A Novel Cost Optimization Framework for Multi-Cloudlet Environment over Optical Access Networks

Sourav Mondal, Goutam Das, and Elaine Wong
1Dept of Electrical and Electronic Engineering, The University of Melbourne, VIC 3010, Australia.
2G. S. Sanyal School of Telecommunications, Indian Institute of Technology Kharagpur, 721302, India.

Abstract—In the post-4G era, “low latency” has become one of the most important network requirements along with support for ultra-high capacity and ultra-high reliability. This has led to the evolution of “cloudlets” to support similar services provided by legacy cloud technology viz., storage capacity and computational capability. In this paper, we propose a novel framework for cloudlet-empowered-cloud network design and planning, based on optical access infrastructures. Our focus is on network planning to optimize network infrastructure cost by formulating a nonlinear programming model to identify placement locations of cloudlet servers subjected to capacity and latency constraints. We demonstrate the feasibility of the proposed model against urban, suburban and rural scenarios, providing guidance on the installation and maintenance costs. Furthermore, we assess the percentage of incremental energy arising from the presence of cloudlets in the optical access network. The proposed framework is a first in yielding insights that will serve as a foundation for further cloudlet network planning strategies.

Keywords—cloudlet; low latency; network planning; nonlinear programming; optical network; tactile cloudlet

I. INTRODUCTION

In the past decade, the massive commercialization and distribution of smart gadgets like iPhones, Android phones, Google glasses and other mobile device along with smart applications that run on these devices, have exponentially fueled the necessity of more computational powers and storage towards the edge of the network [1]. With the advent of Internet of Things (IoT), billions of devices are expected to connect to the Internet in the near future, igniting the explosive requirement of very high throughput, and seamless network access with low end-to-end network latency [2]. These demands on the network will continue to increase exponentially as the network evolves from supporting machine-to-machine information exchange to supporting tactile human-machine information exchange. For example, applications such as augmented reality, autonomous transport, and real-time environmental monitoring and response, require a low network latency of 1-10 ms. Cloud servers located in the core network are usually very far from the edge and thus fail to meet this strict latency requirement and the mobile devices are always expected to suffer from resource poverty due to their inherent design constraints on weight, battery life and heat dissipation. Thus, researchers are exploring the integration of cloud and mobile computing, so that mobile devices can offload some of their computationally intensive and high memory consuming tasks to the cloud, known as cyber-foraging [3]. These fundamental issues have led to the proposal of cloudlets by Satyanarayanan et al. [1]. Nonetheless, it must be noted that cloudlets cannot completely replace legacy cloud technology because clouds have several significant advantages, most notably network security and the scope for massive data storage and data processing power. When cloud and cloudlets coexist in the network, the combination can however improve end-user experience due to reduced latency.

The term “cloudlet” is commonly referred to as “data center in a box” that “brings the cloud closer”. Defined as a trusted, resource-rich computer or cluster of computers that’s well-connected to the Internet and available for use by nearby mobile devices [1], cloudlets have several advantages and opportunities to improve network latency. Cloudlets can be installed with a wireless access point (WAP) or in the closest vicinity of a group of WAPs. This creates a three-tier network architecture, i.e., cloud – cloudlet – edge devices. Thus, a cloudlet becomes a single hop node from the end-user and can thus provide computational and memory based services much faster than a remote cloud. Cloudlets run virtual machines (VMs) to support the end-user and send periodic updates to the cloud. However, a cloudlet can work independently as well, which makes it a very robust solution even in hostile environments [3].

Since the initial proposal [1], several studies have been carried out to investigate cloudlets. In [4], the authors provided a very compact and concise survey of the existing literature. While discussing the evolution of cloudlet-based mobile computing, they summarized four aspects, namely, concept extension, motivation consolidation, performance improvement and application fields’ expansion. A relative comparison of performances among recently developed frameworks, such as MAUI (Mobile Assistance Using Infrastructure), CloneCloud and COMET (Code Offload by Migrating Execution Transparently) were also presented. The authors also compared key features like CPU, memory and storage of potential cloudlet equivalent devices, e.g., Samsung Galaxy S6, Xiaomi Smart Router and Lenovo ThinkStation P900.

