Traffic-Profile and Machine Learning Based Regional Data Center Design and Operation for 5G Network

Udita Paul, Jiamo Liu, Sebastian Troia, Olabisi Falowo, and Guido Maier

Abstract: **Data center in the fifth generation (5G) network will serve as a facilitator to move the wireless communication industry from a proprietary hardware based approach to a more software oriented environment. Techniques such as Software defined networking (SDN) and network function virtualization (NFV) would be able to deploy network functionalities such as service and packet gateways as software. These virtual functionalities however would require computational power from data centers. Therefore, these data centers need to be properly placed and carefully designed based on the volume of traffic they are meant to serve. In this work, we first divide the city of Milan, Italy into different zones using K-means clustering algorithm. We then analyse the traffic profiles of these zones in the city using a network operator's Open Big Data set. We identify the optimal placement of data centers as a facility location problem and propose the use of Weiszfeld's algorithm to solve it. Furthermore, based on our analysis of traffic profiles in different zones, we heuristically determine the ideal dimension of the data center in each zone. Additionally, to aid operation and facilitate dynamic utilization of data center resources, we use the state of the art recurrent neural network models to predict the future traffic demands according to past demand profiles of each area.**

Index Terms: **Big data, cellular traffic, data centers, recurrent neural networks, traffic prediction, 5G.**

I. INTRODUCTION

The next evolution of wireless networks is confirmed to be the 5G. Its deployment has been necessitated due to the immense surge in data demand that the wireless communication sector has witnessed in recent years. Demand for data is forecast to reach close to 50 Exabytes per month in the year 2021 [1]. This demand would be a 7-fold growth from the data that was consumed per month in 2016. The expenditure (capital and operational) associated with present wireless service providing infrastructure would simply be astronomical if this forecast demand is to be met. Along with traditional users of the mobile broadband internet, 5G will involve consumers from other vertical industries such as automotive, health, energy and other industries. To meet demands of this diverse clientele, 5G is set to incorporate a concept called the 'network slicing' that will allow the network operators to provide dedicated virtual networks with functionality specific to particular services or customers over a common network infrastructure. Thus it will be able to support the numerous and varied services envisaged in 5G.

Network slicing would enable operators to separate a physical network into multiple virtual networks tailored to meet requirements of different user groups [2]. As the current mobile core network mainly consists of hardware dependent network functions, flexible and scalable way of providing service becomes an issue. Two main concepts in form of NFV and SDN have come to the forefront to aid the transition of the core architecture to a more softwarized domain. NFV allows network functions such as service gateway to be deployed as virtual network functions (VNFs) implemented in form of software on commercial off the shelf (COTS) hardware (such as servers in data centers) [3]. This allows service providers to be less dependent on hardware and aids seamless scaling in case changes are needed to be made. This further reduces the operator's expenses associated with establishment of the core architecture. SDN [4], on the other hand splits the control and data plane of the network functionalities thereby producing a programmable environment that greatly simplifies network management. SDN allows service providers to have more control over the various moving pieces of a software based network.

Data centers hosting various VNFs are likely going to be placed at the edge of the current networks to supplement those deployed in traditional cloud data centers [5]. These edge/mini data centers would be geographically distributed in different regions and could potentially be located at the point of presence (PoP) level [6]. These regional data centers would also possess smaller capacities in terms of storage, networking and compute resources in comparison to significantly larger data centers deployed by corporations such as Amazon [7]. Fig. 1 shows the higher level physical infrastructure of 5G networks [8] and illustrates the vital role these mini data centers would play in in processing the traffic in different regions/zones. Adequate planning, therefore, needs to take place to determine the optimal positioning and dimensioning of these data centers. Existing research works either focus on the problem of determining resource requirements and placement of individual VNFs [9]–[11], or architectural design of data centers [12]–[14]. The dimensioning and placement problem of data centers were addressed in [15] with the aid of optimization models. However, a key element in form of cellular traffic that originates from users is not considered in the design of data centers in the literature. As 5G will involve different types of subscribers, each data cen-

Manuscript received May 23, 2019; approved for publication by Young-June Choi, Division III Editor, October 20, 2019.

U. Paul, J. Liu, and O. Falowo are with Communication Research Group, University of Cape Town, South Africa, email: {plxudi001, lxxjia007, olabisi.falowo}@uct.ac.za.

S. Troia and G. Maier are with Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano, Italy, email: {sebastian.troia, guido.maier}@polimi.it.

U. Paul is the corresponding author.

Digital object Identifier: 10.1109/JCN.2019.000055

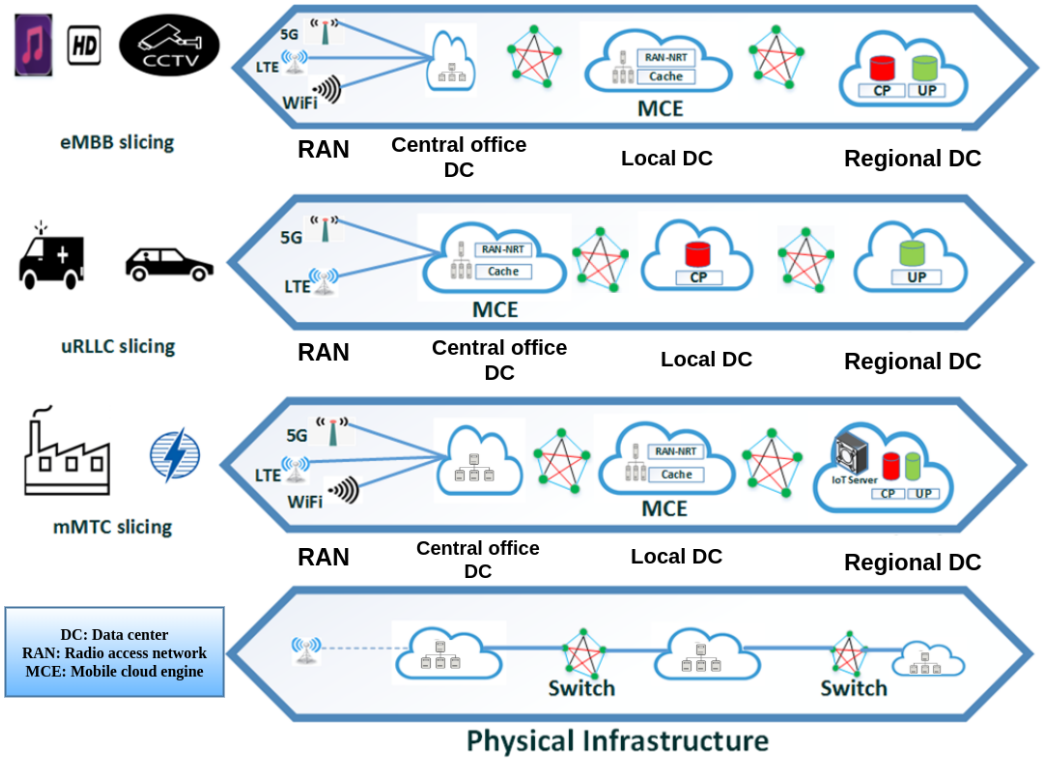


Fig. 1. Network slicing and Data Centers in 5 G Networks

ter needs to be designed based on the volume of traffic it will be required to deal with to guarantee optimal performance. In addition, to ensure service requests are met within the delay constraints, the data centers need to be also optimally located from the base stations. Furthermore, to guarantee dynamic utilization of data center resources, accurate prediction models need to be employed to forecast future traffic demands.

In this work, we exploit the open telecommunications data set of Telecom Italia (TIM) to first study the data traffic profiles of different regions of the city of Milan. We process the big data set and with the aid of K-mean clustering algorithm separate the city into twenty regions. This iterative algorithm assigns different numbers of base stations to various clusters with the aim of reducing the overall aggregated distance. We propose establishment of a data center in each region to provide computational power to various NFV and SDN functionalities of one or multiple network operators. These small sized data centers would be in place to handle the traffic experienced by the base stations in their vicinity to provide better quality of service to the end users. Based on the traffic experienced, we evaluate a weight to be assigned to each TIM base station within a region. We utilize the weights of these base stations to heuristically determine the ideal positioning of a data center within a region. This ensures that the data centers are located closer to the base stations that experience the most traffic, while also catering for the requirements of other base stations. Furthermore, we analyse the hourly traffic pattern within each region under consideration and heuristically determine the dimension of the data centers in terms of the number of CPU cores needed to handle the traffic within their coverage. This analytic approach towards determining the size of the data centers would incur less cost to the infrastructure provider while establishing these facilities. Additionally, based on the traffic pattern and previous information, we employ machine learning algorithms to forecast the next days' traffic within each region under consideration. With the aid of these forecast values, data center's resource manager (usually a SDN controller) would be able to determine the amount of computational power that will be needed to handle the future traffic from different wireless service providers, in advance. This would greatly reduce the operational cost for the infrastructure providers and ensure optimal utilization of data center resources. The main contributions of this work can be summarised as follows:

- Analytically divide the city of Milan into twenty different regions/zones.
- Conduct a detailed analysis on the cellular traffic dataset of Telecom Italia (TIM).
- Determine heuristically the ideal placement of a regional data center according to the traffic handled by different base stations within that region.
- Determine the overall capacity of a data center based on a heuristic approach.
- Apply recurrent neural network (RNN) models to the TIM data set to predict next day's traffic demand on the considered regions.
- Validate and compare the performances of different RNN models in predicting the next-day traffic profile.

The rest of the paper is organized as follows. Section II provides an overview of the background and work related to the fields of interest of this work. In Section III, we analyse the data set that has been used in this work and provide the methodology

employed in dividing the city into different zones. Section IV presents a detailed analysis of the traffic profiles of the three chosen zones from the data set. This is followed by Section V that presents the design of the data centers. In Section VI, machine learning algorithms are employed and evaluated to predict future traffic profiles. Finally, conclusions are presented in Section VII.

