# How beneficial are intermediate layer Data Centers in Mobile Edge Networks?

Mehta, Amardeep; Tärneberg, William; Klein, Cristian; Tordsson, Johan; Kihl, Maria; Elmroth, Erik

*Published in:*
Workshops on Fog and Mobile Edge Computing at Foundations and Applications of Self* Systems

*DOI:*
10.1109/FAS-W.2016.55

2016

2016 IEEE 1st International Workshops on Foundations and Applications of Self-* Systems

# How beneficial are intermediate layer Data Centers in Mobile Edge Networks?

Amardeep Mehta*, William Tärneberg†, Cristian Klein*, Johan Tordsson*, Maria Kihl† and Erik Elmroth*
*Department of Computing Science, Umeå University, Sweden
Email: $(amardeep,\ cklein,\ tordsson,\ elmroth)@cs.umu.se$
†EIT, Lund University, Sweden
Email: $(william.tarneberg,\ maria.kihl)@eit.lth.se$

***Abstract*—To reduce the congestion due to the future bandwidth-hungry applications in domains such as Health care, Internet of Things (IoT), etc., we study the benefit of introducing additional Data Centers (DCs) closer to the network edge for the optimal application placement. Our study shows that the edge layer DCs in a Mobile Edge Network (MEN) infrastructure is cost beneficial for the bandwidth-hungry applications having their strong demand locality and in the scenarios where large capacity is deployed at the edge layer DCs. The cost savings for such applications can go up to 67%. Additional intermediate layer DCs close to the root DC can be marginally cost beneficial for the compute intensive applications with medium or low demand locality. Hence, a Telecom Network Operator should start building an edge DC first having capacity up to hundreds of servers at the network edge to cater the emerging bandwidth-hungry applications and to minimize its operational cost.**

***Keywords*—Application Placement, Cost Optimization, Infrastructure Resource Placement, Mobile Edge Computing, Fog Computing**

## I. Introduction

The current telco networks may not be able to support the bandwidth requirements of network-intensive cloud applications such as video surveillance for home security, high definition video conferencing, telemedicine, etc., that may require bandwidths up to hundreds of megabits per second. For example, Internet of Things (IoT) applications such as virtual reality, will demand bandwidths up to 100 Mb/s per application [1]. Interestingly, the problem is neither in the last-mile nor between Internet Service Providers (ISPs), but the internal part of the telco networks when the aggregated demand exceeds the network capacity, which may lead to congestion [2]. Hence, the internal network may not be able to meet the challenges of the future bandwidth-hungry Internet-based emerging applications due to expensive backhauling and increasing wired network congestion. Telecom Network Operators (TNOs) have been using bandwidth usage-based pricing as a congestion control tool [3]. The usage-based pricing has several drawbacks, as it may get adverse response from the end-users due to uncertainty of the network budget expenses. It will also increase the management and billing costs to the ISP substantially. Finally, it may discourage the use of Internet, a notion that many in research communities find objectionable [4].

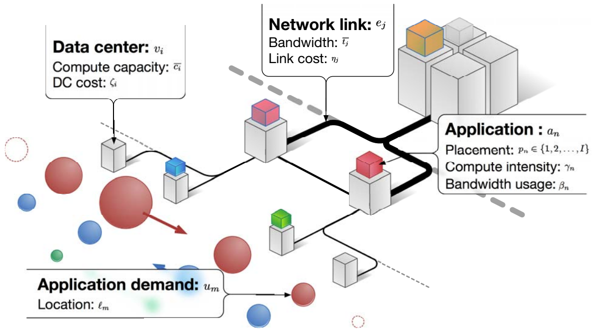

Data center: $v_i$
Compute capacity: $\overline{c_i}$
DC cost: $\zeta_i$
Network link: $e_j$
Bandwidth: $\overline{t_j}$
Link cost: $\eta_j$
Application : $a_n$
Placement: $p_n \in \{1,2,\ldots,I\}$
Compute intensity: $\gamma_n$
Bandwidth usage: $\beta_n$
Application demand: $u_m$
Location: $\ell_m$

Fig. 1. Heterogeneous Data Center compute and link bandwidth capacity distributed over the Mobile Edge Network.

One approach to address the challenge is to move the computations close to the end-users in the ISP's internal network to avoid congestion. Augmenting an existing large-scale distributed cloud infrastructures with heterogeneous compute capacities inside the ISP's internal network, called Mobile Edge Network (MEN) as shown in the Figure 1, can alleviate the congestion problem and meet the performance challenges due to the emerging applications. The compute cost at a large-scale cloud Data Center (DC) is cheaper compared to the edge DC considering economy of scale [5]. The incurred overall bandwidth usage cost is higher for an application placed at a large-scale distant DC compared to an edge DC closed to the end-users. This trade-off leads to an essential question, where and how much compute capacity needs to be allocated to an intermediate layer DC of a MEN. This paper investigates the various system parameters of a MEN to answer the question from the TNO perspective. Previous work study the optimal application placement for a given infrastructure. In contrast, we study where DCs should be located in the infrastructure for the cost-optimal operations.

We investigate the benefit of introducing additional DCs close to the network edge for the optimal application placement. We also identify and model the relevant system parameters to capture the dynamics of a MEN, and find the most sensitive parameter among them. The models are described in Section II. The cost benefits of having the intermediate layer DCs for different application types is described in

978-1-5090-3651-6/16 $31.00 © 2016 IEEE
DOI 10.1109/FAS-W.2016.55

Section III. Our evaluation shows that the edge layer DCs in a MEN infrastructure are cost beneficial for the bandwidth-hungry applications having strong demand locality and in the scenarios where large capacity is deployed at the edge layer DCs.