Most of the existing research on cloudlets has been performed from an application point of view and with an underlying assumption that the communications infrastructure technology is based on wireless access. In [5], the authors proposed a novel cloud and cloudlet empowered fiber-wireless (FiWi)-heterogeneous network architecture for LTE-Advanced (LTE-A). The authors provided an overview of the optical fiber based network support architecture for cloudlets. In [6], a cloudlet with control server is implemented at the central office (CO) of a fiber-based local area network for human-machine applications. Nonetheless, the lack of insights on computation offload algorithms and cloudlet network planning e.g., using optimal objective function to either reduce latency, cost and energy consumption or enhance performance, warrants thorough study.

In this paper, we propose a novel framework to plan a complete network infrastructure for cloudlets based on the time division multiplexed (TDM) passive optical network (PON). The TDM-PON can support either the standardized G(E)-PON, EPON or 10-G(E)PON technologies. To the best of our knowledge, the design and planning issue of TDM PON for
Cloudlet support has not been examined so far in existing literature. The prime motivation for this approach is the popularity of the PON as the access network technology due to its low cost per bit, and with the increasing demand of high mobile traffic, radio-over-fiber (RoF) and FiWi technologies appear to be the frontrunners in supporting x-hauling of 5G wireless access and small-cells [5].

Our primary objective is to extract the best performance from the TDM-PON network in terms of latency, while using optimized capacity per cloudlet whereby cloudlet resources are used efficiently with minimal under-utilization or overloading.

In this paper, our key contributions are the following:

- In a TDM-PON based cloudlet network architecture we propose a method to identify suitable field locations for cloudlet placement. Our method of cloudlet placement guarantees latency constraints for data offloading services while optimizing the overall cost of network augmentation. We consider the most common tree-branch topology for the PON architecture with 1:N splitter at the remote node (RN). However, this is scalable to other topologies (ring and bus) [6] of PONs as well.

- We formulate a mixed integer nonlinear programming (MINLP) problem for network planning that identifies cost efficient cloudlet placement locations and assignment of optical network units (ONUs) to cloudlets. We further linearize the objective function and nonlinear constraints to reduce the computational complexity of our design tool. We also include an additional feature that different cloudlets can have different number of racks to improve the utilization of the installed computational facility. In this work, to create a baseline, we focus mainly on static planning [8].

- Our primary design constraint of the optimization problem is to achieve a very low end-to-end latency, i.e., 1 ms, 10 ms and 100 ms, for the planned network. We consider the cloudlets as finite capacity queueing systems and once all the racks are occupied, they offload the rest of the incoming task requests to a remote cloud [10].

- We validate our model for urban, suburban and rural scenarios and display comparative results for deployment cost, average number of racks per cloudlet and the increase of energy consumption over a typical TDM-PON without cloudlets, for targeted latency figures of 1 ms, 10 ms and 100 ms.

The rest of this paper is organized as follows. In Section II, the details of TDM-PON based cloudlet network are described. In Section III, the system model and the MINLP formulation are presented. Simulation results are compared and discussed in Section IV. Finally, a summary of the main findings of this work is presented in Section V.
II. VISION OF OPTICAL CLOUDLET NETWORK ARCHITECTURE

Fig. 1 illustrates the envisioned end-to-end physical infrastructure of cloudlet networks based on TDM-PON access. Cloud servers are typically placed at a remote, geographically secured location towards the core network beyond backhaul metro networks. Cloudlets in comparison are located in close proximity to the end-user. As illustrated in Fig. 1, cloudlets are placed in the field with direct connections to suitably chosen ONUs. In a conventional wireless network, all WAPs are connected with the aggregation node by star topology [8], but, with a TDM-PON based access network, it can be either ring, bus or tree-branch. In Fig. 1, a tree-branch topology is chosen for illustration purposes. The power splitting ratio of each TDM-PON shown in Fig. 1 is configured to support the required population density and network coverage [6]. However, for greater user coverage, the TDM-PONs can serve as optical backhaul or fronthaul for WAPs, so that compatibility with future technologies like 5G, LTE-A, FiWi etc. is maintained.

A. Cloudlet network architecture with cloudlets in field

The cloudlet network architecture under our consideration, is based on the architecture proposed in [5], where cloudlet servers are placed in the field and the detailed picture is shown in Fig. 2. A key feature of such a network is that all cloudlets are orchestrated via M optical line terminal units (OLTs) located at the CO or equivalently, the aggregation node. Each CO is connected to the cloud server via a core node and backhaul network and each OLT is connected via a feeder fiber (maximum 10 km) to the passive splitter situated at the remote node (RN). In turn, the splitter can have a split ratio 1:4 or 1:8 (maximum 10 km) to the passive splitter situated at the remote CO is connected to the cloud server via a core node and backhaul network, and each OLT is connected via a feeder fiber (maximum 10 km) to the passive splitter situated at the remote RN. In turn, the splitter can have a split ratio 1:4 or 1:8 (maximum 10 km) to the passive splitter situated at the remote RN. In turn, the splitter can have a split ratio 1:4 or 1:8 (maximum 10 km) to the passive splitter situated at the remote RN.

