_{FR}^{total}$ incurred while offloading a task from the client to a FAP is similarly derived as shown in (8a). The processing energy depends on the FAP energy efficiency $E_F$ in joules per cycle, the size of task $T$ in bits, and the F-RAN computation intensity $\omega_{FR}$ in CPU cycles per bit. Like OF-RAN, the $\omega_{FR}$ can be found using (8b) where $\gamma$ is the task complexity, $\delta_F$ is the FAP performance in FLOPs per cycle per core, and $\beta_F$ is the number of cores.

$$E_{FR}^{total} = \frac{\partial P_C^{tx}}{R_{C,R}} + \frac{\varphi P_R^{tx}}{R_{R,C}} + \frac{T P_C^{tx}}{R_{C,F}} + \omega_{FR} T E_F + \frac{T' P_F^{tx}}{R_{F,C}} \quad (8a)$$

$$\omega_{FR} = \frac{\gamma}{\beta_F \delta_F} \quad (8b)$$

In C-RAN, the total energy $E_{CR}^{total}$ incurred while offloading a task from the client to BBU is given by (9a), in which the communication energy includes not only the energy for sending a task of size $T$ (or $T'$ after processing) between the client and RRH, but also between the RRH and BBU via the optical fronthaul, where $P_{R(op)}^{tx}$ and $P_{B(op)}^{tx}$ denotes the optical transmit power of RRH, and BBU, respectively. Similarly, the processing energy depends on the BBU energy efficiency $E_B$ in joules per cycle, the size of task $T$ in bits, and the C-RAN computation intensity $\omega_{CR}$ in CPU cycles per bit given by (9b).

$$E_{CR}^{total} = \frac{T P_C^{tx}}{R_{C,R}} + \frac{T P_{R(op)}^{tx}}{R_{R,B}} + \omega_{CR} T E_B + \frac{T' P_{B(op)}^{tx}}{R_{B,R}} + \frac{T' P_R^{tx}}{R_{R,C}} \quad (9a)$$

$$\omega_{CR} = \frac{\gamma}{\beta_B \delta_B} \quad (9b)$$

### C. FAILURE

The percentage of offloading failure is another performance metric evaluated. Two possible factors of failure considered are: (i) link failure; and (ii) completion time failure. In OF-RAN, there are wireless links between the client, RRH, seed node, and service nodes of a v-FAP. A wireless link from a transmitting node $u$ to a receiving node $v$ (where $u$ and $v$ can be the client, RRH, seed node, or service node) is considered to fail when the received power $P_v^{rx}$ is below the receiver sensitivity $\tau_v$.

Even when all the links are successful, an offloading can still fail when the total delay $D_{OF}^{total}$ incurred to complete a task is longer than the completion time requirement $\phi$ of the task. Hence, for a given client $C$, the offloading is deemed to have failed when either a link or completion time failure occurs, as shown by the failure conditions given in (10):

$$\left(\forall v \in V, P_v^{rx} < \tau_v\right) \vee \left(D_{OF}^{total} > \phi\right) \quad (10)$$

where $V$ is the set of receiving nodes for wireless links used in offloading for $C$ in OF-RAN.

In F-RAN, there are wireless links between the client, RRH, and FAP. Similarly, the offloading is deemed to have failed when the failure conditions in (11) are satisfied, where $D_{FR}^{total}$ is the total delay incurred to complete a task for the client in F-RAN.

$$\left(\forall v \in V, P_v^{rx} < \tau_v\right) \vee \left(D_{FR}^{total} > \phi\right) \quad (11)$$