## II. System Models

We consider the future infrastructure built over the current existing ones for a MEN. To study the conditions in which intermediate layer DCs for a MEN are cost beneficial, we propose the following models that capture the most important costs. We also model the concerned values along with the abstract models. The network topology is modeled as a tree for a MEN, where the vertices are DCs and the edges are network links, each with a set of finite resources. Applications are hosted in DCs and subject to demand through the network links, originating at the leaf nodes. The graph $G = (V, E)$ represents the MENs network topology, where $V = \{v_i \mid i = 1, 2, ..., I\}, E = \{e_j \mid j = 1, 2, ..., J\}$.

### A. Data center Model

Heterogeneous compute capacities for the DCs, which diminishes with depth, are distributed over the MEN. The cost for running a CPU in a DC is not linearly proportional to that of the other heterogeneous capacity DC considering economy of scale [5]. A DC is represented by a vertex $v_i$, $i \in \{1, 2, \ldots, I\}$ in the graph and has the following characteristic:

- **Compute capacity** $\overline{c_i}$ is a number describing the total compute capacity of the DC.

Other, characteristics such as Memory, Storage can be modeled in similar ways. We omit them to simplify the model, as their dynamics is similar to the compute capacity parameter. Further, a vertex $v_i$, $i \in \{1, 2, \ldots, I\}$ is associated with the following operational cost:

- **DC compute cost** $\zeta_i$ is a linear function of compute resource usage that approximates the DC's running cost per time unit.

For example, the DC compute cost function can be approximated as $\zeta_i(c_i) = w_i c_i$, where $w_i$ is the **cost-per-unit-compute** resource and $c_i$ is the compute capacity used by the running applications at the DC $i$. The unit for $w_i$ is cost per unit CPU-hour usage.

To model the cost-per-unit-compute resource of a DC, one can use Walker's model proposed in [6]. The model computes the cost of a CPU-hour as a ratio of the net present value (NPV) and the net present capacity (NPC) of a DC. It not only considers investment amortization but also the change in the CPU performance over time in accordance with Moore's law. The NPV is all the expenses incurred during the technological lifetime of the equipments and consists of Capital Expenditure (CAPEX) (initial investment in purchasing computers, softwares, networking equipments, power generator and cooling systems, etc.) and Operational Expenditure (OPEX) (operating and cooling energy, staff costs, space renting for DC hosting, software licenses, etc.), which is equivalent to Total Cost of Ownership (TCO). NPC depends on exploitable total useful capacity (TC) in terms of the obtained CPU hours and operational life time of the installed equipments of the DC.

TABLE I
Cost of a CPU-hour for a DC capacity based on Walker's Model

| Capacity (#servers) | 1 | 15 | 150 | 1500 | 15000 | 150000 |
|---|---|---|---|---|---|---|
| Cost ($ per CPU-h) | 0.206 | 0.148 | 0.115 | 0.091 | 0.075 | 0.064 |
| Relative cost ratio | 3.22 | 2.31 | 1.80 | 1.42 | 1.17 | 1 |

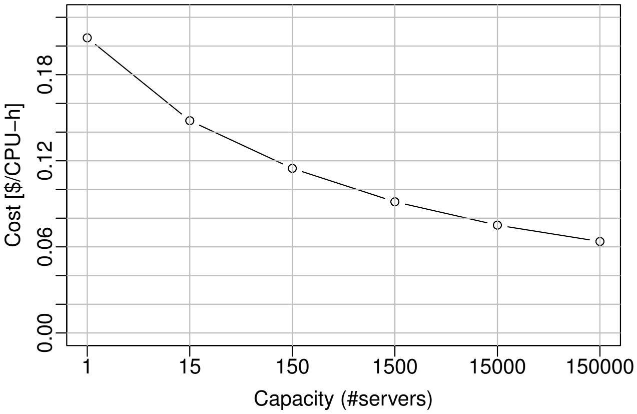


Fig. 2. Visualization of data from Table I

The OPEX does not scale linearly with scaled capacity of a DC considering economy of scale. For example, if the capacity of a DC is 10 times higher than another DC, then its OPEX cost is only 7 times higher than the cost of the other DC [7]. For 15000 servers with quad-core processors, the annual operating cost is $7 \times 10^6$ (unit \$) and CAPEX is \$$30 \times 10^6$ (unit \$) considering each server costs 2000 (unit \$) [6]. We assume 60% server utilization, 7% annual cost of capital and 5 years amortization period [7]. Table I shows the cost of CPU-h vs capacity for a DC. The non-linear increase in the cost due to economy of scale is shown in the Figure 2 for different capacity of DCs.

### B. Network Model

Each link $e_j$, $j \in \{1, 2, ..., J\}$, in the graph has the following characteristic:

- **Bandwidth** $\overline{t_j}$ is a number specifying the maximum throughput over the edge.

In addition, each edge has the following operational cost

- **Link cost** $\eta_j$ is a linear function of throughput that returns the link's running cost per time unit.

For example, the link cost function can be approximated as $\eta_j(t_j) = y_j t_j$, where $y_j$ is the **cost-per-unit-link** resource and $t_j$ is the bandwidth used by the running applications at the link $j$. The unit for $y_j$ is cost per unit GB data transfer.

We assume that links connected to a large DC have higher bandwidth capacity and hence lower cost for transferring data.

TABLE II
COST FOR TRANSFERRING 1 GB DATA OVER A LINK

| Cable / Capacity (Mbps) | OC3 / (155) | OC12 / (622) | OC48 / (2500) |
|---|---|---|---|
| Cost ($ per GB) | 0.3078 | 0.0980 | 0.0759 |
| Relative cost ratio | 4.06 | 1.29 | 1 |

The link's capacity decreases with the depth of the network topology whereas the cost per unit resource usage over the link increases. The flat rate pricing method for a network link is transparent, easy to implement and widely practiced across the world by several TNOs [2], [8]. The actual cost of transferring data through a link of the ISPs' internal network depends on the network technology, link capacity, labor cost for laying cables, maintenance cost, network usage, number of subscribers, and the properties of the backhaul links.