Each cloudlet contains one or more racks of virtual machines (VMs). Up to 2500 VMs can be supported by each rack in the cloudlet, therefore, the higher the number of racks, the higher the number of VMs [8]. At the same time, each cloudlet is connected to multiple ONUs via point-to-point fiber. Each of these links, indicated in brown in Fig. 2, can support up to 1 Gbps. Also, as per definition of cloudlets in [1], the connection between the cloud and each cloudlet is mandated. To facilitate this, additional point-to-point fiber links (indicated in green in Fig. 2) between each cloudlet and CO, are implemented. More importantly, these links facilitate VM migration, as cloudlets can coordinate among themselves through the CO.

III. SYSTEM MODEL

The system model and MINLP formulation are presented here. To design a cloudlet network, we begin with an existing optical access network and then pre-identify a set of possible field cloudlet locations by using some clustering methods. A summary of network optimization parameters and their definitions is outlined in Table I.

A. Network Optimization Parameters

In this work, we consider cloudlets as M/M/1 queueing systems having $m \in K$ racks operating in parallel and service rate of each rack is $\mu$. Note that ideally the system should be $M/M/m$ queue if complete parallel computing is not possible. However, $M/M/1$ queue provides a lower limit on latency for $M/M/m$ queue as long as the single server has aggregated service rate of $m$ servers. The latency values are very close at high load with higher utilization and only deviates at lower load. Since cost optimization is performed in this work, it is expected that the system will be operating at a higher load and therefore this assumption is valid. The parameter $m$ is chosen as a variable so that the number of racks can vary from cloudlet to cloudlet. If a particular site $a \in A$ has not been chosen as a cloudlet location, its number of racks is zero. Further, the total incoming service request rate at a cloudlet $(\lambda_a)$ is the sum of service requests arriving from all ONUs $(\lambda_b)$ connected to that cloudlet. All the activated cloudlets have option to process the total incoming requests locally or offload a certain fraction of it to the remote cloud. Decision to offload tasks to remote cloud helps to minimize the overall cost but meeting the latency constraint becomes challenging. The average processing latency at cloudlet is expressed as follows [9].

$$d_{\text{cloudlet}} = 1/(\mu - \lambda_a)$$

We assume that incoming VM requests are homogeneously processed by all the $m$ racks. Thus, the total latency against service requests from all $b^{th}$ ONUs at $a^{th}$ field cloudlet can be expressed as sum of transmission latency and processing latency. The core cloud is assumed to be $M/M/\infty$ queuing system. If a service request is offloaded to the remote cloud, the total latency can be expressed as follows [10].

$$d_{\text{cloud}} = \Lambda + 1/\mu$$

B. Decision Variables

We choose a set of binary and fractional decision variables to formulate the MINLP model as follows. The parameter $x_a$ is a binary variable to indicate our decision to install a cloudlet at site $a \in A$.

$$x_a = \begin{cases} 
1; & \text{if a cloudlet is installed at location } a \in A \\
0; & \text{otherwise} 
\end{cases}$$

The parameter $n_{am}$ is a binary variable to indicate whether $m$ number of racks are chosen at cloudlet $a \in A$.

$$n_{am} = \begin{cases} 
1; & \text{if the cloudlet at } a \in A \text{ contains } m \text{ racks} \\
0; & \text{otherwise} 
\end{cases}$$

To linearize the objective function, we combine the product term of two binary variables $x_a$ and $n_{am}$ into another binary variable $\bar{x}_a (= n_{am} x_a)$.