**TABLE 1. Notations and Definitions.**

| Notation | Definition |
|---|---|
| $PL_{u,v}$ | path loss of link from node $u$ to node $v$ in dB |
| $f$ | signal frequency |
| $d_0$ | close-in free space reference distance in meters |
| $c$, $c_{(op)}$ | propagation speed in free space, and optical fiber, respectively in meters per second |
| $\alpha$ | path loss exponent |
| $d_{u,v}$ | distance between node $u$ and node $v$ in meters |
| $X_\sigma$ | shadowing component in dB with standard deviation $\sigma$ |
| $R_{u,v}$ | data rate of link from node $u$ to node $v$ in bits per second (bps) |
| $b$ | channel bandwidth |
| $P_u^{tx}$ | transmitted signal power of node $u$ in dBm |
| $P_v^{rx}$ | received signal power of node $v$ in dBm |
| $P_v^{no}$ | average noise power at node $v$ in dBm |
| $T$, $T'$ | size of original, and processed task, respectively in bits |
| $D_{OF}^{total}$, $D_{FR}^{total}$, $D_{CR}^{total}$ | total delay incurred in OF-RAN, F-RAN, and C-RAN, respectively in seconds |
| $M_n$, $M'_n$ | size of original, and processed sub-task, respectively assigned to $n_{th}$ service node in bits |
| $N$ | number of service nodes in a v-FAP |
| $\eta$ | number of clients |
| $\partial$ | size of request from OF-RAN/F-RAN client to RRH in bits |
| $\varphi$ | size of notification from RRH to seed node/FAP in bits |
| $\gamma$ | task complexity in floating-point operations (FLOPs) per bit |
| $E_{OF}^{total}$, $E_{FR}^{total}$, $E_{CR}^{total}$ | total energy consumed in OF-RAN, F-RAN, and C-RAN, respectively in joules |
| $E_{Sv}$, $E_F$, $E_B$ | processing energy efficiency of a service node in OF-RAN, FAP in F-RAN, and BBU in C-RAN, respectively in joules per cycle |
| $\omega_{OF}$, $\omega_{FR}$, $\omega_{CR}$ | computation intensity of OF-RAN, F-RAN, and C-RAN, respectively, in cycles per bit |
| $\delta_{Sv}$, $\delta_F$, $\delta_B$ | processor performance of service node, FAP, and BBU, respectively, in FLOPs per cycle per core |
| $\beta_{Sv}$, $\beta_F$, $\beta_B$ | number of processor cores in a service node, FAP, and BBU, respectively |
| $C_{Sv}$, $C_F$, $C_B$ | computation capacity of a service node, FAP, and BBU, respectively in FLOPs per second (FLOPS) |
| $\tau_v$ | receiver sensitivity of node $v$ in dBm |
| $\psi$ | probabilistic state of optical fiber link between RRH and BBU in C-RAN |
| $\psi_{fail}$ | expected failure rate of optical fiber link between RRH and BBU in C-RAN |
| $\phi$ | task completion time requirement in seconds |
| $\zeta_{Sv}$, $\zeta_F$, $\zeta_B$ | processor frequency of service node, FAP, and BBU, respectively, in hertz (or cycles per second) |

[1]Nodes $u$ and $v$ can be any transmitting node, and receiving node, respectively, in the RAN
[2]$u$ or $v$ can be replaced by $C$ (client), $S$ (seed node), $Sv$ (service node), $F$ (FAP), $R$ (RRH), or $B$ (BBU)

In C-RAN, there are not only wireless links between the client and RRH, but also optical fiber links between the RRH and BBU. An optical fiber link is considered to fail when a random probability $\psi$ representing the state of the link is below an expected failure rate $\psi_{fail}$ of the link. Consequently, the offloading failure conditions can be given by (12), where $D_{CR}^{total}$ is the total delay incurred to complete a task for a client in C-RAN.

$$\left[\left(\forall v \in V, P_v^{rx} < \tau_v\right) \vee \left(\psi < \psi_{fail}\right)\right] \vee \left(D_{CR}^{total} > \phi\right) \quad (12)$$

Table 1 lists the notations used in the analytical model and their definitions.

## V. SIMULATION ENVIRONMENT

The analytical model developed in Section IV is implemented in MATLAB to evaluate the offloading performance of all three RAN architectures under varying task complexity ($\gamma$) and number of clients ($\eta$). For the proposed OF-RAN, the impact of varying number of service nodes ($N$) in a v-FAP is also investigated. Table 2 lists the simulation parameters and their realistically chosen values based on the real-world devices or operation settings.

All the wireless links between RRH, clients, FAPs, seed nodes, and service nodes operate at 38 GHz, which is one of the 5G mmWave frequencies. Each wireless node pair communicates over a line-of-sight (LOS) channel with a path loss exponent slightly higher than 2 (free space path loss exponent) and a bandwidth of 500 MHz. The parameters $d_{u,v}$, $T$, $M_n$, and $\phi$ are assigned with random values uniformly distributed on a range as shown in Table 2. Unless otherwise specified, the default values of the following parameters are used: $\gamma = 6250$; $\eta = 15$, and $N = 4$. All results are averaged over 100 simulations and their 95% confidence interval are shown when the margins of error are more than 5% of the mean value, as otherwise they are hardly visible in the graphs.