II. BACKGROUND AND RELATED WORK

The literature related to this work can be classified into two areas. The first area focuses on the traffic analysis and prediction models that have been proposed in the literature. The second area is related to the design of various components of the cloud infrastructure in cellular and wired network.

A. Traffic Analysis and Prediction in Cellular Network

A lot of work has thus far been conducted to determine the traffic profile that a single or group of base stations experience over a period of time. Given the proliferation of bandwidth intensive applications, cellular data trace reveals more related to network behaviours than traditional voice traffic. Furno *et al.*[16] developed a cognitive framework to analyse traffic profiles using the TIM and Orange open data sets. The idea behind this work was to identify anomalies that usually occur while considering network-wide usages. Wang *et al.*[17] studied the traffic experienced by multiple geographically separated base stations to demonstrate the strong spatial-temporal relationship that exists in the domain of cellular traffic. Sinusoidal superposition and log-normal distribution methods were employed to describe the temporal and spatial traffic variations respectively. Troìa *et al.*[18] proposed a novel method to identify typical traffic patterns exploiting matrix factorization methods. They were able to extract typical 24 hours patterns experienced by the mobile network in Milan city. The learning and prediction model proposed by Li *et al.* in [19] used a big data set to study the parametric differences that exists between different types of cellular network applications. They further proposed a predictive algorithm to make forecasts related to application-level traffic. The same authors in [20] used entropy theory to demonstrate the inherent pattern that exists in cellular traffic. They also concluded that the data traffic prediction is solely reliant on temporal and spatial relevance.

Among tools used in the prediction of traffic pattern in cellular networks, few have been commonly used. Linear models such as auto-regressive integrated moving average (ARIMA) and its modified version fractionally- ARIMA (FARIMA) have been used in [22],[23]. Kalman filtering has been used in [24],[25] with mobility and network traffic model utilized in [26],[27]. These shallow learning architectures however, have proven incapable of accurately modelling deep and complex non-linear relationships that are usually present in cellular traffic traces. Recently, deep learning-based predictive algorithms have emerged and proven effective in the prediction of traffic pattern. Oliveira *et al.* in [28] compared the performances of recurrent neural network (RNN) with stacked auto-encoder to forecast internet usage. Their analysis showed that RNN is superior to auto-encoder in making accurate prediction in this use case. The work in [29] used two state of the art neural network models: RNN and convolutional neural network (CNN), alone and in conjunction with each other to forecast maximum, average and minimum traffic volume of different regions in the TIM dataset. Their evaluation presented prediction accuracy between 70 to 90 percent. The work in [30] compared the performances of neural network models and linear models to make network traffic predictions with neural network models outperforming others.

The existing literature in the domain of traffic prediction and analysis, either used models that have proven to be less effective with complex data or they have not performed hourly time series analysis with state of the art RNN models. While many form of statistical tools have been used to perform time series analysis and predictions in the literature, the recurrent neural network has demonstrated superior performance over others [31]. RNN in its architecture contains a recurring loop which allows it to combine present information with the past. This makes RNN and its models suitable for analysis of time series sequence. As cellular traffic is highly dependent upon time, RNN's ability to recognize patterns in such data is highly desirable. In addition, the more recent models of RNN have significantly enhanced the performance of the simple RNN model. However, the effect of activation functions on the prediction accuracy of the RNN models has not gained much attention in the literature. In this work, we analyse the performances of different RNN models with activation functions to obtain future traffic demands.

B. Design and operation of Cloud Infrastructure

Dominicini *et al.* in [32] designed a NFV oriented architecture for edge data centers. They implemented a server centric data center architecture that produced better results when compared to traditional network-centric architectures. The authors of [14] proposed optimization models to reduce the total energy consumption in both data centers and data center networks. Gebert *et al.* proposed potential solutions that accommodates a sudden increase in traffic demand with the aid of dynamic and optimal placements of various VNFs [33]. The work in [34] proposed modification on the existing CU algorithms to determine characteristics of different traffic that a data center network experiences. They validated the proposed algorithm by using a real data set and achieved improvement on error performance, space cost and time complexity. The problem of optimal placement of several VNFs was explored further in [35]. In this work, the authors proposed mathematical frameworks that determined the ideal location for several VNFs within the boundaries of network capacity and latency limits. Shi *et al.* in [36] combined a decision making approach with Bayesian learning to dynamically allocate computing cloud resource in data centers to various NFV components. In [15], the authors proposed an optimization model that addressed the issue of optimal placement of both SDN and NFV components along with ideal size of data centers. Their work considered network cost with load to determine the best location for various data centers in the cases of Germany and USA.

Several work have also been carried out to determine the capacity that data centers are required to be equipped with to be able to provide desired performance. The authors in [37] uti-

Table 1. Symbols and their descriptions

Symbol	Description
B	Set of base station coordinates
b	Total number of base stations
K	K-means centroid coordinates
k	Total number of clusters
O	Overall aggregated distance
d	Distortion value
Y	Established location of Data center
C	Base station cluster
$w_i z$	Weight of the i^{th} base station in zone z
q_z	Total number of base stations in zone z
A_z	Set of volumes of traffic of each station in zone z
v_{iz}	Volume of traffic of the i^{th} base station in zone z
x_{jz}, y_{jz}	Geographical coordinates of the j^{th} candidate location of the data center in zone z
d_{oj}	Distance between the candidate location of the data center in a zone to a point
V_z	Design Capacity of the data center in zone z
D_z	A capacity multiplier between $[1, 2]$ for the data center in zone z
$f(std)_z$	Corresponding standard deviation value of the hour which has maximum sum of mean and one standard deviation in zone z
$g(mean)_z$	Corresponding mean value of the hour which has maximum sum of mean and one standard deviation in zone z
$P_h(mean + 1std))_z$	Probability of traffic volume of the h^{th} hour in zone z to be less than the sum of mean and on standard deviation of the h^{th} hour
α	Capacity design constant, -0.6
P_s	Capacity design constant, 0.6

lized optimization framework to determine the optimal dimensions of several types of geographically distributed data centers. Their results demonstrated that key performance metrics such as latency and cost can be best satisfied if the data centers are sized and located optimally. In [38], the authors presented an analytical model to optimize cost while determining the ideal dimension of a mobile network operator's data center. Their proposed model was able to determine the optimal number of physical machines that a cloud data center would require while meeting up with different amount of subscriber's demand. Carvalho *et al.* in [39] proposed an admission control mechanism to determine the minimum capacity for a cloud infrastructure such as data center. Their results showed that with aid of different service level agreements between infrastructure providers and subscribers, admission control can be employed to reduce data center's capacity, thereby lowering capital and operational expenditure. In [40], authors employ deep learning techniques to perform predictions on network traffic experienced in data centers.

After reviewing the existing research works so far conducted in this area, we can observe that they do not focus on the design of data centers using real world cellular traffic. As suggested in the literature, adaptive utilization and placement of cloud computing resources and functionalities are essential to provide optimal services for the end users. As traffic pattern changes with space and time, it is therefore crucial to determine the traffic profile that exists in various regions within a geographical area. With proper analysis and forecast of traffic patterns in different locations, data centers can be optimally placed and their resources can be properly utilized. To the best of our knowledge, no existing literature has studied the TIM data set in this context, which is the main focus of our work. The symbols used throughout the paper have been listed in Table I.

III. DATASET ANALYSIS & REGION FORMATION

In this section, we first present the dataset utilized for the purpose of our work. We then present the algorithm used in forming various regions within the city to facilitate our design of data centers.

A. Dataset

The data used in this work was released by Telecom Italia in 2015 and has been made available for public use [41]. It contains call detail records (CDRs) of different areas over the period of November 1, 2013 to January 1, 2014 within the Italian city of Milan and Province of Trento. The data set breaks down the considered area into 10,000 square cells (geographically located), with each cell representing an area of 235m by 235m. The CDRs contain information related to each cell's different telecommunication activities such as number of calls, sms and internet activity within a ten minute period. As internet activities certainly demand more resources in today's wireless communication, for the purpose of this work, we only consider the amount of internet activities that occur within a cell in a given time frame. As the data set also contains 62 days worth of records, we separated the holidays (22 days) from the working days (40 days). Another important attribute of this data set is in its recording of the CDRs. Each entry of the internet activity in the dataset represents the number of times a connection is initiated or terminated. A new CDR is also registered every time a previous connection exceeds 5 Megabytes (MB) of traffic. Since we can neither identify exact number of new and terminated internet connections nor the exact volume of traffic exchanged within a

connection, we, therefore, assume half of the records to be new connections with each connection having a volume of 5 MB of internet traffic. The cells of the TIM data set are presented in Fig. 2.

9901	9902	9903	…	…	…	…	…	…	10000
9801	9802	9803	…	…	…	…	…	…	9900
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
101	102	103	…	…	…	…	…	…	200
1	2	3	…	…	…	…	…	…	100

Fig. 2. The cells in TIM dataset

This dataset does not reveal information about the geographical positions of the base stations in the Milan area. Therefore, we processed another dataset [42] that collects information about the base stations of Telecom Italia deployed in Milan. The results obtained from the analysis of this data set was used to match the traffic volume information provided in the TIM data set. By matching these two data sets we were able to obtain further information such as: the total number of TIM base stations in the Milan grid (2554), geographical coordinates of each base station and hourly number of CDRs experienced by each base station during the time period the dataset was formulated [18].

B. Region Formation

In order to analyse the mobile internet traffic that different areas in Milan experience, it is critical to divide the city into different zones. The division of zones needs to take into consideration the distribution of base stations within the city. Certain number of base stations need to be clustered to form a single zone. To achieve this clustering of base stations to create different zones within the city, we employ a popular clustering method known as the K-means clustering algorithm.

In this work, the objective of the K-means algorithm is to determine k number of centroids which are to be associated with a certain number of members which would result in least overall aggregated distance. Depending on the locations of the members, the positions of these centroids and their members are changed iteratively to obtain the best possible location for each centroid. The process is repeated until the membership of each centroid remains unchanged from previous iteration, resulting in the algorithm to converge to the local minima of the overall aggregated distance.