The parameter $x_{ab}$ is a binary variable to indicate the connectivity between cloudlet $a \in A$ and ONU $b \in B$, which forms the adjacency matrix of the network.

$$x_{ab} = \begin{cases} 
1; & \text{if ONU } b \in B \text{ is connected to cloudlet } a \in A \\
0; & \text{otherwise} 
\end{cases}$$

The parameter $p_a$ is a binary variable that indicates if the total incoming task request is processed entirely by a cloudlet or only a fraction of it is processed by cloudlet and the rest is offloaded to the remote cloud.

$$p_a = \begin{cases} 
1; & \text{if task request is processed completely at cloudlet } a \in A \\
0; & \text{otherwise} 
\end{cases}$$
The parameter $\phi_a$ is a fraction that indicates the fraction of the incoming task processed at cloudlet $a \in A$ and lies within $[0, 1]$. Each cloudlet offloads the remaining tasks, i.e., $(1 - \phi_a)$ to the remote cloud.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>Set of potential field cloudlet locations</td>
</tr>
<tr>
<td>$B$</td>
<td>Set of locations of ONUs in the existing TDM-PON network</td>
</tr>
<tr>
<td>$K$</td>
<td>Set of number of racks in a field cloudlet within range $[1, 100]$</td>
</tr>
<tr>
<td>$\alpha_a$</td>
<td>Cost of installing a cloudlet at site $a \in A$ with one rack</td>
</tr>
<tr>
<td>$\xi_2$</td>
<td>New infrastructure setup cost of installing cloudlet at $a \in A$</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Installation cost of optical fiber per kilometer</td>
</tr>
<tr>
<td>$I_{max}$</td>
<td>Maximum allowed optical fiber length between cloudlet $a \in A$ and ONU $b \in B$</td>
</tr>
<tr>
<td>$L_{ab}$</td>
<td>Optical fiber length between cloudlet $a \in A$ and ONU $b \in B$</td>
</tr>
<tr>
<td>$BW_{ab}$</td>
<td>Bandwidth of the link between cloudlet $a \in A$ and ONU $b \in B$</td>
</tr>
<tr>
<td>$R_{max}$</td>
<td>Maximum supported bandwidth of the link between cloudlet $a \in A$ and ONU $b \in B$</td>
</tr>
<tr>
<td>$D_{ab}$</td>
<td>Transmission latency between cloudlet $a \in A$ and ONU $b \in B$</td>
</tr>
<tr>
<td>$D_{QoS}$</td>
<td>Maximum allowed latency to maintain desired QoS</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Service rate of each homogeneous rack of a cloudlet</td>
</tr>
<tr>
<td>$\lambda_a$</td>
<td>Total task arrival rate at cloudlet $a \in A$ for all $b$ ONUs connected to cloudlet $a$</td>
</tr>
<tr>
<td>$\lambda_b$</td>
<td>Total task arrival rate from a ONU at $b \in B$</td>
</tr>
<tr>
<td>$\mu_a$</td>
<td>Total service rate at cloudlet $a \in A$ containing $m \in K$ racks</td>
</tr>
<tr>
<td>$\Lambda$</td>
<td>Average latency to transmit a service request to remote cloud</td>
</tr>
</tbody>
</table>

### C. Objective Function and Constraints

Network operators are typically focused on minimizing capital and operational expenditure whilst maximizing revenue based on an expected number of users connected within the network. Driven by this, we formulate the objective function in (3) to minimize cloudlet installation and fiber cost (CAPEX) whilst maintaining the strict latency limit.

$$\min \left( \sum_{a \in A} \sum_{k \in K} \left( m \alpha_a \xi_2 + \sum_{a \in A} \sum_{b \in B} \eta L_{ab} x_{ab} + \sum_{a \in A} \xi_2 x_a \right) \right)$$

(3)

To linearize the product of two binary variables in the first term of the objective function ($\bar{x}_a = n_{an} x_a$) in (3), the constraint shown in (4) is invoked.

$$\bar{x}_a \leq n_{an} \bar{x}_a, \bar{x}_a \leq x_a, \bar{x}_a \geq n_{an} + x_a - 1, \forall a \in A, m \in K$$

(4)

The constraints described in (5)-(6) ensures that only one value of $m$ (the number of racks in a particular cloudlet) is chosen. As discussed in section III.A, if a particular cloudlet is not activated at location $a \in A$, then the corresponding number of racks is zero for that cloudlet.

$$x_a \leq \sum_{m \in K} n_{an}, \forall a \in A \tag{5}$$

$$\sum_{a \in A} n_{an} \leq 1, \forall a \in A \tag{6}$$

Constraints described in (7) and (8) ensure that every ONU $b \in B$ is connected to only one cloudlet $a \in A$. If $L_{ab} \geq L_{max}$, then $x_{ab} = 0$ and this constraint is ensured by (9).