## VI. RESULTS AND DISCUSSION

### A. EFFECT OF VARYING N IN OF-RAN

Table 3 shows the OF-RAN performance in terms of the total delay, total energy consumption, and offloading failure under the effect of varying number of service nodes ($N$) in a v-FAP. The results are obtained for a default $\eta = 15$ clients and task complexity $\gamma = 6250$ FLOPs per bit. The total failures are further broken down into link and completion time failures. In addition, their 95% confidence interval (CI) are shown as the calculated margins of error are mostly not negligible (>5%).

It can be observed that the total delay decreases as $N$ increases. This is because a larger $N$ splits the service task into smaller sub-tasks, resulting in each service node to incur a smaller processing delay. Since all service nodes are processing in parallel, and the processing delay is more

**TABLE 2. Simulation Parameters.**

| Parameter | Value | Unit |
|---|---|---|
| $P_C^{tx}$, $P_S^{tx}$, $P_{Sv}^{tx}$, $P_F^{tx}$, $P_R^{tx}$ | 30 | dBm |
| $P_{R(op)}^{tx}$, $P_{B(op)}^{tx}$ | 2, 5 | dBm |
| $P_v^{no}$ | 82 | dBm |
| $\tau_v$ | 79 | dBm |
| $d_o$, $d_{u,v}$ | 1, 1 100 | meter(s) |
| $T$, $T'$ | 52,000 68,000 | bits |
| $M_n$, $M_n'$ | 13,000 17,000 | bits |
| $N$ | 1 10 | |
| $\gamma$ | 2,500 10,000 | FLOPs/bit |
| $\eta$ | 1 30 | |
| $\partial$, $\varphi$ | 8,000 | bits |
| $\sigma$ | 3.2 | dB |
| $f$ | 38 | GHz |
| $b$ | 500 | MHz |
| $\alpha$ | 2.05 | |
| $\psi_{fail}$ | 0.998 | |
| $\phi$ | 5 50 | milliseconds |
| $E_{Sv}$, $E_F$, $E_B$ | $1\times10^{-10}$, $5\times10^{-10}$, $1.46\times10^{-8}$ | joules/cycle |
| $\zeta_{Sv}$, $\zeta_F$, $\zeta_B$ | 2.4 | GHz |
| $\beta_{Sv}$, $\beta_F$, $\beta_B$ | 1, 8, 24 | |
| $\delta_{Sv}$, $\delta_F$, $\delta_B$ | 8, 16, 16 | FLOPs/cycle |
| $c$, $c_{(op)}$ | $3\times10^8$, $2\times10^8$ | meters/second |
| $R_{R,B}$, $R_{B,R}$ | 15 25 | Gbps |

**TABLE 3. Effect of $N$ on the OF-RAN Performance ($\gamma = 6250$, $\eta = 15$).**

| $N$ | Delay (ms) | Energy (mJ) | Failure (%) | | | |
|---|---|---|---|---|---|---|
| | | | Link failure | Completion time failure | Total | CI |
| 1 | 25.68 | 71.91 | 4.34 | 33.13 | 37.47 | 1.46 |
| 2 | 20.83 | 72.02 | 7.40 | 20.13 | 27.53 | 1.94 |
| 3 | 16.89 | 72.29 | 10.27 | 12.13 | 22.40 | 2.22 |
| 4 | 13.45 | 72.07 | 12.07 | 6.13 | 18.20 | 2.27 |
| 5 | 11.14 | 72.13 | 17.47 | 3.53 | 21.00 | 2.21 |
| 6 | 9.486 | 72.23 | 19.66 | 2.47 | 22.13 | 2.46 |
| 7 | 8.02 | 72.36 | 21.73 | 0.27 | 22.00 | 2.39 |
| 8 | 7.069 | 72.19 | 22.93 | 0.13 | 23.07 | 2.70 |
| 9 | 6.283 | 72.18 | 26.46 | 0.07 | 26.53 | 2.70 |
| 10 | 5.612 | 72.04 | 29.27 | 0 | 29.27 | 2.51 |

dominant than the transmission and propagation delays in the considered scenario, the total delay for servicing a client is largely dependent on the maximum processing delay among the service nodes in a v-FAP. Hence, increasing $N$ decreases this maximum delay, which in turn decreases the total delay.