In our case, given a set of base stations B containing the b (2554) pairs of geographical coordinates (in form of latitude(x) and longitude (y)), we can assign K as the set containing the geographical locations of the k clusters' centroids. These sets can be represented as:

$$B = \{B_1, B_2, ..., B_b\}, \tag{1}$$

$$K = \{K_1, K_2, ..., K_k\}. \tag{2}$$

The overall aggregated distance, O, associated with k centroids and their members can be represented as:

$$O = \sum_{i=1}^{k} \sum_{B \in K_i} ||B - K_i||^2 \quad \text{for} \quad i = 1, 2, ..., k. \tag{3}$$

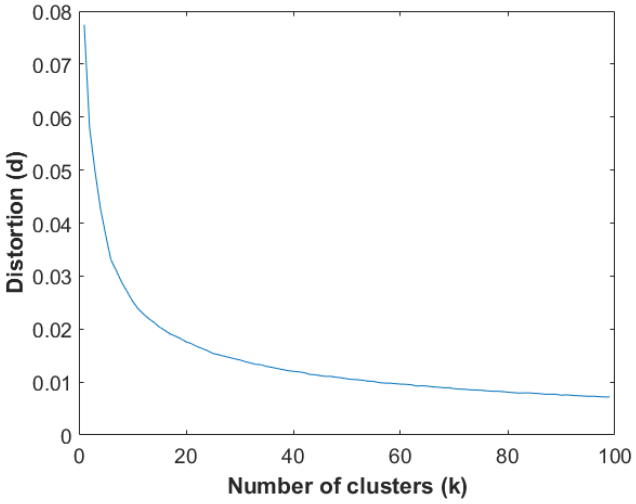


Fig. 3. The relationship between distortion values and number of clusters as obtained using Elbow

In order to determine the optimal number of clusters k, we experimented with different values of k. For each value of k, we obtained a distortion value, d, which is defined by $d = O/k$. The relationship is demonstrated in Fig. 3.

From Fig. 3, it can be seen that even as the value of k increases beyond 20, the value of d does not decrease significantly. As such, we use 20 as the value of k and thereby divide the base stations in the city to form 20 zones. For each of these 20 zones, we propose establishment of a data center to process the traffic that the base stations within a zone experiences. Fig. 4 shows the locations of the base stations within different zones in Milan. Entries of similar color and shape represent the base stations within a specific zone. For the sake of illustration, zones 6, 14 and 19 are labeled in Fig. 4. Fig. 5 demonstrates the architecture of our proposed model. Essentially, the proposed model depicted in Fig. 5 breaks the city down into 20 zones and for each zone, designs a location and capacity for a data center that will be charged with handling the traffic within that zone.

IV. TRAFFIC PROFILE ANALYSIS

In this section, we first choose to present three zones and analyse the traffic profiles in each of these zones. Fig. 6 shows the total volume of traffic that each of the 20 zones experience in the duration of time the dataset was formed. It can be observed from Fig. 6 that among these zones, zone 6 experiences the least amount of traffic while zone 19 experiences the highest volume of traffic. Zone 14 can be seen to have average amount of traffic. As such, we present the traffic profile analysis of these three zones.

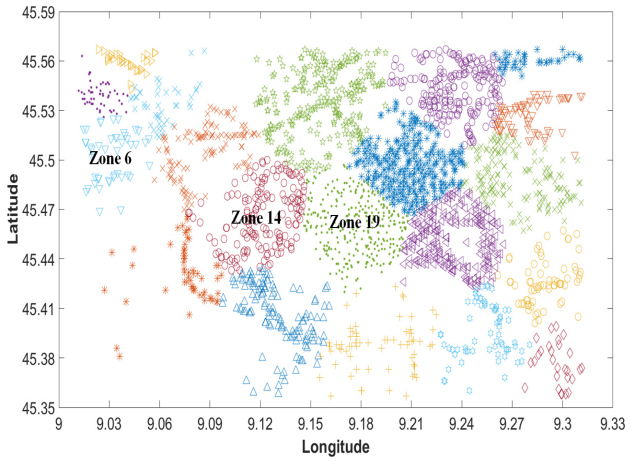


Fig. 4. Base station locations in the zones formed by K-Means Clustering in Milan. Locations of zones 6,14 and 19 are annotated.

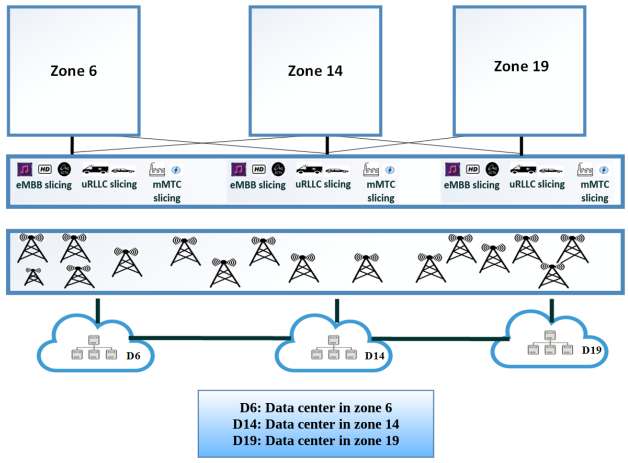


Fig. 5. Proposed Model illustrating the different zones, slices and data centers

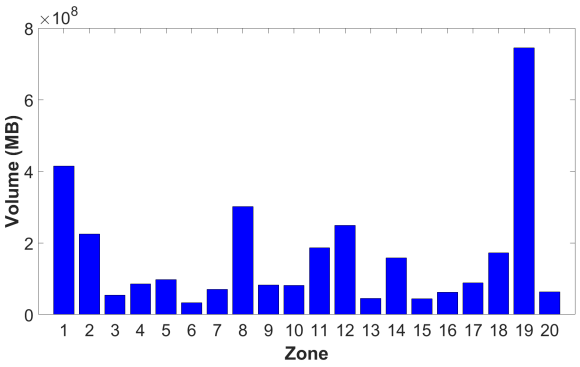


Fig. 6. Aggregated Traffic Volume in Each Zone formed using K-Means Clustering Algorithm

Based on the aggregated hourly data obtained in the Section III A, we determine the maximum and average hourly traffic volumes of these zones during holidays and workdays. In addition, we also present the sum of the mean and one standard deviation to illustrate the amount of variation in each hour of traffic. Despite the differences in the traffic profiles of these zones, there exists some common phenomena among them. The detailed analysis of the traffic profiles of these zones are presented below.

Common Characteristics of the Zones: Figs. 7 (a), (c) and (e) represent the holiday traffic profiles of Zone 6, 14 and 19 respectively. Conversely, (b), (d) and (f) of the same figure present the workday traffic profiles of these zones. Table VI, X and XIX in [43] further provide additional information that aid the analysis of traffic profiles of these zones. For both holiday and workday traffic, the presence of the 'tidal effect' is evident. Traffic is gradually seen to decline during late night hours (starting from around 8 pm) and the minimum is reached around 4 am. The traffic consumption is seen to gradually increase from early morning hours (5 am to 8 am) in both types of days. The physical quantity of the standard deviation observed in these traffic volumes also increases as the traffic volume increases. Another important quantity of interest is the probability with which the traffic volumes of past days tend to fall within the sum of the mean and deviations. It is observed that, in the cases of both holidays and workdays, the hourly traffic of past days mostly fall within the sum of the mean and one standard deviation with a probability of 75 percent and more. However, this probability increases to the range of 90 to 100 percent when the sum of the mean with two standard deviations is considered. This probability analysis enables us to understand the hourly variation in traffic volume which would further aid in the subsequent design of the dimension of the data centers. The inherent traffic characteristics of each of the three zones are explained below.

Zone 6: This zone experiences the lowest amount of traffic among all others. The zone is located in the outskirts of the city and is sparsely populated. Fig. 4 shows the geographical locations of the base stations that are within this zone. There are a total of 47 TIM base stations (BSs) within this area. This relatively low number of BSs highlights the low level of traffic this zone experiences for both holidays and workdays as seen from Figs. 7 (a) and 7 (b). Some key features associated with the traffic volumes in this zone are presented in Table VI in [43].

Zone 14: This zone experiences medium amount of traffic and covers areas that are somewhat in the center of the city. There are 166 BSs in this zone which is significantly more when compared to Zone 6. This is expected given the larger volume of traffic that is experienced in these areas over time. Figs. 7 (c) and (d) also present some traffic information with Fig. 4 presenting the geographical locations of the BSs within zone 14 . Table XIV in [43] presents some of the attributes noticed for the traffic in zone 14.

Zone 19: This zone experiences the highest volumes of traffic in comparison to others. This zone, as seen from Fig. 4, is in the heart of Milan and experiences heavy traffic volumes. Services and financial companies form this predominantly commercial zone. To meet up with the traffic demands in these areas, there are 309 TIM base stations located in this area. Figs. 7 (e) and (f) present some key statistics regarding the traffic experienced on holiday and workday in this zone. Fig. 4 also illustrate the locations of these BSs within zone 19 with Table XIX in [43] presenting some key features of this zone's traffic volumes.