To model the cost-per-unit-link resource, we use the cost estimation from [2], where the author has estimated the effective cost to transfer 1 GB of traffic based on Bell's Network cost model. Table II shows the cost for 1 GB transfer through links using High Speed Packet Access (HPSA) network technology and considering subscriber growth over the years of operation [2]. The computed link cost also matches that of [8].

### C. Application Model

An application $a_n$, $n \in \{1, 2, \ldots, N\}$ has the following characteristics:

- **Placement** $p_n \in \{1, 2, \ldots, I\}$ is a number specifying that the application is running on the DC at vertex $v_{p_n}$.
- **Compute intensity** $\gamma_n$ is a linearly increasing function of the demand of the application $n$ that describes the amount of computational resources required by the application (the unit being CPU-h per request),
- **Bandwidth usage** $\beta_n$ is a linearly increasing function of the demand of the application $n$ with respect to the application's end-user locations that returns the bandwidth usage for the application (the unit being GB per request).

For an application, we define the **compute-intensity-to-bandwidth-usage-ratio**, $A_{cl} = \frac{\gamma_n}{\beta_n}$ (the unit being CPU-h/GB) per request. This ratio is used to group applications with similar resource consumption ratios.

### D. Application Demand Model

Finally, let $U = \{u_m \mid m = 1, 2, \ldots, M\}$ be the set of users of a MEN. Each user $u_m$, $m \in \{1, 2, \ldots, M\}$ has the following characteristic:

- **Location** $\ell_m \in \{1, 2, \ldots, I\}$ is a number specifying that the user is currently served by the DC at leaf vertex $v_{l_m}$.

We define $U_n = \{u_m \in U \mid a_n \in A\}$ to be the demand for an application $n$ and let $U_{n,i} = \{u_m \in U_n \mid \ell_m = i\}$ be the demand of application $n$ being served by the DC at the leaf vertex $v_i$. For each application $a_n$, we define $\lambda_{n,i} = \frac{U_{n,i}}{U_n}$ as the **fraction-of-total-demand** coming through leaf node $v_i$.

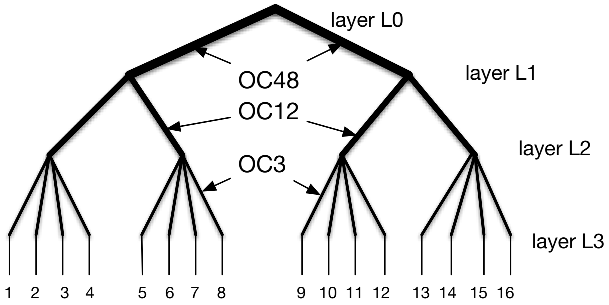


Fig. 3. Mobile Edge Network Topology for Experimentation where DCs can be placed at a layer.

## III. EXPERIMENTS AND RESULTS

In this section, we investigate the application placement and the cost savings for having intermediate layer DCs in a MEN topology. We assume that each application is mapped to one DC. For the rest of this paper, we consider a scenario where the main objective for a TNO is to place the applications in order to minimize the cost of overall infrastructure. Total application execution cost for a TNO is the sum of costs of consumed resources of DCs and links by all the running applications.

### A. Experimental setup

We study the impact of aspects such as DC compute capacity distribution, application types, and application demand locality parameters for a MEN topology with a known network cost model using a simulator.

**Topology**: We model a MEN as a tree having four layers named L0, L1, L2 and L3 from the root to the leaf having link capacity OC48, OC12 and OC3 between subsequent layers, as shown in the Figure 3. The Global Internet have mesh topology whereas an ISP's internal network resembles as tree [9]. Hence, the tree topology is general enough to capture the main characteristics of the problem to find out the cost benefits of the intermediate layer DCs in an ISP's internal network. Having a different topology than a perfectly balanced tree will affect the results, but our study look at the main differences in application request origin so the obtained results still hold true for the general scenarios. Each layer has intersection points among the edges where a DC is placed in our experiments. Application demand from end-users comes from edge nodes at layer L3. For example, the leaf nodes numbered 1-16, as shown in the Figure 3, represent the application demand from the end-users.

**Compute cost distribution**: We consider two extreme cases for the DC cost distribution over the intermediate layers of a MEN to capture all possible capacity distribution scenarios for the intermediate layers' DCs. Examples of the low and high cost distribution scenarios are shown in the Table III. We assume 150000 servers for the layer L0 DC size [10]. For the low cost distribution scenario, the capacity is reduced by 10 fold for the subsequent layers, whereas it is reduced by 50 fold for the high cost distribution scenario. The resource cost

TABLE III
DC resource cost distribution scenarios based on two extreme capacity distributions

| Layers | Low cost distribution | | High cost distribution | |
|---|---|---|---|---|
| | Capacity (#servers) | Cost ($/CPU-h) | Capacity (#servers) | Cost ($/CPU-h) |
| L0 | 150000 | 0.064 | 150000 | 0.064 |
| L1 | 15000 | 0.075 | 3000 | 0.086 |
| L2 | 1500 | 0.091 | 60 | 0.127 |
| L3 | 150 | 0.115 | 1 | 0.206 |

TABLE IV
$A_{cl}$ (CPU-h/GB) per request of the profiled applications

| Application Name | #requests | CPU time(ms) | Bandwidth (MB) | $A_{cl}$ (CPU-h/GB) |
|---|---|---|---|---|
| Web | 8545 | 521050 | 746.2 | 0.19 |
| Streaming | 33892 | 60690 | 107070.5 | 0.157 |

at an intermediate layer is higher in the high cost distribution scenario compared to the low cost distribution scenario.