On the other hand, the total energy consumption is found to be relatively unaffected by $N$. This is again due to the total energy being dominated by processing energy over communication energy. The processing energy is dependent on the total service task size, i.e. sum of all sub-task sizes, the service node's energy efficiency and OF-RAN computation intensity, which do not change with $N$. The minute changes in total energy are attributed to small differences in communication energy caused by some randomness in the path loss and consequently data rate of the links between nodes.

However, $N$ has a significant impact on the type of failure occurrence. As seen in Table 3, increasing $N$ increases the proportion of link failures, but decreases that of completion time failures. The reason is that a higher $N$ increases the number of links, but decreases the total delay that in turn reduces the number of completion time failure. The total failure rate is minimized when $N = 4$, which explains our choice of setting the default number of service nodes in a v-FAP to this value.

Since the service nodes are only used in OF-RAN, we do not evaluate the effect of $N$ on other types of RAN. In the next two sections, we further evaluate the performance of OF-RAN under the effect of varying task complexity $\gamma$ and number of clients $\eta$, and compare it with the performances of current F-RAN and C-RAN architectures.

### B. EFFECT OF VARYING $\gamma$

Fig. 3 and Fig. 4 show the total delay, and total energy consumption, respectively, of all three RAN architectures under varying task complexity ($\gamma$). Results are obtained under a default number of clients ($\eta = 15$) for an average scenario, and a large number of clients ($\eta = 30$) for a stress scenario. The 95% confidence interval of the results are found to have a margin of error between 0.2−3.3%, which are hardly visible and thus omitted in the graphs.

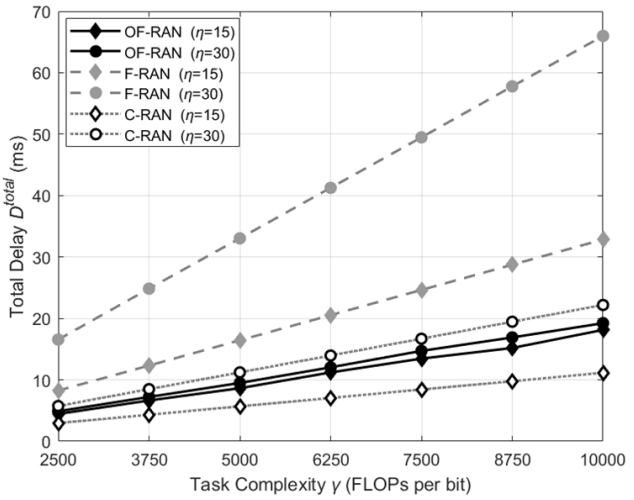


**FIGURE 3. Effect of $\gamma$ on total delay of the RANs.**

Expectedly, both delay and energy consumption increase with $\gamma$, as higher-complexity tasks demand more processing time and energy. In Fig. 3, it can be seen that C-RAN incurs the least delay in the average case ($\eta = 15$), followed by OF-RAN and F-RAN. However, in the higher-stress case ($\eta = 30$), the OF-RAN outperforms both C-RAN and F-RAN. This is because the computation capacity available to each client in C-RAN and F-RAN decreases with higher $\eta$ due to the finite fixed capacity of the BBU, and FAP, respectively. On the contrary, OF-RAN can expand its computation capacity when needed by establishing more v-FAPs (one for each new client) subject to the service nodes availability. This illustrates the inherent scalability of the OF-RAN architecture.

Fig. 4 shows that OF-RAN outperforms C-RAN and F-RAN in total energy consumption for both average and stress scenarios. This is despite the OF-RAN utilizing more nodes, i.e. service nodes, which can lead to higher communication energy consumption. The reason is due to OF-RAN's much lower consumption of processing energy, which dominates the total energy consumption. While the BBU and FAP (processing nodes in C-RAN, and F-RAN, respectively) have higher computation capacity, they are also more power-hungry and consume more energy per CPU cycle. On the other hand, being often battery-powered user devices, OF-RAN's service nodes are operating with better processing energy efficiency or less energy in joules per cycle.