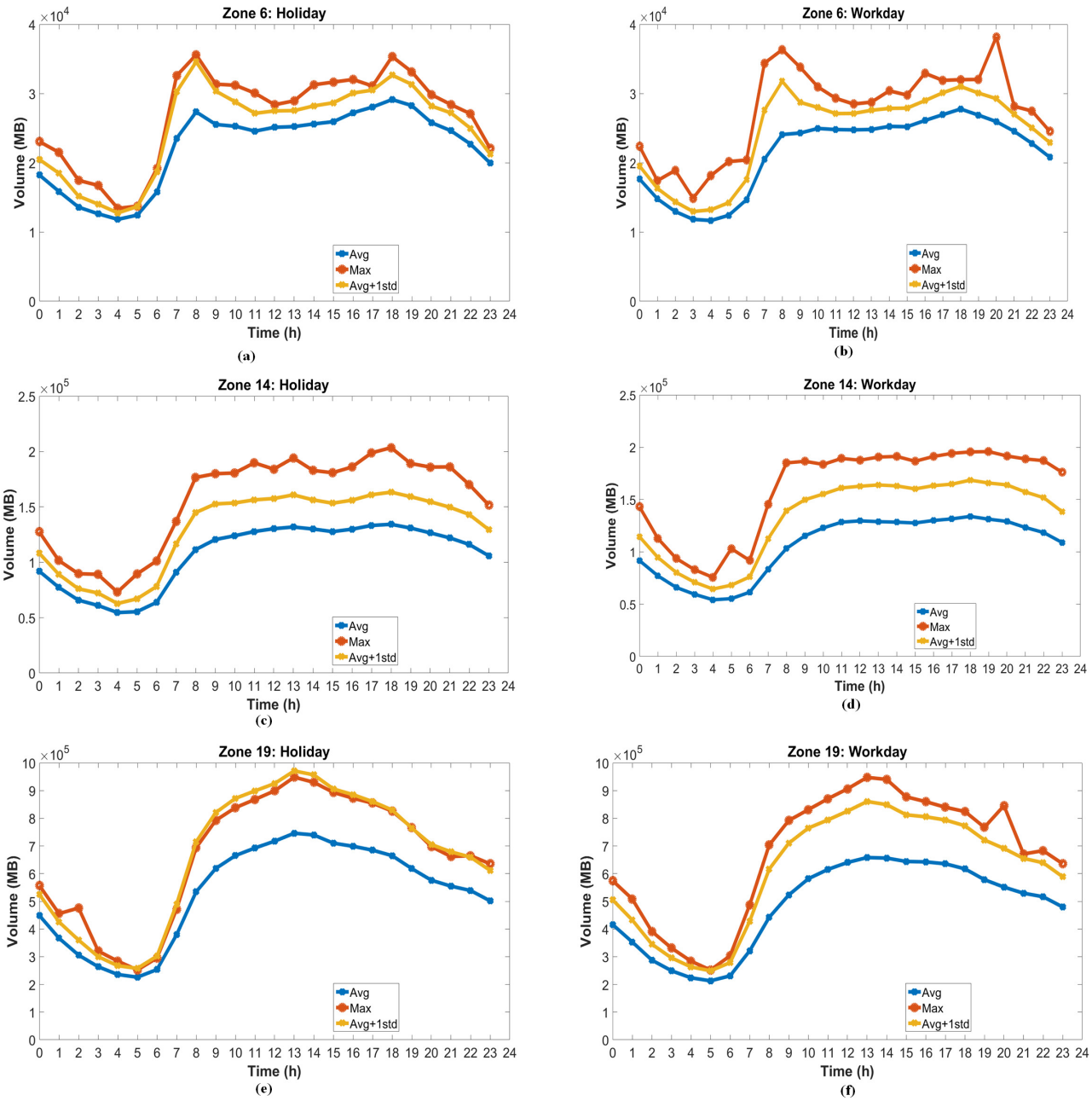


Fig. 7. Traffic profiles in Zones 6, 14 and 19 during holidays and workdays. The profile displays the average, maximum and average with one standard deviation traffic volume in the zones.

V. DESIGN OF DATA CENTERS

In this section, we use the locations of the base stations to first heuristically determine the ideal location for a data center in each of the considered zones. We then proceed to determine the dimension of the data center to meet up with the traffic demand from the zones.

A. Placement of the Data Centers

One consideration while determining the position of a data center is its distance from the base stations. Minimizing the aggregate distances between the data centers and the base stations would reduce the cost of front haul links and also would lower the delay in propagation. The problem of determining the ideal location for such a facility can be identified as the *Weber's problem* which also is a special case of the *Facility Location Problem* [44].

The aim of any facility location problem is to determine the most suitable place to establish one or multiple facilities in the presence of many candidate locations. The facilities are usually required to provide services to meet demands that are imposed by their customers (whose locations are known). In our case, the facilities are the data centers which provide the base stations (the customers) with computational power to process the traffic experienced in each base station. The Weber problem looks to reach a point that ensures that the weighted sum from the point to the known base stations' locations reaches its minimum

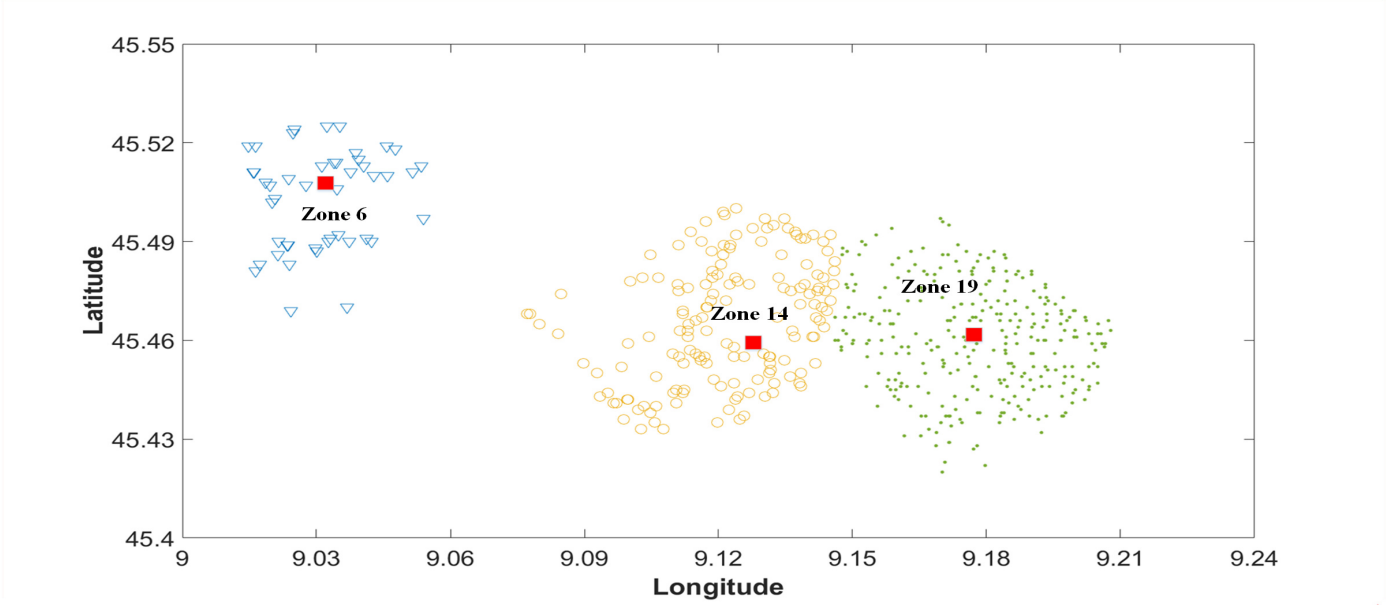


Fig. 8. Proposed location of Data Centers in Zones 6, 14 and 19. The data center in a zone is shown in red while the base stations in that zones are spread all over.

[45]. Since we are considering the distances of the base stations from the data center as the parameter to minimize, the following mathematical model can be employed to represent it:

$$\arg\min_{Y\in R^2} f(Y,\theta) = \arg\min \sum_{i=1}^{m} w_i \theta(C_i, Y), \tag{4}$$

here w_i denotes the weights assigned to the i^{th} base station, among m base stations, belonging to the base station cluster C. The cost function θ in this optimization framework aims to minimize the overall distance between the base stations in a particular cluster C from a candidate location for data center Y. Some base stations within a zone often experience significantly higher volumes of traffic than others. Therefore, it is logical to place greater weights on the base stations that experience heavier load and place the data center closer to these base stations. As we have 20 zones with each having a certain number, q, base stations in it, we can then define a set A_z that contains the volume of traffic of each base station within that zone z. A_z therefore can be represented as:

$$A_z = \{v_{1z}, v_{2z}, ..., v_{qz}\}, \tag{5}$$

where v_{iz} is the volume of traffic of the i^{th} base station in zone z. To determine the weight, w_{iz}, of the i^{th} base station within zone z, we use the following equation:

$$w_{iz} = 1 - \frac{max(A_z) - v_{iz}}{max(A_z) - min(A_z)} \quad \text{for} \quad i = 1, 2, ..., q. \tag{6}$$

The values of $max(A_z)$ and $min(A_z)$ are highest and lowest volume of the traffic that the most and least loaded base station within the zone z experiences respectively. Knowing the weight of each base station in a zone, we can utilize Weiszfeld procedure [46] to determine the data center location in each zone. We use this algorithm due to its proven efficiency as well as low computational complexity [47]. Weiszfeld algorithm is based on the gradient descent algorithm which essentially minimizes the sum of the weighted l_2 norm of each element of the base station group C_i and iterates to obtain the best possible location for the establishment of the data center.

As we consider three zones, each with its own boundaries in terms of latitude and longitude, the possible location for a data center in z^{th} zone has to be contained in a 2-dimensional vector space, J, which also includes the location of all the BSs within that zone . The algorithm begins at a random coordinate point having latitude (x_{1z}) and longitude (y_{1z}) and attempts to locate the optimal point within the set J to minimize the sum of Euclidean distances from the BSs within that zone. The x and y values are calculated using the formula:

$$x = \frac{\sum_{j\in J}(w_{jz}x_{jz})/d_{oj}}{\sum_{j\in J}(w_{jz})/d_{oj}}, \quad y = \frac{\sum_{j\in J}(w_{jz}y_{jz})/d_{oj}}{\sum_{j\in J}(w_{jz})/d_{oj}}, \tag{7}$$

where x_{jz} and y_{jz} represent the j^{th} candidate's location for the data centre in zone z with d_{oj} representing the distance between the candidate location for the data center to a point in set J. w_{jz} represents the weight assigned to the j^{th} base station in zone z. The iterations are continued until either a convergence is reached or if the maximum number of evaluations is completed.