**Application demand locality**: Application placement is only determined by the relative demand coming from a leaf node, hence, there is no need to explicitly track each user nor the total absolute demand. The application demand locality can be modeled using the relative demand ($\lambda$) from leaf nodes that aggregate at specific layers in MEN. For example, **High locality** applications can be modeled using demand from leaf nodes that aggregate at layer L2. We choose leaf nodes 1 and 2 for this purpose. Similarly, **Medium locality** applications can be modeled using demand from leaf nodes that aggregate at layer L1. We choose leaf nodes 1 and 5 for this purpose. Finally, **Low locality** applications can be modeled using demand from leaf nodes that aggregate at layer L0. We choose leaf nodes 1 and 9 for this purpose.

The relative demand ($\lambda$) of an application can be 1, i.e. the whole demand is coming from a single leaf node, or 0.5 when the demand is coming equally from the two leaf nodes. In general, if demand $\lambda$ is coming from the first leaf node, then $(1 - \lambda)$ will be the demand from the second leaf node. To simulate the application end-user distribution, certain values of $\lambda$ can be chosen in the range from 1 to 0.5. For the current experiment, 10 values are chosen for $\lambda$ in the range from 0.5 to 1 to simulate the applications' end-user distribution scenarios.

**Application types**: Today's existing applications can typically be mapped in $A_{cl}$ range from 0.1 to 10. To choose the relevant values, we profiled two network intensive applications, web-server and Media streaming from the CloudSuite 3.0 benchmark [11], and computed their $A_{cl}$ values. The 3-tier web server application consists of a web server, a Memcached server, and a Database server. The total CPU time for the web application is computed by aggregating the CPU times consumed by all the tiers. The total bandwidth requirements include both upload and download data for all the requests. The $A_{cl}$ values for the Web-server and Media streaming applications are 0.19 and 0.157 respectively as shown in the Table IV. Future bandwidth-hungry applications may require bandwidths up to 100 Mbps [1]. Hence, the $A_{cl}$ range for the existing and future bandwidth-hungry applications may be in the range from 0.01 to 1.

TABLE V
Application types based on $A_{cl}$ values

| Application type | Name | $A_{cl}$ (CPU-h/GB) |
|---|---|---|
| Network intensive | $A_{cl,1}$ | 0.01,1 |
| Compute intensive | $A_{cl,2}$ | 1, 10 |
| Network and compute intensive | $A_{cl,3}$ | 0.01, 10 |

TABLE VI
Experimentation Scenarios combining the DC cost distribution and the application demand locality

| Scenario Names (DC cost distribution, demand locality) | Compute cost distribution | Leaf nodes |
|---|---|---|
| Low cost distribution, high locality | Low | 1, 2 |
| Low cost distribution, medium locality | | 1, 5 |
| Low cost distribution, low locality | | 1, 9 |
| High cost distribution, high locality | High | 1, 2 |
| High cost distribution, medium locality | | 1, 5 |
| High cost distribution, low locality | | 1, 9 |

We study three application groups namely Network intensive, Compute intensive, and Both network and compute intensive applications based on $A_{cl}$ range as shown in Table V.

We simulate 100 applications for each application group. Applications are uniformly distributed over the logarithmic $A_{cl}$ range. The idea is to have more applications close to the lower limit of the $A_{cl}$ range.

We thus obtain 6 scenarios, combining 2 capacity distributions and 3 application localities, as presented in Table VI. In the next section, we evaluate each scenario for each application range.

### B. Impact of Optimal Application Placement

One way to understand the cost benefits of a MEN is to evaluate the impact of an application placement. For this purpose, we compute the percentage of applications ending up in each intermediate layer DC for the different DC cost distribution and applications' demand locality scenarios mentioned in Table VI. The plots are shown in the Figure 4. The x-axis of each plot for the application placement shows the relative demand from a leaf node, whereas the y-axis shows the $A_{cl}$ values for the applications. Each color region in the plot shows the DC layer where applications with given $A_{cl}$ values and demand locality would be optimally placed. Table VII shows the percentage of applications ending up in the intermediate layer DCs for the MEN topology for the scenarios mentioned in Table VI.

*Effect of DC cost distribution parameter:* Figures 4a–4c show applications with $A_{cl} \leq 3$ can be placed at the intermediate layers' DCs for the high cost distribution scenarios, whereas Figures 4d–4f show applications with $A_{cl} \leq 6$ can be placed at the intermediate layers' DCs for the low cost distribution scenarios. Hence, more applications with higher $A_{cl}$ values can be optimally placed close to the network edge for the low cost distribution scenarios. A change from high to

TABLE VII
PERCENTAGE OF THE APPLICATIONS IN $A_{cl,3}$ BEING PLACED AT THE INTERMEDIATE LAYER DCS FOR THE OPTIMAL PLACEMENT

| Scenarios | | **Application percentage for layer** | | | |
|---|---|---|---|---|---|
| Cost distribution | demand locality | L0 | L1 | L2 | L3 |
| Low | high | 4.8 | 1 | 12.8 | 81.4 |
| | medium | 5 | 15.5 | 0 | 79.5 |
| | low | 21.2 | 0 | 0 | 78.8 |
| High | high | 16 | 3.8 | 13.3 | 66.9 |
| | medium | 16 | 18.6 | 0 | 65.4 |
| | low | 34 | 0.6 | 0 | 65.4 |

TABLE VIII
MEDIAN PERCENTAGE COST SAVINGS FOR DCS AT THE INTERMEDIATE PLACEMENT LAYERS FOR THE HIGH COST DISTRIBUTION SCENARIOS

| **Application type** | **demand locality** | **Placement layers** | | |
|---|---|---|---|---|
| | | L0,L3 | L0,L2,L3 | L0-L3 |
| Network-intensive | high | 64.25 | 64.41 | 64.41 |
| | medium | 54.21 | 54.37 | 54.46 |
| | low | 46.73 | 46.73 | 46.73 |
| Compute-intensive | high | 0.51 | 0.50 | 1.31 |
| | medium | 0.48 | 0.48 | 1.31 |
| | low | 0.48 | 0.48 | 0.48 |
| Network and compute intensive | high | 56.35 | 56.85 | 56.85 |
| | medium | 46.59 | 47.09 | 47.35 |
| | low | 39.93 | 39.93 | 39.93 |