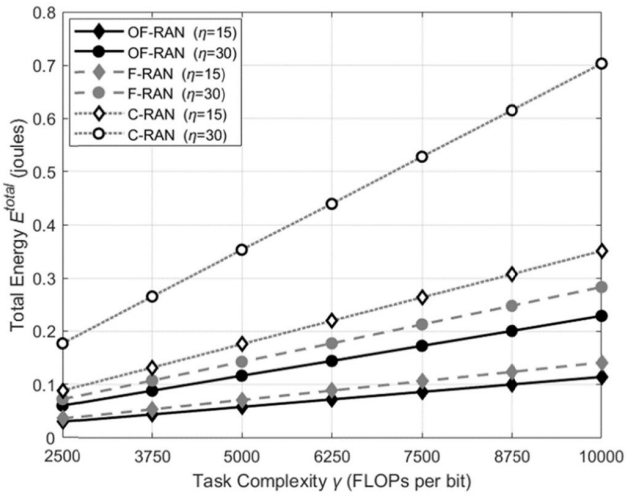


**FIGURE 4.** Effect of $\gamma$ on total energy consumption of the RANs.

Fig. 5 further shows impact of task complexity on the offloading failure, which is broken down into link and completion time failures. The results are shown for the stress scenario ($\eta = 30$) with their 95% confidence interval as the margins of error are not negligible (>5%). Expectedly, the failure rate of all RANs increases with task complexity, caused by an increase in completion time failures due to longer processing time. At low task complexity ($\gamma = 2500$), C-RAN has the lowest failure rate, followed by OF-RAN and F-RAN. Both failures in C-RAN and OF-RAN are mainly due to link failures. However, as task complexity increases to $\gamma = 7500$, OF-RAN begins to outperform as the number of its completion time failures increases at a slower rate than C-RAN and F-RAN. This is consistent with the observation in Fig. 3 where the delay of OF-RAN (predominantly processing delay) increases at a slower rate than C-RAN and F-RAN under the stress scenario. This illustrates once again that OF-RAN is better suited for stress scenarios with not only high number of clients but also high task complexity.

### C. EFFECT OF VARYING $\eta$

In this section, we present a more detailed analysis on the effect of varying number of clients ($\eta$) under average and stress scenarios defined by task complexity. Fig. 6 and Fig. 7 show the total delay, and total energy consumption, respectively, of all three RAN architectures, as $\eta$ varies from 1 to 30. Results are obtained under a default task complexity ($\gamma = 6250$) for an average scenario, and high task complexity ($\gamma = 10000$) for a stress scenario. The corresponding 95% confidence intervals have a margin of error between 0.15−4.3%, which are again hardly visible and thus omitted in the graphs.

For low number of clients ($\eta \leq 10$), C-RAN has the lowest delay in both average and stress scenarios, which is attributed to its high computation capacity, resulting in much smaller processing time that dominates the total delay. F-RAN has a lower initial delay than OF-RAN, but it increases with $\eta$ at a rate faster than OF-RAN and C-RAN. OF-RAN starts to outperform F-RAN at $\eta = 10$, and then C-RAN at $\eta = 30$, in both average and stress cases. Moreover, it exhibits a