$$\sum_{a \in A} x_{ab} = 1, \forall b \in B \tag{7}$$

$$x_{ab} \leq \max \left[ 0, \left( \frac{\left( L_{max} - L_{ab} \right)}{\left( L_{max} - L_{ab} \right)} \right), \forall a \in A, b \in B \right) \tag{8}$$

The constraint described in (10) denotes that bandwidth of the link between each ONU and corresponding cloudlet is always more than the requested bandwidth $B_{req}$ but less than $B_{max}$.

$$BW_{ab} \geq B_{req} \text{ and } BW_{ab} \leq B_{max}, \forall a \in A, b \in B \tag{9}$$

We calculate the arrival rate at cloudlet $a \in A$ as $(\lambda_a = \sum_{b \in B} x_{ab} \lambda_b)$ and service rate as $(\mu_a = \sum_{a \in A} mn_{an} \mu)$. The next set of constraints described (11)-(12) provides the lower and upper bounds of $p_a$, where $\Gamma$ is a large constant number.

$$p_a \geq \left( \lambda_a - \mu_a \right) / \Gamma, \forall a \in A \tag{11}$$

$$p_a \leq \left( \left( \lambda_a - \mu_a \right) / \Gamma \right) + 1, \forall a \in A \tag{12}$$

Finally, the constraint described in (13) provides the upper limit of service request at each cloudlet for all ONUs.

$$p_a \left[ \left( \left( \mu_a \mu_{\phi_a} + x_{ab} D_{ab} \right) + (1 - p_a) \right) \left( \phi_a \left( \left( \mu_a \mu_{\phi_a} + x_{ab} D_{ab} \right) + (1 - \phi_a) \right) / \mu_a + d_{\text{round}} \right) \right] \leq D_{QoS}, \forall a \in A, b \in B \tag{13}$$

In (13), the term multiplied with $p_a$ represents the sum of processing latency at cloudlet and the transmission latency. The term multiplied with $(1 - p_a)$ represents the sum of processing latency at cloudlet and the transmission latency for the fraction of task $\phi_a$ and sum of processing latency at cloud and transmission latency for rest of the tasks. In summary, (13) ensures that total latency from transmission and processing does not exceed the maximum allowed latency to maintain a desired QoS $D_{QoS}$.

The most suitable choice for the cost optimization framework is to execute the majority of the incoming service requests at cloudlet and offload a very small fraction to remote cloud. Therefore, without loss of generality, we can consider the value of $p_a$ to be zero and hence we can easily move the terms from denominator to numerator and can linearize the product terms of binary variables as objective function in (3). Thus, we have an optimization problem with a linear objective function and linear constraints, which we can nomenclate as field cloudlet placement programming (FCPP) problem.

### IV. FRAMEWORK EVALUATION

#### A. Dataset

In order to validate our FCPP model, implemented with the A Modeling Language for Mathematical Programming (AMPL) platform and the COUENNE package for MINLPs [11], we consider cloudlet deployment scenarios over urban, suburban and rural areas with population densities of 4000, 2500 and 1000 per square kilometre, respectively, as shown in Fig. 3. Due to the lack of readily available spatial demographic data, we adopt commonly used statistical models such as the Poisson point process to identify ONU locations for city areas with different population densities [12]. We consider a 10 km x 10 km total area that consists of 100 small grids of dimension 1 km x 1 km. The random population of each small grid is determined using Poisson distribution and those many points are then uniformly distributed within them. Next, $k$-means clustering algorithm [12] is used to identify possible field cloudlet locations. In Fig. 3, each blue dot indicates an ONU, capable of serving around 1000 users and the red asterisks
indicate field cloudlet locations. The position of red asterisks are dependent on the value of $k$ used in $k$-means algorithm. We also consider that these ONUs are supported by 10G-PONs and a network span of to 10 km. This enables each ONU to enjoy 10 Gbps downlink and 2.5 Gbps uplink bandwidth.

We consider the normalized cost of installing a cloudlet rack $\alpha$ is 1 [13]. The point-to-point links between each ONU and corresponding cloudlet has a maximum bandwidth support ($B_{\text{max}}$) of 1 Gbps. The average request size from each user can vary from 0.2 Mbits to 2.0 Mbits, depending on nature of application [4]. The task arrival rate from each ONU, serving 1000 users can be maximum 1000 and for our static analysis we consider $\lambda_{\text{a}}$ as 1000 VMs/sec. The service rate of each rack ($\mu$) is chosen to be 2500 VMs/sec. We arbitrarily choose total 30 clustering points by $k$-means clustering algorithm on 10 km x 10 km grid. Thus to maintain the feasibility of optimal solution, the value of maximum allowed link length between ONU and cloudlet $L_{\text{max}}$ has to be greater than 50% of the dimensions of the grid and we choose it as 7 km. The normalized costs of fiber link installation per km ($\eta$) are 2.5, 1.75 and 1 for urban, suburban and rural areas, respectively [14]. The normalized additional cost of installing a cloudlet ($\xi$) is 3 [13]. The average latency to offload a task request to remote cloud is ($\Lambda$) 0.8 sec [10]. Using these values, we solve the FCPP problem against three values of QoS latency ($D_{\text{QoS}}$) constraint i.e., 1 ms, 10 ms and 100 ms.