TABLE IX
MEDIAN PERCENTAGE COST SAVINGS FOR DC AT THE INTERMEDIATE PLACEMENT LAYERS FOR THE LOW COST DISTRIBUTION SCENARIOS

| **Application type** | **demand locality** | **Placement layers** | | |
|---|---|---|---|---|
| | | L0,L3 | L0,L2,L3 | L0-L3 |
| Network-intensive | high | 66.11 | 66.16 | 66.16 |
| | medium | 56.07 | 56.12 | 56.16 |
| | low | 48.40 | 48.40 | 48.40 |
| Compute-intensive | high | 24.34 | 26.17 | 26.17 |
| | medium | 19.02 | 19.15 | 20.75 |
| | low | 16.38 | 16.38 | 16.38 |
| Network and compute intensive | high | 62.09 | 62.24 | 62.24 |
| | medium | 52.33 | 52.48 | 52.58 |
| | low | 45.09 | 45.09 | 45.09 |

low cost distribution for an intermediate layer's DCs results into a shrinking of the application placement region for all the DC layers except the edge DC layer as shown in the plots of the Figure 4. Applications with $A_{cl}$ values up to 5 can be optimally placed at the edge layer DC for the low cost distribution scenarios. Therefore, the edge layer DC would be the most cost beneficial for the low cost distribution scenarios among the intermediate layers' DCs.

*Effect of Application demand locality:* For the high DC cost distribution scenarios, Figures 4a–4c show that an application placement is sensitive to the first intermediate layer close the network edge where its demand is aggregated, i.e., $\sum \lambda_k = 1$, where $k$ is the application demand leaf node. As shown in Figure 4 when the application demand locality decreases and $\lambda$ goes closer to 0.5 from a leaf node, then the applications with relatively higher $A_{cl}$ range would be optimally placed at the intermediate layer DCs where their demand is aggregated, compared to the applications in the other intermediate layers. For example, the application demand is aggregated at the layer L1 for the high cost distribution, medium locality scenario, applications with $A_{cl} \in [0.03, 3]$ can be placed in layer L1 DC for $\lambda = 0.5$. However, no low locality applications would be placed at the layer L1 DC.

Low cost distribution scenarios shown in the Figure 4d–4f show similar analysis except the placement region shrinks more compared to the high cost distribution scenarios.

The application percentage decreases from 13.3% to 0% due to the change in the demand locality from high to medium for the high distribution cost scenarios as shown in the Figures 4a and 4b, whereas it decreases from 13.3% to only 12.8% due to the change in the cost distribution from high to low for the high demand locality applications as shown in the Figure 4a and 4d, also mentioned in the Table VII. Hence, application demand locality is more sensitive compared to the DC cost distribution parameter.

*Effect of $A_{cl}$:* Figure 4 shows that the network intensive applications with lower $A_{cl}$ values would be optimally placed closer to the edge DC, whereas the compute intensive applications with higher $A_{cl}$ values would be optimally placed closer to the root DC to minimize the overall cost. However, only the $A_{cl}$ value does not decide the exact DC layer for the optimal application placement.

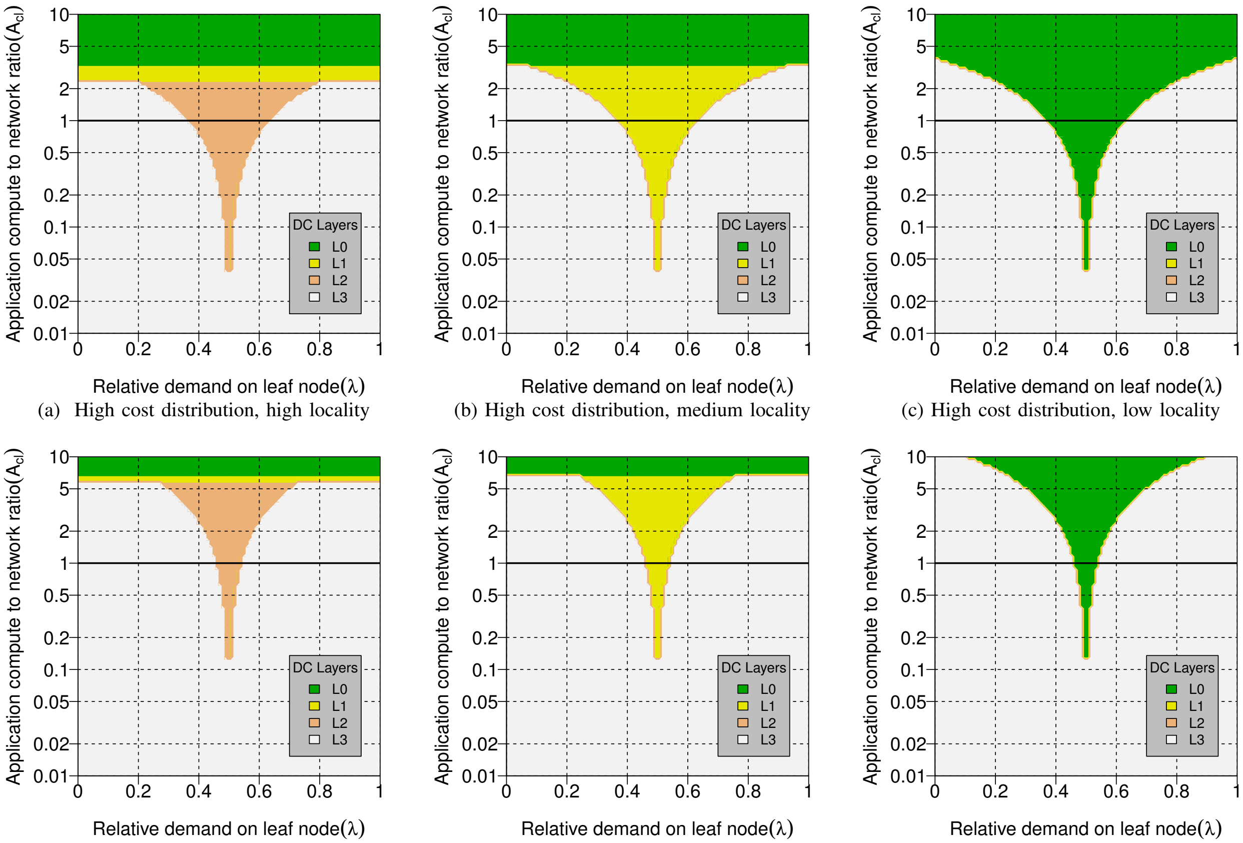
For the high DC cost distribution scenarios, applications with $A_{cl} \leq 0.03$ would be optimally placed at the edge DCs, whereas applications with $A_{cl} \leq 0.12$ would be optimally placed at the edge DCs for the low DC cost distribution scenarios as shown in the Figure 4. Hence, applications in a relatively larger $A_{cl}$ range would be placed at the edge DCs for the low cost distribution scenarios compared to the high cost distribution scenarios. For example, more than 79% of the applications would be placed at the edge DCs for the low cost distribution scenarios compared to only 65% of the applications for the high cost distribution scenarios as shown in the Table VII.