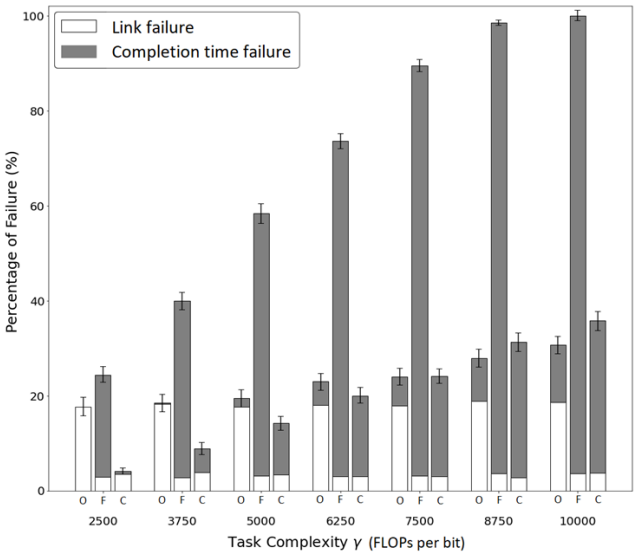


**FIGURE 5.** Effect of $\gamma$ on failure rate of the OF-RAN (O), F-RAN (F) and C-RAN (C).

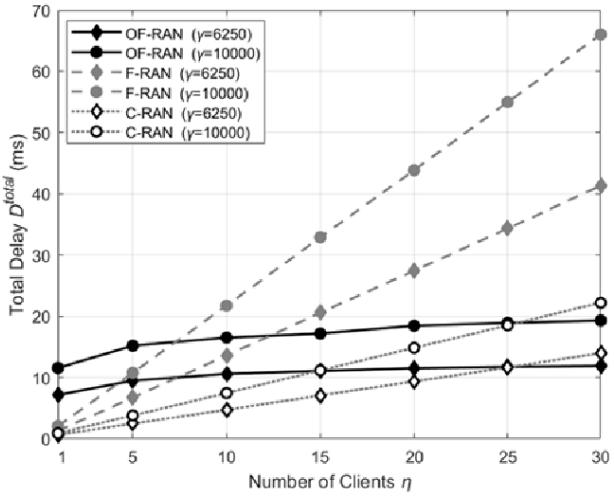


**FIGURE 6.** Effect of $\eta$ on total delay of the RANs.

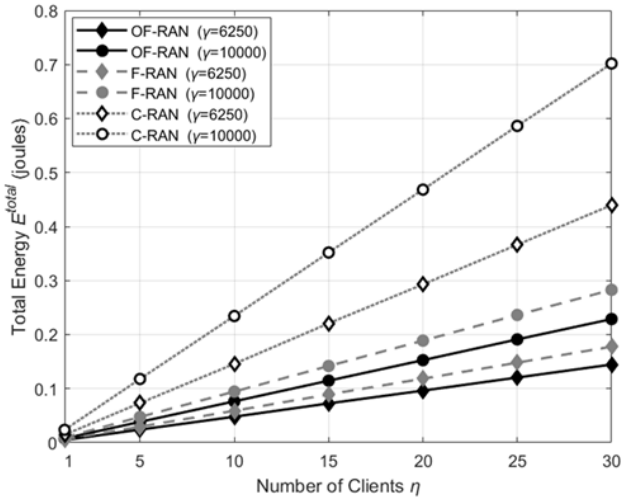


**FIGURE 7.** Effect of $\eta$ on total energy consumption of the RANs.

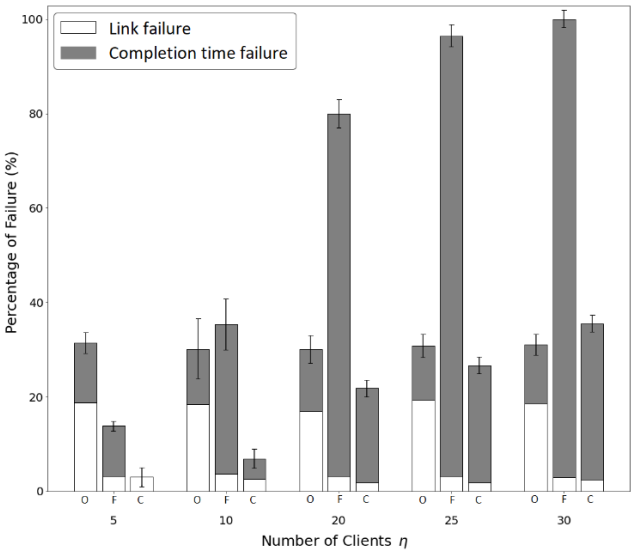


**FIGURE 8.** Effect of $\eta$ on failure rate of the OF-RAN (O), F-RAN (F) and C-RAN (C).

relatively flat delay response to $\eta$, due to its ability to expand computation capacity when needed as explained in previous section. In terms of energy, OF-RAN consistently consumes the least for all $\eta$ and in both average and stress cases. On the other hand, C-RAN consistently consumes the most, mainly due to its power-hungry BBUs that result in high processing energy consumption.

Fig. 8 shows the impact of $\eta$ on the offloading failure. The results are shown for stress scenario ($\gamma = 10000$) with their 95% confidence interval as the margins of error are non-trivial (>5%). Similar to the delay result, the failure in OF-RAN is relatively unaffected by $\eta$. Moreover, it starts to outperform F-RAN at $\eta = 10$, and then C-RAN at $\eta = 30$. On the other hand, the failure in C-RAN and F-RAN increase with $\eta$ due to more completion time failures. This is because a higher $\eta$ reduces the computation capacity available to each client in these RANs, and the impact is greater on F-RAN since FAPs are not as computationally powerful as BBUs in C-RAN. Overall, the results show that the OF-RAN is a promising and scalable architecture.