With the aid of this algorithm, we determine the ideal location for each data center in Zones 6, 14 and 19. The location of the data centers among the base stations are illustrated in rectangular boxes in Fig. 8.

B. Dimension of the Data Centers

Once the data center is established, it becomes critically important to determine its dimension. This largely depends upon the traffic demand that the data center is expected to cater for. For the purpose of this work, as we only have the information of the single mobile operator (TIM), we design the size of the data centers based on the traffic volumes experienced by its BSs in the zones under consideration. In a real life design case, it is

expected that the infrastructure providers would lease their services to multiple operators and as such would require relevant information from other operators as well. Also, we only focus on the resources in terms of the computational power required to process the traffic in these zones.

The computational power is provided by servers within a data center. An area with large number of BSs would require higher amount of computational resources (CPU cores provided by the servers) to serve the traffic demands as well as to host various VNFs such as SDN controllers and virtual gateways. We assume that the VNFs for a particular zone are all hosted in a single centralized data center rather than being distributed all over. This approach requires less number of servers, and subsequently cores, as opposed to having a distributed VNF architecture [15]. VNFs that are intended to serve both data and control plane functionalities require more computational power than SDN controllers that deal with only control plane functionalities. The authors in [48] demonstrated that 20 cores of CPU processing power are required to handle 1 unit of data traffic demand (i.e. 1 Gbps) and only 6 cores are required by the SDN controllers. Therefore, their work shows that a total of 26 cores are needed to process 1Gbps of traffic load and overhead. As such, in our model, we adopt this specification from [48] to design the dimensions of each zone's data center based on the traffic profile analysis conducted previously.

While determining the processing power required by a data center (data center's capacity), we need to carefully evaluate the traffic profiles that the base stations under its coverage experiences. As we have only 62 days worth of data, certain traffic characteristics might not have been captured within this time frame. Allocating resources to meet just the maximum of peak-hour traffic would lead to over provisioning of resources that would remain underutilized most of the times. Similarly, having enough servers to meet only the average demand would lead to shortage of resources during peak demand hours thereby resulting in poor quality of services (QoS). Referring to Tables in [43], we can see that a good metric to determine the volume of traffic that a data center needs to be designed for can be based on the sum of the average and standard deviations of the traffic volume. The hourly traffic volume of zones had surpassed the sum of the mean and one standard deviation in considerable number of occasions. However, most of these volumes fell well within the sum of the mean and two standard deviations. The ideal traffic volume that a data center need to cater for, therefore, lies somewhere in the range between these two. We therefore heuristically determine that the ideal design capacity, V_z, of the data center in zone z to be:

$$V_z = D_z \times (f(std)_z + g(mean)_z), \tag{8}$$

$f(std)_z$ and $g(mean)_z$ are the corresponding values of the standard deviation and mean for the maximum hourly sum of mean and one standard deviation for a particular zone z. $D_z \in [1, 2]$ is the multiplier which aids in determining the maximum traffic volume, V_z, that the data center in the z^{th} zone would be capable of serving at any given time. This multiplier is inversely proportional to the probability of the h^{th} hour's traffic volume to fall within the sum of the mean and one standard deviation in zone z, $P_h(mean+1std)_z$. In this work, the D_z value is 1.6 when the $P_h(mean+1std)_z$ is 0.6 and D_z is 1 when $P_h(mean+1std)_z$ is 1. Then, D_z can be obtained using the following heuristically obtained mathematical relationship:

$$D_z = (1-\alpha) + \left(\frac{(max(P_h(mean+1std))_z + \alpha)) \times \alpha}{1 - P_s} \right), \tag{9}$$

where α for this dataset is -0.6 and P_s is 0.6 as mentioned above. $max(P_h(mean+1std))_z$ denotes the probability value for the hour that demonstrates the highest sum of mean with one standard deviation ($mean + 1std$) of traffic volume in that zone. For example, in zone 6, we can see from Table VI in [43] that the highest value of $mean + 1std$ is observed for the hour 8 (between 8 am and 9 am) holiday traffic which corresponds to a $P_8(mean + 1std)_6$ value of 0.95. Therefore, to determine the deviation factor in this zone, we use this value as our $max(P_h(mean+1std))_6$. With this, we can then determine the traffic volume that the data center in zone z needs to be designed for using equation 8.

As mentioned previously, based on the specification in [48], to process one unit of traffic i.e 1 Gbps, 26 cores of processing power is required. Given the aggregated nature of the data set we have, it is not possible to evaluate the demand volume that is experienced per second for each zone. Therefore, in this work, we assume that at every second, same amount of demand is generated resulting in a cumulative volume of V_z in the z^{th} zone per hour. Note also that the capacity of the front haul links between the base stations and the data center also play a crucial role in the processing of the traffic demand. This, however, is beyond the scope of this work. Using the above equations and the specifications, we can proceed to determine the capacity of each data center.

Zone 6 Data Center: This light traffic volume zone, as mentioned above, possesses $max(P_8(mean+1std))_6$ value of 0.95 based on the maximum of sum and 1 std that is noticed at the 8^{th} hour. Using equation 9, we obtain the multiplier of zone 6, D_6, to be 1.07. The corresponding mean and standard deviation values of this hour's of traffic are $27,350.04$ MB and $7,237.04$ MB respectively. Therefore, using equation 8, we obtain the value of $37,181$ MB (290 Gb) as the maximum volume of traffic that this data center would be required to handle at any given hour. Note that this value is greater than any of the peak hourly traffic for both working days and holidays based on the available data. This value is also much smaller than the sum of mean and two standard deviation value. As such, it is a value with ample tolerance to meet the highest traffic demand that might be encountered in this zone. Using our assumptions and specifications, we evaluate that the data center for this zone would require maximum of 2 cores to process the traffic volume for the TIM subscribers in this zone.

Zone 14 Data Center: Zone 14's medium level traffic has a $max(P_h(mean + 1std))_{14}$ value of 0.8 corresponding to the 18^{th} hour of the workday traffic. The deviation factor of this zone, D_{14} is evaluated to be 1.30 using equation 9. With corresponding values of mean and standard deviation of $133,907.90$ MB and $34,688.45$ MB respectively, we determine the maximum volume of traffic that the data center of this zone would have to handle at any given hour to be $219,175$ MB ($1,712$Gb).

To fulfill this volume of traffic demand, the servers in the data center in this zone need to have 12 cores of CPU power.

Zone 19 Data Center: Traffic level of this area surpasses others and possess a $max(P_h(mean + 1std))_{19}$ value and deviation factor, D_{19}, of 1 for the 13^{th} hour of the holiday traffic. The mean and standard deviation value corresponding to this hour are $746,016.40$ MB and $225,068.60$ MB respectively. As expected, the maximum volume of traffic that the data center in this zone needs to cater for, $971,085$ MB ($7,586$ Gb) is also the highest among all others. To satisfy this level of demand, the capacity of this data center would also have to greater than others. This zone's data center would require 55 cores to process the traffic demands from the subscribers.

VI. MACHINE LEARNING BASED TRAFFIC DEMAND PREDICTION

In this section, we devote to employ several state-of-the-art recurrent neural network (RNN) models to forecast next day's traffic of each zone based on previously collected data. The idea is to have the future traffic demand of these areas in hand to facilitate the operation of these data centers. Accurate prediction models will aid operation of the data centers and can lower the operational cost of the infrastructure providers as unused capacity of these data centers can be put on sleep mode, resulting in reduced energy consumption.

RNN has proved to be an effective tool to perform prediction on time-series data. Given that we have a inherently seasonal data of hourly aggregated traffic demand of different zones, RNN models can be used to make forecasts of future demand. We use two RNN models: long short term memory (LSTM) and gated recurrent unit (GRU). We also test the fitness of two activation functions : rectified linear unit (ReLU) and hyperbolic tangent (tanh) to determine the combination that produces the result with highest accuracy. As the holiday demand is different from the workday one, we tested these models on each type of day for the considered zones. Below we briefly explain the concepts of RNN and its models that have been utilized for this section of the work.

A. Recurrent Neural Networks

RNN's main idea is to capture and store relevant amount of information from the input in a memory to use it while making a future prediction for the output. This is a fundamental difference between RNN and traditional feed forward neural networks that simply make use of only the present input to produce an output. RNNs are termed as recurrent as they perform the same operation on every element of a sequence whereby the output of the present step is heavily impacted by that of previous steps. Fig. 9 illustrates a typical RNN model and its ability to include previous input with present one to predict the future output.

RNN takes in the input i, captures the hidden state a and produces an output of o at every time step t. The information from one step to the following is carried on by a loop. The W's stand for various weight matrices during the time steps. These matrices are changed during the training phase as the network is 'unrolled' for a certain number of time steps. As shown in Fig. 9, this unrolling of network in time steps allow the RNN to learn

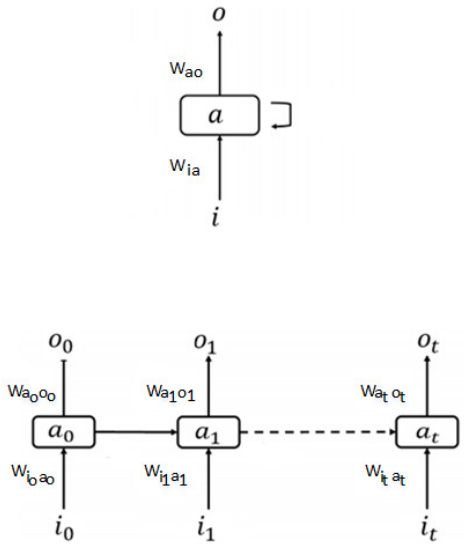


Fig. 9. Recurrent Neural Network architecture

information present in sequential data. The computation that takes place in every time step can be summarised as follows:

1. i_t serves as the input in time step t.
2. The hidden state a_t at time step t is calculated based on the previous hidden step and the present input. These two pieces of information are combined through the use of activation functions such as ReLU and tanh.
3. The output step at time step t is termed as o_t.