The optimal application placement region for an intermediate layer DC is more sensitive to $A_{cl}$ than the application demand locality parameter. For example, more than 79% of the applications would be placed at the edge DCs for the low cost distribution scenarios for the given $A_{cl}$ range for the optimal placement as shown in the Table VII. The rest of the applications, which is less than 21%, would be optimally placed at the other DCs layers where their demands are aggregated.

## C. Cost Savings Analysis

In this section, we quantify the cost benefits by computing the average cost savings for the applications in a given $A_{cl}$ range when adding DCs at the intermediate layers compared to a scenario having a DC only at the root of a MEN.

Figure 5 shows the violin plots for the cost benefits for the scenarios mentioned in Table VI. Violin plots are similar to box plots, except that they also show the probability density of the data. Each plot in Figure 5 shows the percentage cost benefits for a TNO for the three application types mentioned in Table V. The x-axis of each plot shows three groups of



(a) High cost distribution, high locality

(b) High cost distribution, medium locality

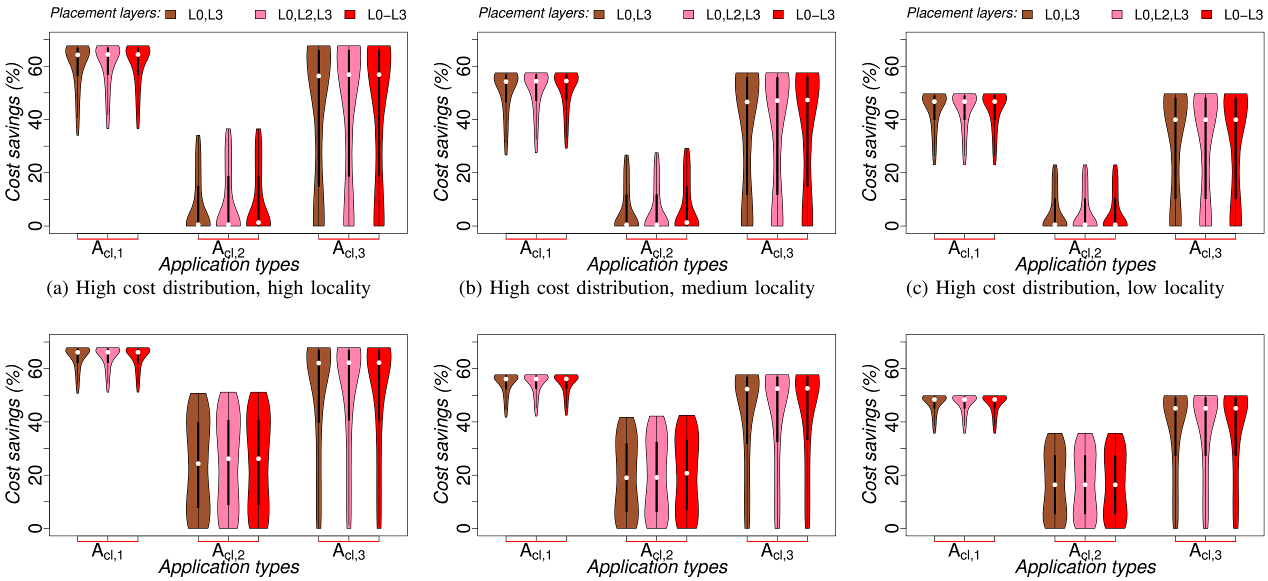
(c) High cost distribution, low locality

(d) Low cost distribution, high locality

(e) Low cost distribution, medium locality

(f) Low cost distribution, low locality

Fig. 4. Applications in $A_{cl,3}$ being placed at the intermediate layer DCs for the optimal placement. The application demand locality decreases for the plots from left to right, whereas the DC cost distribution decreases for the plots from top to bottom.



(a) High cost distribution, high locality

(b) High cost distribution, medium locality

(c) High cost distribution, low locality

(d) Low cost distribution, high locality

(e) Low cost distribution, medium locality

(f) Low cost distribution, low locality

Fig. 5. Cost savings for DCs at the intermediate placement layers. The application demand locality decreases for the plots from left to right, whereas the DC cost distribution decreases for the plots from top to bottom.

plots, one each for the three application types, whereas the y-axis shows the percentage cost savings. Each group contains three violin plots, each for having DCs at the layers $(L0, L3)$, $(L0, L2, L3)$ and $(L0, L1, L2, L3)$ for the optimal application placement. The thickest region of each violin plot for an application type shown in Figure 5 corresponds to the most frequent cost savings. The order for choosing the intermediate layers for having DCs for the application placement is based on the existing cache server solutions for the Content Delivery Networks (CDNs) [12]. Table VIII–IX show the median of the percentage cost savings for the high and low cost distribution scenarios for 3 application types, 3 demand locality and 3 cases for having DCs at the layers $(L0, L3)$, $(L0, L2, L3)$ and $(L0, L1, L2, L3)$.