### B. Optimal Solution Evaluation

We use our FCPP model on the dataset shown in Fig. 3 to identify optimal cloudlet installation locations out of all possible identified locations. In Fig. 4, we show the optimally selected cloudlet placement locations for $D_{\text{QoS}} = 1$ ms. Similarly, optimal cloudlet placement locations can also be identified for $D_{\text{QoS}} = 10$ ms and $D_{\text{QoS}} = 100$ ms. However, note that different cloudlets have different number of total racks. Parameters like $\alpha$, $\eta$, $L_{\text{max}}$ and $\Lambda$ play major roles in impacting the optimal locations, whereas $D_{\text{QoS}}$ and total number of ONUs connected to a cloudlet mainly responsible for choosing optimal number of racks in each cloudlet. Thus, varying these parameters from the values we have chosen here, will lead to different types of optimal solutions.

### C. Results and Discussions

The normalized cost values per 1000 users in the optimized urban, suburban and rural areas, plotted for $D_{\text{QoS}} = 1$ ms, 10 ms and 100 ms are shown in Fig. 5. Note that a unit cost refers to the fiber link installation per km for rural areas [13]. As described in (3), these costs take into account only the additional cost of installing the cloudlet network. Results in Fig. 5 highlight that the higher fiber installation cost per km in the urban areas as compared to that in suburban and rural areas, result in an overall higher cost in the former. The optimization framework aims to offload the arriving tasks as much as possible to remote cloud, but, when the latency requirement is very tight, the solution therefore converges to one that would
require these tasks to be performed at the cloudlets. Consequently, a higher number of cloudlets and possibly with a higher number of racks is needed to meet stringent latency constraints. Installing the same number of cloudlets in an urban area will cost more than in suburban and rural areas.

Fig. 6 shows the average number of racks required per cloudlet for urban, suburban and rural scenarios against $D_{lat} = 1$ ms, 10 ms and 100 ms. In all deployment scenarios considered, the number of racks required per cloudlet is highest for the 1 ms latency constraint and lowest for the 100 ms latency constraint. It is also evident from Fig. 6 that as the VM serving capacity per cloudlet rack remains unchanged for all deployment scenarios, a lower number of racks will be required to serve all users in rural areas, whereas more racks will be required to serve all users in urban areas. It is also evident that the difference between the number of racks per cloudlet required in the urban and suburban scenarios is lower than the difference between urban and rural areas. This is due to the lower cost of fiber installation per km and lower population density in rural area than urban and suburban areas.

Finally, Fig. 7 shows the percentage of incremental energy consumption arising from installing cloudlets over a conventional 10G-PON infrastructure [13]-[14]. The network infrastructure we have considered is based on [5]. For every field cloudlet we activate, the incremental energy increases and the higher the number of racks, the higher the consumed energy. For all deployment scenarios considered, the percentage of incremental energy decreases as the latency constraint is relaxed. This is due to the fact that a lower number of racks per cloudlet can process the incoming service requests. Interestingly, we observe that the percentage of incremental energy is the highest for rural areas as compared to suburban and urban areas, for all latency constraints considered. In rural areas, the number of ONUs are lower than that of the suburban and urban areas and hence, the energy budget of the overall network is also lower. Thus, activation of cloudlets in a rural area will result in the highest incremental percentage of energy as compared to the urban and suburban areas.

V. SUMMARY

In this work, we proposed a novel optimization framework for field cloudlet placement over existing PON infrastructures. We have formulated a problem that can simultaneously optimize cloudlet placement locations as well as number of racks on each active cloudlet. Using this framework on diverse deployment scenarios e.g., urban, suburban and rural, cost optimality for different latency constraints is achieved for the first time. In summary, our framework is a first in yielding insights that will serve as a foundation for further cloudlet network planning strategies.

REFERENCES