With different inputs i_t in different time steps same computations are performed with unrolled parameters W_{ia}, W_{aa} and W_{ao}. This attribute of the RNNs makes them extremely useful for smaller data set by avoiding over fitting. Two common RNN models in use now are the LSTM and GRU. We provide brief explanation of the working principles of these models along with the activation functions.

LSTM: The hidden state in traditional RNN does not provide enough control over how much of the past information should be kept and this leads to problems such as vanishing and exploding gradients [49]. To overcome such problems, LSTM models were designed to have two additional gates termed as the input and forget gates. The gating mechanism allows LSTMs to adequately model long-term dependencies present in complex non linear data. LSTM essentially learns the optimal parameters for its gates during the training phase, thereby determining the behavior of its memory. Interested readers are addressed to read [50] for more details on LSTM.

GRU: Due to the presence of both input and forget gates, the LSTM model often becomes computationally expensive. GRU, a more recent edition of the RNN models, presents a simpler architecture where the input and forget gates are combined into a update gate. The basic idea of capturing and learning long term dependencies on time series data is however maintained in GRU as well. Detailed explanation regarding the GRU model can be found in [51].

Activation functions: the activation functions play an important role in RNN and its models' ability to accurately make future predictions. The two activation functions we have used in

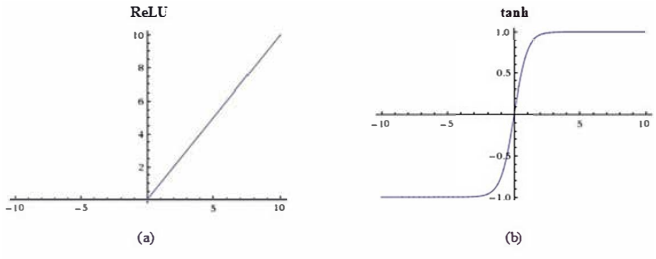


Fig. 10. Activation functions: ReLU and tanh

this work are ReLU and tanh. Fig. 10 (a) shows the ReLU activation function and Fig. 10 (b) demonstrates the tanh activation function.

We test the fitnesses of the LSTM and GRU models on the zonal data to predict future demands. We employ these models with relu and tanh activation functions. We follow a $70 : 30$ train-test split convention i.e the first 70 percent of both holiday and workday data is used for training the neural network and the rest is used for testing. In addition, we use average hourly traffic as a baseline for comparison purposes. We use Google's open source machine learning platform Tensorflow on a 2.6 GHz, 4 cores and NVDIA GTX 970 graphics card enabled computer to analyse the performances of these algorithms on holiday and workday data. The neural network was designed with 2 hidden layers with each having 50 neurons. The input and the output dimensions are both 1×1 The obtained results are explained below:

Holiday Prediction: Figs. 11 (a), (c) and (e) demonstrate the performances of the considered algorithms on holiday data of zone 6, 14 and 19 for a three-day period respectively. Note that holidays consist of less amount of data points compared to workdays (528 data points for holidays compared to 960 data points in workdays). LSTM and GRU models generally perform well with accuracy of 90 percent and more across all zones with both activation functions. The average hourly traffic, however, is clearly seen to be incapable of capturing the traffic trend of these three day period. Figs. 12 (a) and (c) show the average of root mean square errors (RMSE) and symmetric mean absolute percentile error (SMAPE) of these algorithms on holiday data set. It is also observed different model emerges as the best when predictions are made across different zones. Therefore, it can be concluded that no single prediction model can be used to obtain accurate forecast of future traffic across all zones.

Workday Prediction: The performances of the considered machine learning algorithms on a slightly larger workday dataset are presented in Figs. 11 (b), (d) and (f). Once more, the GRU and LSTM algorithms performed similarly to each other and were able to forecast traffic with great accuracy. The average once again can be seen to be insufficient for this purpose. Figs. 12 (b) and (d) shows the RMSE and SMAPE values of each algorithm on workday data set. The GRU and LSTM models predict with least error while maintaining accuracy of greater than 95 percent on this data set. Similar to holiday prediction, the prediction model that makes the best prediction varies in different zones.

Figs. 13 (a) and (b) demonstrates the time it takes for each of these algorithms to complete training and make prediction for both holiday and workday data respectively. As expected, due to the presence of additional gate in the LSTM architecture, it takes slightly longer runtime when compared to a simpler GRU architecture.

With the aid of machine learning algorithms, it would be possible for infrastructure providers to determine the hourly demand that will be encountered from a particular MNO within a zone. In the absence of an accurate traffic forecasting mechanism, allocation of data center resources would be reactive i.e resources would be allocated once the demand arises. This could often lead to congestion and degradation of QoS as it is difficult to allocate proper amount of resources if the allocation is based on reaction. Furthermore, from the data center's point of view, knowing future demand values can aid utilization of its resources. During hours when relatively low volume of traffic is predicted in a certain zone, its data center can effectively keep the amount of resources needed to cater for that predicted volume of traffic operational, with rest being aggressively put on idle/sleep mode [52],[53]. This eliminates the need to constantly keep the data center resources active during all hours. By keeping additional resources inactive will lower the energy consumed by the data center and significantly reduce operational expenses for the infrastructure provider.

VII. CONCLUSION

In this paper, we analysed the open Big data set of Telecom Italia to determine the traffic profiles that exist in different zones within the city of Milan. We processed the data set to have a hourly cellular traffic demand that arises in different parts of the city during the course of the day. Using K-mean clustering algorithm, we split the city of Milan into 20 zones and from that isolated three zones (Zone 6, Zone 14, and Zone 19) that demonstrate the least, medium and most volume of traffic respectively. Based on the location and traffic handled by each base station in a zone, we proposed the establishment of a data center to host the VNFs and SDN controllers in each zone. We identified the problem of the placement of data center as a facility location problem which was solved using Weiszfeld's algorithm. Furthermore, based on the traffic profile of each zone, we heuristically determined the ideal dimension of a data center that will be capable of handling the traffic within that zone. Finally, we used machine learning algorithms to predict the future demand to enhance the operation of the data center in each of the considered zones. Results showed the ability of the LSTM and GRU models to predict future demand values with considerably high accuracy.

Acknowledgment

The research leading to these results has received funding from the European Commission for the H2020-ICT-2016-2 METRO-HAUL project (G.A. 761727). The authors also acknowledge the support received from Telkom SA and the Jasco Group via the Telkom Centre of Excellence (CoE) in Broadband Networks at the University of Cape Town.

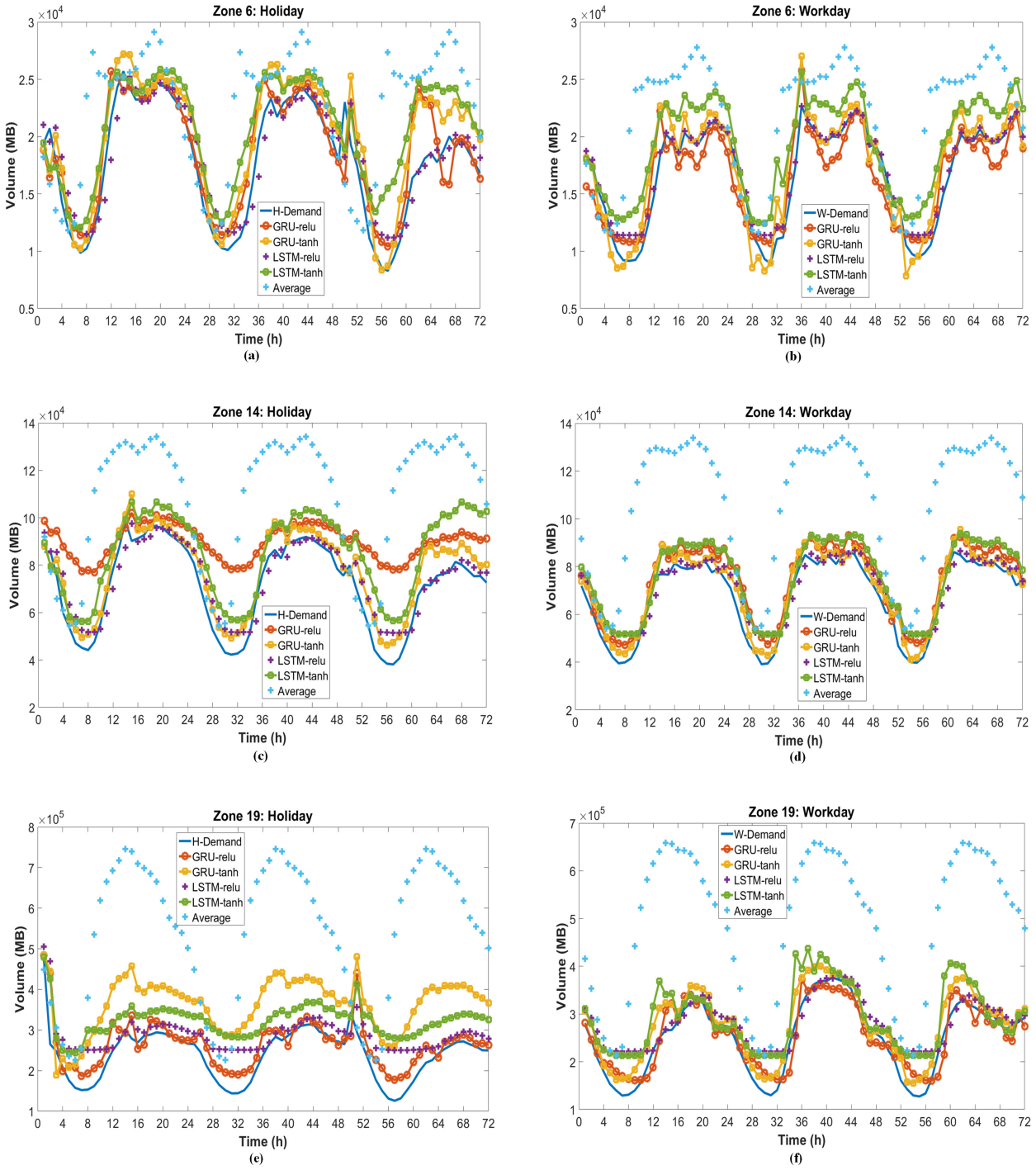


Fig. 11. Machine learning algorithms' predictions for Zones 6, 14 and 19 for holiday and workday traffic

REFERENCES

[1] *Cisco Visual Networking Index (VNI) Update Global Mobile Data Traffic Forecast 2016-2021*, Vni, 2017.