*Effect of DC cost distribution:* Figures 5a and 5d shows the plots for the cost savings for the network intensive applications for the high and low cost distribution scenarios respectively. The percentage cost savings are between 33% and 68% for the high cost distribution scenario due to the additional edge layer (L3) DCs, while they are between 50% and 68% for low cost distribution scenarios as mentioned in the Tables VIII–IX. The change in the median cost savings for the network intensive applications increases merely up to 2%, whereas it can go up to 24% for the compute intensive applications due to the change in the cost distribution from high to low and the additional edge layer (L3) DCs as shown in the Tables VIII–IX. The increase in the percentage cost savings is due to applications with relatively higher $A_{cl}$ values being placed at the intermediate layer DCs for the low cost distribution scenario.

Adding DCs at more intermediate layers do not result into any significant cost benefits as shown in Tables VIII–IX.

*Effect of Application demand locality:* The applications' median cost savings decrease as the their demand locality changes from high to low as shown in Figures 5a–5c. It decreases from 64% to 46% and from 66% to 48% for high and low cost distribution scenarios as shown in Tables VIII–IX. The cost savings for the high demand locality applications can increase slightly by adding DCs at the layer L2 along with L0 and L3 layers as shown in Figure 5.

The application demand locality is more sensitive than the cost distribution parameter as only a 2% increase can be achieved in the cost savings for the change in the cost distribution from high to low as shown in Tables VIII–IX.

*Effect of $A_{cl}$:* Figure 5 shows that the network intensive applications gain the most for both cost distribution scenarios, whereas the compute intensive applications gain mostly for the low cost distribution scenarios due to the addition of DCs at the edge layer L3. The median cost savings for the network intensive application goes from 46% to 67% due to the addition of DCs at the edge layer L3, whereas it goes from 1% to 26% for the compute intensive applications as shown in Tables VIII–IX. The higher cost savings are more frequent for the network intensive applications, whereas the low and average cost savings are more frequent for the compute intensive applications for the high and low cost distribution scenarios respectively.

In case of both network and compute intensive applications, the median cost savings is lower than that of the network intensive applications.

### D. Summary:

The addition of DCs at the edge layer of a MEN is cost beneficial for the network intensive applications having high application demand locality and low DC cost distribution scenarios. For example, adding edge layer DCs can result into cost savings up to 67% for the network intensive applications. However, adding DCs at more intermediate layers can give an insignificant cost benefit to a TNO. Adding an intermediate layer DCs close to the root DC can be cost beneficial for the medium or low demand locality compute intensive applications.

The most important aspect when determining whether intermediate layer DCs are beneficial is application type. Demand locality is the second most sensitive parameter. The DC resource cost spread across the MEN topology is the third most sensitive parameter.

## IV. Related Work

As far as we know, the previous works look into the optimal placement algorithms for a given infrastructure, whereas our work investigates the placement and the optimal amount of the infrastructure resources for a given optimal placement algorithm. There are several introductory related works describing the challenges and defining the concept as Mobile Edge computing [13], Fog computing [14]–[16], Mobile cloud computing [17] and Telco-cloud [18]. [14] has defined characteristics of Fog as: a) Low latency and location awareness; b) Wide spread geographical distribution; c) Mobility; d) Very large number of nodes; e) Predominant access of wireless; f) Strong presence of streaming and real time applications; g) Heterogeneity. Another similar concept, CDN, addresses the static content placement by deploying cache servers at the edge of the Internet to reduce the download delay of contents from the remote sites [12], [19].

The model for the cost of a CPU-hour is explained in [6]. The author uses the model to compare the cost of leasing CPU time from online cloud service providers or purchase and use a server cluster of equivalent capacity. [5] explains the model for all the cost constituents for a DC. The cost for running a CPU in a large DC is lower to that of a smaller capacity DC considering economy of scale [5], [7]. Our computed CPU-h cost for a large DC is close to the cost charged by Amazon which is $0.026 per vCPU-hour for an EC2 small instance [20].

A detailed survey and analysis for broadband static and dynamic data pricing plans has been done in [3], [4]. They have pointed out that the latest network technological advances, e.g., 4G/LTE and wifi offloading can not meet the challenges of the future data demand especially from mobile data and video traffic due to expensive backhauling and increasing wired network congestion. The ISPs have been using usage-based pricing as a congestion control tool that may discourage users due to the uncertainty of their network budget expanses [4]. The cost for transferring 1 GB data through Bell's network is computed in [2] and static pricing is also motivated in the work. The computed link cost also matches with the cost charged by Amazon, which is between $0.05 and $0.09 for transferring out 1 GB of data from an EC2 instance for United States (U.S.) region [20]. A holistic cost model for an operator is explained in [21] where the total cost of running a network is modeled as the sum of the fixed network cost and the usage-based costs. [22] examines the cost and speed relationship for the broadband Internet access across in 24 cities in U.S. and abroad. Other leading reports in this area are Akamai's *State of the Internet* [23], the International Telecommunication Union's *The State of broadband*, the Organization for Economic Co-operation and Development's broadband portal, and the Federal Communication Commission's *Measuring Broadband America*.

IoT applications such as video surveillance for home security, virtual reality, high definition video conferencing, telemedicine, etc. will generate significant upstream demand up to several hundred megabits per second [1]. [24] has performed a survey about the integration of Cloud Computing and IoT. Many of the IoT applications deployed in the cloud can benefit from Machine-to-Machine (M2M) communications. Some other emerging application domains are Health care, Smart cities and communities, Automotive and smart mobility, Smart energy and smart grid, Smart logistics and Environmental monitoring.