## VII. CONCLUSION

This article analyzes and compares the offloading performance of OF-RAN with that of existing C-RAN and F-RAN. For each RAN, we develop an analytical model to evaluate its offloading performance in terms of completion delay, energy consumption, and failure rate. The performances are evaluated under the effect of varying number of service nodes, number of clients, and task complexity.

The results show that there exist an optimal number of service nodes for which the failure rate of OF-RAN is minimized. OF-RAN also outperforms C-RAN and F-RAN in all three performance metrics under high-stress scenarios where the task complexity and number of clients are high. This illustrates the scalability of our OF-RAN, which can co-exist with and complement the C-RAN and F-RAN to support computation-intensive and delay-sensitive offloading services.

As future work, we plan to investigate our OF-RAN for distributed deep learning where the client devices offload computation-intensive deep learning tasks to the v-FAPs for time and energy efficient processing. It will be also interesting to investigate how the use of cognitive radio in OF-RAN can expand its notion of opportunistic access to device resources to include opportunistic access to spectrum resources.

## REFERENCES

[1] X. Xu, Y. Li, T. Huang, Y. Xue, K. Peng, L. Qi, and W. Dou, "An energy-aware computation offloading method for smart edge computing in wireless metropolitan area networks," *J. Netw. Comput. Appl.*, vol. 133, pp. 75–85, May 2019.

[2] J. Du, L. Zhao, J. Feng, and X. Chu, "Computation offloading and resource allocation in mixed fog/cloud computing systems with min-max fairness guarantee," *IEEE Trans. Commun.*, vol. 66, no. 4, pp. 1594–1608, Apr. 2018.

[3] Y.-J. Ku, D.-Y. Lin, C.-F. Lee, P.-J. Hsieh, H.-Y. Wei, C.-T. Chou, and A.-C. Pang, "5G radio access network design with the fog paradigm: Confluence of communications and computing," *IEEE Commun. Mag.*, vol. 55, no. 4, pp. 46–52, Apr. 2017.

[4] J. Tang, W. P. Tay, T. Q. S. Quek, and B. Liang, "System cost minimization in cloud RAN with limited fronthaul capacity," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3371–3384, May 2017.

[5] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog-computing-based radio access networks: Issues and challenges," *IEEE Netw.*, vol. 30, no. 4, pp. 46–53, Jul. 2016.

[6] L. Lilien, A. Gupta, Z. Kamal, and Z. Yang, "Opportunistic resource utilization networks—A new paradigm for specialized ad hoc networks," *Comput. Electr. Eng.*, vol. 36, no. 2, pp. 328–340, 2010.

[7] J. Jijin, B.-C. Seet, P. H. J. Chong, and H. Jarrah, "Service load balancing in fog-based 5G radio access networks," in *Proc. IEEE Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Montreal, QC, Canada, Oct. 2017, pp. 1–5.

[8] J. Jijin and B.-C. Seet, "Opportunistic fog computing for 5G radio access networks: A position paper," in *Proc. 3rd EAI Int. Conf. Smart Grid Innov. Frontiers Telecommun. (SmartGIFT)*, Auckland, New Zealand, Apr. 2018, pp. 82–92.

[9] S. Luo, X. Chen, Z. Zhou, X. Chen, and W. Wu, "Incentive-aware micro computing cluster formation for cooperative fog computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2643–2657, Apr. 2020.