[2] Next generation mobile networks alliance. (Feb. 2015). *NGMN 5G Initiative White Paper*. [Online]. Available: https://www.ngmn.org/uploads/media/NGMN-5G-White-Paper-V1-0.pdf.

[3] *Network functions virtualisation (NFV); Management and orchestration*, ETSI, Sophia Antipolis, France, 2014. [Online]. Available: https://www.etsi.org/deliver/etsi-gs/NFV-MAN/001-099/ 001/01.01.01-60/gs-NFV-MAN001v010101p.pdf

[4] *SDN architecture*, Open Netw. Found., Palo Alto, CA, USA, 2014. [Online]. Available: https://www.opennetworking.org/images/stories/downloads/sdn-resources/technical-reports/TR-SDN-ARCH-1.0-06062014.pdf

[5] B. Martini, F. Paganelli, P. Cappanera, S. Turchi and P. Castoldi, "Latency-aware composition of virtual functions in 5G," in *Proc. NetSoft*, Apr. 2015, pp. 1–6.

[6] R. Pal, S. Lin, and L. Golubchik , "The cloudlet bazaar dynamic markets for the small cloud," ArXiv preprint arXiv: 1704.00845, 2017.

[7] M. Satyanarayanan, "The emergence of edge computing," *Computer*, vol. 50, no. 1, pp. 30–39, Jan. 2017.

[8] *5G network architecture: A high-level perspective*, Huaweii, 2016. [Online]. Available: http://www.huawei.com/minisite/hwmbbf16/insights/5G-Nework-Architecture-Whitepaper-en.pdf

[9] W. Rankothge, F. Le, A. Russo and J. Lobo, "Optimizing resource alloca-

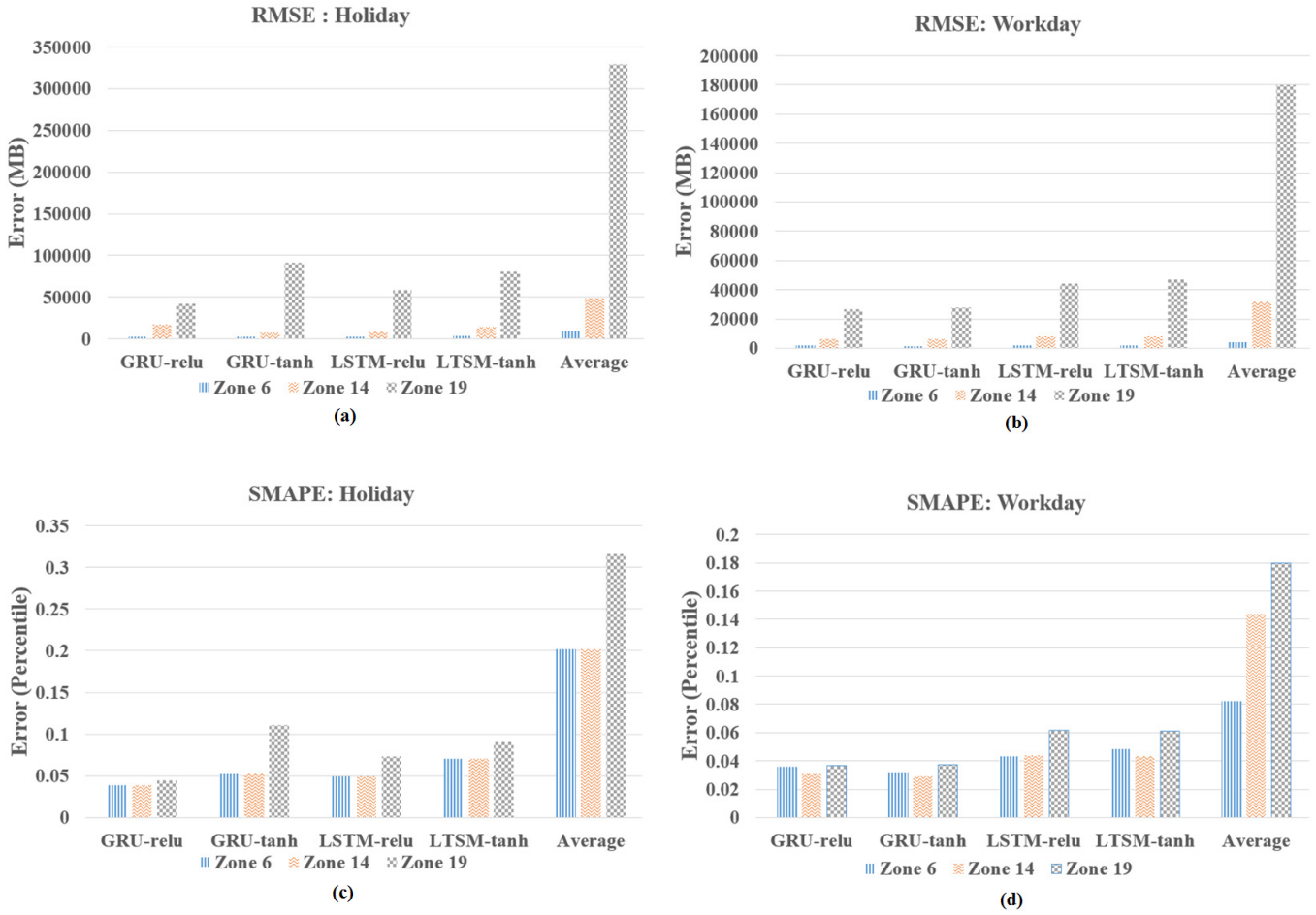


Fig. 12. RMSE and SMAPE of the algorithms in predicting Holiday and Workday traffic.

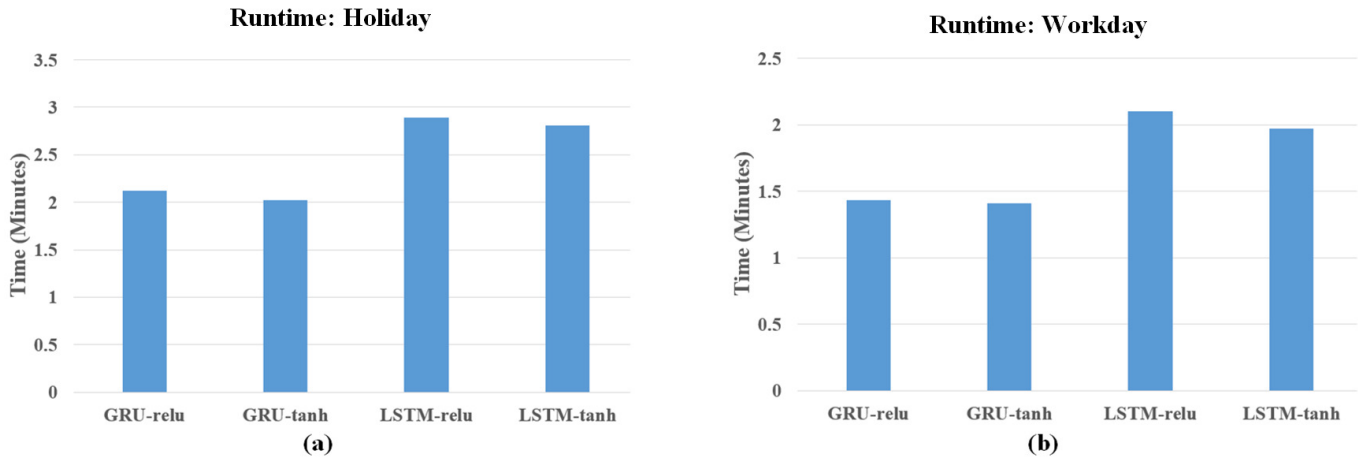


Fig. 13. Runtime of the algorithms to predict Holiday and Workday traffic

tion for virtualized network functions in a cloud center using genetic algorithms," *IEEE Trans. Netw. Service Manag.*, vol. 14, no. 2, pp. 343–356, June 2017.

[10] R. Mijumbi *et al.*, "Topology-aware prediction of virtual network function resource requirements," *IEEE Trans. Netw. Service Manag.*, vol. 14, no. 1, pp. 106–120, Mar. 2017.

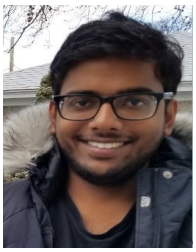
[11] K. Suksomboon, M. Fukushima, M. Hayashi, R. Chawuthai, and H. Takeda, "LawNFO: A decision framework for optimal location aware network function outsourcing," in *Proc. NetSoft*, June 2015, pp. 1—9.

[12] A. Laghrissi, T. Taleb, M. Bagaa and H. Flinck, "Towards edge slicing: VNF placement algorithms for a dynamic & realistic edge cloud environment," in *Proc. IEEE GLOBECOM* , Dec. 2017, pp. 1–6.

[13] R. Tripathi, S. Vignesh, V. Tamarapalli and D. Medhi, "Cost efficient design of fault tolerant geo-distributed data centers," *IEEE Trans Netw. Service Manag.*, vol. 14, no. 2, pp. 289–301, June 2017.

[14] K. Zheng, W. Zheng, L. Li and X. Wang, "PowerNetS: Coordinating data center network with servers and cooling for power optimization," *IEEE Trans. Netw. Service Manag.*, vol. 14, no. 3, pp. 661–675, Sept. 2017.