## V. Conclusion and Future work

Moving the computations of the future bandwidth-hungry applications close to the end-users can alleviate the network congestion problem. One of the essential question for a TNO is to find out how much compute capacities need to be allocated at the intermediate layers' DCs to meet the challenge as well as maximize the cost benefit with the optimal placement. In this work, we study the benefit of introducing additional DCs closer to the network edge for the optimal application placement. The main parameters that affect the study are the DC cost distribution, application types and application demand locality. The application type is defined as the compute to network resource requirement ratio for an application request. The study shows that a TNO should start building an edge DC first having capacity up to hundreds of servers at the network edge to cater the emerging bandwidth-hungry applications and to minimize its operational cost. In future, we would like to investigate the effect of capacity constraints and applications Service Level Agreement (SLA) requirements on the cost benefits for multi-tier applications.

## Acknowledgment

Financial support has been provided by the Swedish Research Council (VR) under contract number C0590801 for the Cloud Control project.

## References

[1] C. F. Lam, "Fiber to the Home: Getting Beyond 10 Gb/s," *Optics and Photonics News*, vol. 27, no. 3, pp. 22–29, 2016.

[2] "Canada's Usage Based Billing Controversy: How to address the Wholesale and Retail Issues," http://www.michaelgeist.ca/wp-content/uploads/2011/03/GeistonUBB.pdf, accessed: 2016-05-09.

[3] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "A survey of smart data pricing: Past proposals, current plans, and future trends," *ACM Computing Surveys (CSUR)*, vol. 46, no. 2, p. 15, 2013.

[4] L. A. DaSilva, "Pricing for QoS-Enabled Networks: A Survey." *IEEE Communications Surveys and Tutorials*, vol. 3, no. 2, pp. 2–8, 2000.

[5] L. A. Barroso, J. Clidaras, and U. Hölzle, "The datacenter as a computer: An introduction to the design of warehouse-scale machines," *Synthesis lectures on computer architecture*, vol. 8, no. 3, 2013.

[6] E. Walker, "The Real Cost of a CPU Hour," *Computer*, vol. 42, no. 4, pp. 35–41, April 2009.

[7] S. Brumec and N. Vrček, "Cost effectiveness of commercial computing clouds," *Information Systems*, vol. 38, no. 4, pp. 495–508, 2013.

[8] "Don't worry- Mobile broadband is profitable," http://www.ericsson.com/ericsson/corpinfo/publications/ericsson_business_review/pdf/209/209_BUSINESS_CASE_mobile_broadband.pdf, accessed: 2016-05-09.

[9] C. Labovitz, S. Iekel-Johnson, D. McPherson, J. Oberheide, F. Jahanian, and M. Karir, "Atlas internet observatory 2009 annual report," *Arbor Networks Inc., University of Michigan and Merit Network Inc*, 2009.

[10] "How Big is AWS? Netcraft Finds 158,000 Servers," http://www.datacenterknowledge.com/archives/2013/06/04/how-big-is-aws-new-netcraft-numbers-show-insight/, accessed: 2016-05-09.

[11] M. Ferdman, A. Adileh, O. Kocberber, S. Volos, M. Alisafaee, D. Jevdjic, C. Kaynak, A. D. Popescu, A. Ailamaki, and B. Falsafi, "Clearing the clouds: a study of emerging scale-out workloads on modern hardware," in *Proceedings of the seventeenth international conference on Architectural Support for Programming Languages and Operating Systems*. ACM, 2012, pp. 37–48.

[12] S. Spagna, M. Liebsch, R. Baldessari, S. Niccolini, S. Schmid, R. Garroppo, K. Ozawa, and J. Awano, "Design principles of an operator-owned highly distributed content delivery network," *IEEE Communications Magazine*, vol. 51, no. 4, pp. 132–140, April 2013.

[13] M. Beck, M. Werner, S. Feld, and S. Schimper, "Mobile Edge Computing: A Taxonomy," in *Proc. of the Sixth International Conference on Advances in Future Internet*, 2014.

[14] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*. ACM, 2012, pp. 13–16.

[15] T. H. Luan, L. Gao, Z. Li, Y. Xiang, and L. Sun, "Fog computing: Focusing on mobile users at the edge," *arXiv preprint arXiv:1502.01815*, 2015.

[16] L. M. Vaquero and L. Rodero-Merino, "Finding your way in the fog: Towards a comprehensive definition of fog computing," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 5, pp. 27–32, 2014.

[17] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: architecture, applications, and approaches," *Wireless communications and mobile computing*, vol. 13, no. 18, pp. 1587–1611, 2013.

[18] P. Bosch, A. Duminuco, F. Pianese, and T. Wood, "Telco clouds and Virtual Telco: Consolidation, convergence, and beyond," in *2011 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, May 2011, pp. 982–988.

[19] G. Pallis and A. Vakali, "Insight and perspectives for content delivery networks," *Communications of the ACM*, vol. 49, no. 1, pp. 101–106, 2006.

[20] "Amazon EC2 Instance Pricing," https://aws.amazon.com/ec2/pricing/, accessed: 2016-05-09.

[21] M. Motiwala, A. Dhamdhere, N. Feamster, and A. Lakhina, "Towards a cost model for network traffic," *ACM SIGCOMM Computer Communication Review*, vol. 42, no. 1, pp. 54–60, 2012.

[22] N. Russo, R. Morgus, S. Morris, and D. Kehl, "The cost of connectivity," *New American Foundation*, 2014.

[23] D. Belson, "Akamai state of the Internet report, q4 2009," *ACM SIGOPS Operating Systems Review*, vol. 44, no. 3, pp. 27–37, 2010.

[24] A. Botta, W. de Donato, V. Persico, and A. Pescapé, "Integration of cloud computing and internet of things: a survey," *Future Generation Computer Systems*, vol. 56, pp. 684–700, 2016.