[10] H. M. Abdel-Atty, R. S. Alhumaima, S. M. Abuelenin, and E. A. Anowr, "Performance analysis of fog-based radio access networks," *IEEE Access*, vol. 7, pp. 106195–106203, 2019.
[11] T. C. Chiu, A. C. Pang, W. H. Chung, and J. Zhang, "Latency-driven fog cooperation approach in fog radio access networks," *IEEE Trans. Services Comput.*, vol. 12, no. 5, pp. 698–711, Sep./Oct. 2018.
[12] Q. Li, J. Lei, J. Lin, and X. Wu, "Latency minimization for multiuser computation offloading in fog-radio access networks," 2019, *arXiv:1907.08759*. [Online]. Available: http://arxiv.org/abs/1907.08759
[13] M. Peng and K. Zhang, "Recent advances in fog radio access networks: Performance analysis and radio resource allocation," *IEEE Access*, vol. 4, pp. 5003–5009, Sep. 2016.
[14] X. Zhang and M. Peng, "Testbed design and performance emulation in fog radio access networks," *IEEE Netw.*, vol. 33, no. 3, pp. 49–57, May 2019.
[15] M. Xu, Z. Zhao, M. Peng, Z. Ding, T. Q. Quek, and W. Bai, "Performance analysis of computation offloading in fog-radio access networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, May 2019, pp. 1–6.
[16] L. Xiao, X. Lu, T. Xu, X. Wan, W. Ji, and Y. Zhang, "Reinforcement learning-based mobile offloading for edge computing against jamming and interference," *IEEE Trans. Commun.*, vol. 68, no. 10, pp. 6114–6126, Jul. 2020.
[17] H. Xiang, S. Yan, and M. Peng, "A realization of fog-RAN slicing via deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2515–2527, Apr. 2020.
[18] Z. Zhao, C. Feng, H. H. Yang, and X. Luo, "Federated-learning-enabled intelligent fog radio access networks: Fundamental theory, key techniques, and future trends," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 22–28, Apr. 2020.
[19] J. Jijin, B. C. Seet, and P. H. J. Chong, "Multi-objective optimization of task-to-node assignment in opportunistic fog RAN," *Electronics*, vol. 9, no. 3, p. 14, 2020.
[20] J. Jijin, B.-C. Seet, and P. H. J. Chong, "Blockchain enabled opportunistic fog-based radio access network: A position paper," in *Proc. 29th Int. Telecommun. Netw. Appl. Conf. (ITNAC)*, Auckland, New Zealand, Nov. 2019, pp. 1–3.
[21] Y.-Y. Shih, W.-H. Chung, A.-C. Pang, T.-C. Chiu, and H.-Y. Wei, "Enabling low-latency applications in fog-radio access networks," *IEEE Netw.*, vol. 31, no. 1, pp. 52–58, Jan. 2017.
[22] T. S. Rappaport, G. R. Maccartney, M. K. Samimi, and S. Sun, "Wideband millimeter-wave propagation measurements and channel models for future wireless communication system design," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3029–3056, Sep. 2015.
[23] S. Sun, T. S. Rappaport, T. A. Thomas, A. Ghosh, H. C. Nguyen, I. Z. Kovács, I. Rodriguez, O. Koymen, and A. Partyka, "Investigation of prediction accuracy, sensitivity, and parameter stability of large-scale propagation path loss models for 5G wireless communications," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 2843–2860, May 2016.

**JOFINA JIJIN** (Member, IEEE) received the B.Tech. degree in electronics and communications engineering from Mahatma Gandhi University and the M.Tech. degree in telecommunications from the National Institute of Technology Calicut, in 2010 and 2014, respectively. She is currently pursuing the Ph.D. degree with the Auckland University of Technology, New Zealand. Her research interest includes next-generation mobile cellular technologies with a focus on fog-based radio access networks.

**BOON-CHONG SEET** (Senior Member, IEEE) received the Ph.D. degree in computer communications engineering from Nanyang Technological University, Singapore, in 2005. He was a Research Fellow with the National University of Singapore, under the Singapore–Massachusetts Institute of Technology Alliance Program. Since 2007, he has been with the Department of Electrical and Electronic Engineering, Auckland University of Technology, New Zealand, where he is currently an Associate Professor and also leads the Wireless InnovationS in Engineering (WISE) Research Group. His research interest includes info-communication technologies (ICT), including 5G and B5G communications.

**PETER HAN JOO CHONG** (Senior Member, IEEE) received the Ph.D. degree from The University of British Columbia, Canada, in 2000. He was an Associate Professor (tenured) with the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore. From 2011 to 2013, he was an Assistant Head of the Division of Communication Engineering. From 2013 to 2016, he was the Director of INFINITUS, Centre for Infocomm Technology. He is currently a Professor and the Head of the Department of Electrical and Electronic Engineering, Auckland University of Technology, Auckland, New Zealand. He is also an Adjunct Professor with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. His research interests include mobile communications systems, including MANETs/VANETs, multihop cellular networks, and the Internet of Things/Vehicles.