[15] A. Basta *et al.*, "Towards a cost optimal design for a 5G mobile core network based on SDN and NFV," *IEEE Trans. Netw. Service Manag.*, vol. 14, no. 4, pp. 1061–1075, Dec. 2017.
[16] A. Furno, D. Naboulsi, R. Stanica and M. Fiore, "Mobile demand profiling for cellular cognitive networking," *IEEE Trans. Mobile Comput.*, vol. 16, no. 3, pp. 772–786, Mar. 2017.
[17] S. Wang *et al.*, "An approach for spatial-temporal traffic modeling in mobile cellular networks," in *Proc. IEEE ITC*, Sept. 2015, pp. 203–209.
[18] S. Troia, Gao Sheng, R. Alvizu, G. A. Maier and A. Pattavina, "Identification of tidal-traffic patterns in metro-area mobile networks via Matrix Factorization based model," in *Proc. PerCom Workshops*, Mar. 2017, pp. 297–301.
[19] R. Li *et al.*, "The learning and prediction of application-level traffic data in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3899–3912, June 2017.
[20] R. Li, Z. Zhao, X. Zhou, J. Palicot and H. Zhang, "The prediction analysis of cellular radio access network traffic: From entropy theory to networking practice," in *IEEE Commun. Mag.*, vol. 52, no. 6, pp. 234–240, June 2014.
[21] L. Cui, F. R. Yu and Q. Yan, "When big data meets software-defined networking: SDN for big data and big data for SDN", in *IEEE Netw.*, vol. 30, no. 1, pp. 58-65, Jan. 2016.
[22] J. Dai and J. Li, "Vbr mpeg video traffic dynamic prediction based on the modeling and forecast of time series," in *Proc. IEEE NCM*, Aug. 2009, pp. 1752–1757.
[23] O. Cappe, E. Moulines, J. C. Pesquet, A. P. Petropulu, and X. Yang, "Long-range dependence and heavy-tail modeling for teletraffic data," *IEEE Signal Process. Mag.*, vol. 19, no. 3, pp. 14—27, May 2002.
[24] A. Soule *et al.*, "Traffic matrices: Balancing measurements, inference and modeling," in *Proc. ACM SIGMETRICS*, June 2005, pp. 1–13.
[25] M. C. Falvo, M. Gastaldi, A. Nardecchia, and A. Prudenzi, "Kalman filter for short-term load forecasting: An hourly predictor of municipal load," in *Proc. IASTED ASM*, Aug. 2007, pp. 364—369.
[26] F. Ashtiani, J. A. Salehi and M. R. Aref, "Mobility modeling and analytical solution for spatial traffic distribution in wireless multimedia networks," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 10, pp. 1699–1709, Dec. 2003.
[27] K. Tutschku and P. Tran-Gia, "Spatial traffic estimation and characterization for mobile communication network design," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 5, pp. 804–811, June 1998.
[28] T. P. Oliveira, J. S. Barbar, and A. S. Soares, "Computer network traffic prediction: A comparison between traditional and deep learning neural networks," *International J. Big Data Intelligence*, vol. 3, no. 1, p. 28–37, Jan. 2016.
[29] C. W. Huang, C. T. Chiang and Q. Li, "A study of deep learning networks on mobile traffic forecasting," in *Proc. IEEE PIMRC*, Oct. 2017, pp. 1–6.
[30] M. Barabas, G. Boanea, A. B. Rus, V. Dobrota, and J. D. oPascual, "Evaluation of network traffic prediction based on neural networks with multi-task learning and multiresolution decomposition," in *Proc. IEEE ICCP*, Aug. 2011, pp. 95—102.
[31] G. D'angelo, R. Pilla, J. B. Dean and S. Rampone, "Toward a soft computing-based correlation between oxygen toxicity seizures and hyperoxic hyperpnea", *Soft Comput.*, vol. 22, no. 7, pp. 2421–2427, Apr. 2018.
[32] C. K. Dominicini *et al.*, "VirtPhy: Fully programmable NFV orchestration architecture for edge data centers," *IEEE Trans. Netw. and Service Manag.*, vol. 14, no. 4, pp. 817–830, Dec. 2017.
[33] S. Gebert *et al.*, "Demonstrating the optimal placement of virtualized cellular network functions in case of large crowd events." in *Proc. ACM SIGCOMM*, Aug. 2014, pp. 359–360.
[34] C. H. Liu and J. Fan, "Scalable and efficient diagnosis for 5G data center network traffic," *IEEE Access*, vol. 2, pp. 841–855, Aug. 2014.
[35] M. C. Luizelli, L. R. Bays, L. S. Buriol, M. P. Barcellos, and L. P. Gaspary, "Piecing together the NFV provisioning puzzle: Efficient placement and chaining of virtual network functions," in *Proc. IFIP/IEEE IM*, May 2015, pp. 98—106.
[36] R. Shi *et al.*, "MDP and machine learning-based cost-optimization of dynamic resource allocation for network function virtualization," in *Proc. IEEE SCC*, June 2015, pp. 65–73.
[37] I. Narayanan, A. Kansal and A. Sivasubramaniam, "Right-sizing geo-distributed data centers for availability and latency," in *Proc. IEEE ICDCS*, June 2017, pp. 230–240.
[38] H. Raei, "Capacity planning framework for mobile network operator cloud using analytical performance model," *International J. Commun. Syst.*, vol. 30, no. 17, pp. 1–12, June 2017.
[39] M. Carvalho, D.A. Menasce, and F. Brasileiro "Capacity planning for IaaS cloud providers offering multiple service classes," *Future Generation Comput. Syst.*, vol. 77, pp. 97–111, Dec. 2017.
[40] L. Nie, D. Jiang, L. Guo, S. Yu and H. Song, "Traffic matrix prediction and estimation based on deep learning for data center networks", in *Proc. IEEE Globecom Wkshps*, Dec. 2016, pp 1–6.
[41] Gianni Barlacchi *et al.*, "A multi-source dataset of urban life in the city of Milan and the province of Trentino," *Scientific Data*, vol. 2, Oct. 2015.
[42] http://opencellid.org/
[43] https://www.scribd.com/document/383828390/Traffic-Chracteristics-of-Different-Zones-in-Milan
[44] P. V. Heiningen, E. Reehuis and T. Bäck, "Comparing a Weiszfeld's-based procedure and (1+1)-es for solving the planar single-facility location-routing problem," in *Proc. IEEE SSCI*, Dec. 2015, pp. 1743–1750.
[45] R. Z. Farahani and M. Hekmatfar, *Facility Location Concepts, Models, Algorithms and Case Studies*. SpringerVerlag, Berlin, Heidelberg, 2009.
[46] E. Weiszfeld, "Sur le point sur lequel la somme des distances de n points donnes est minimum," *Tohoku Mathematical J.*, vol. 43 no.1, 1937, pp. 335–386.
[47] K. Aftab, R. Hartley and J. Trumpf, "Generalized Weiszfeld algorithms for Lq optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 728–745, Apr. 2015.
[48] A. Basta, W. Kellerer, M. Hoffmann, H. J. Morper, and K. Hoffmann, "Applying NFV and SDN to LTE mobile core gateways, the functions placement problem," in *Proc. ACM SIGCOMM*, Aug. 2014, pp. 33—38.
[49] R. Pascanu, T. Mikolov, Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. ICML*, June 2013, pp. 1310–1318.
[50] Hochreiter, Sepp and Schmidhuber, Jürgen, "Long Short-Term Memory", *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[51] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, " Empirical evaluation of gated recurrent neural networks on sequence modeling". in *Proc. NIPS*, Dec. 2014.
[52] C. Gu, Z. Li, H. Huang and X. Jia, "Energy efficient scheduling of servers with multi-sleep modes for cloud data center," in *EEE Trans. Cloud Comput. (Early Access)*,doi: 10.1109/TCC.2018.2834376.
[53] L. Fan, C. Gu, L. Qiao, W. Wu and H. Huang, "GreenSleep: A multi-sleep modes based scheduling of servers for cloud data center," in *Proc. IEEE BIGCOM*, Aug. 2017, pp. 368–375 .



Udita Paul received his M.Sc in Electrical Engineering at the University of Cape Town in 2018. He is currently pursuing Ph.D. degree in Computer Science at the University of California, Santa Barbara with the MOMENT lab. His primary research interests include 5G, network slicing, network function virtualization, edge computing and machine learning algorithms.

Jiamo Liu received his B.Sc in Electrical and Computer Engineering and a M.Sc in Electrical Engineering at the University of Cape Town in 2017 and 2018 respectively. His primary research interests include machine learning, resource allocation, 5G and network slicing.

Sebastian Troia received the B.Sc. and M.Sc. degrees in telecommunications engineering from the Politecnico di Milano in 2013 and 2016, respectively, where he is currently pursuing a Ph.D. degree in information technology with the BONSAI LAB. His research interests are related to machine learning algorithms for communications networks, software-defined networking, network orchestration automation-optimization, SD-WAN, and optical multipath routing.

Olabisi E. Falowo received his Ph.D. in Electrical Engineering at the University of Cape Town in 2008. He is currently an Associate Professor in the same department. He has published over 100 technical papers in peer-reviewed conference proceedings and journals. His primary research interest is in radio resource management in heterogeneous wireless networks. Olabisi Falowo is a senior member of the IEEE.

Guido Maier received the Laurea degree in electronic engineering and the Ph.D. degree in telecommunication engineering from the Politecnico di Milano, Italy, in 1995 and 2000, respectively. From 1995 to February 2006, he was a Researcher with CoreCom (research consortium supported by Pirelli and PoliMi in Milan, Italy), where he achieved the position of the Head of the Optical Networking Laboratory. In March 2006, he joined the Department of Electronics, Information and Bioengineering, Politecnico di Milano, as an Assistant Professor, where he became an Associate Professor in January 2015. He has participated to several joint research projects with industrial partners, such as Pirelli, Telecom Italia, Fastweb, and RFI. He has been involved in the IST-FP6 and FP7 European Projects, such as MUPBED, NOBEL2, e-Photon/ONe+, BONE, STRONGEST, and (IRSES) MobileCloud. He has authored or coauthored over 90 papers published in proceedings of international conferences and 35 papers published in international journals on networking and optical networks.