A Virtual Machine Location Problem for IP Multimedia Subsystem

Imen Limam Bedhiaf, Omar Cherkaoui, University of Quebec at Montreal, and Guy Pujolle, University of Pierre and Marie Curie
limam.imen@courrier.uqam.ca
cherkaoui.omar@uqam.ca
Guy.Pujolle@lip6.fr

***Abstract*—Virtualization gives the Mobile Virtual Network Operators (MVNOs) the possibility to deploy their components more rapidly and at a lower cost. By optimizing multiple virtual machines (VMs) in the same physical server, we considerably reduce the cost and power consumption. Therefore, server virtualization can be exploited to run VMs on servers that provide the lowest delay to their users. Our goal is to install MVNO virtual machines on a set of host operator physical equipment, such that user latencies are minimal while all capacity constraints are satisfied. We formulate our VM location problem as a mixed integer-programming problem. We study the complexity of the problem. Then, we conduct a sensitivity analysis to study the impact of the user latencies change on the optimal solution. To reach a good feasible solution in less time, we propose an heuristic. It reaches solutions close to the optimal in a very little time. The running time improvement compared to the Branch and Bound model exceeds 90% for larger topologies having more than 500 nodes. Our algorithm estimates the maximal number of VMs that can be created and then placed by our heuristic to provide the minimal latency subject to given capacity and cost budget. Optimizing the VM location contributes efficiently in the performance improvement of the SIP signaling delay.**

***Index Terms*—IMS, Latency, Location, MVNO, Optimization, Virtualization.**

I. Introduction

The concept of virtualization was first introduced in mainframe environments in the late 1960s. It is a technology that decouples logical resources from the physical infrastructure supporting them by adding a layer of abstraction between the applications and the hardware [1]. In our paper, we propose to apply virtualization to the IP Multimedia Subsystem (IMS). MVNOs share different parts of the Mobile Network Operator (MNO) equipments. Virtualization provides the MVNO a transparent and secure access to control its virtualized components in a total isolation of other MVNOs and MNO's virtualized components. Virtualization also optimizes resource sharing with the MNO at a low cost.

Universal Mobile Telecommunication System (UMTS) represents the most common architecture used by both MVNOs and MNOs. An important feature of UMTS Release 5 [2] is the IMS which works in conjunction with the Packet Switched Core Network (PS-CN) to support telephony and multimedia services. The Session Initiation Protocol (SIP) is rapidly being adopted as the signaling protocol [3] for these services. In this paper, we will focus on the Proxy Call Session Control Function (P-CSCF), which forms the first contact point of the IMS terminal, authenticates the user and establishes an IPsec security association with the IMS terminal. It also generates charging records. We propose to virtualize the P-CSCF component, as illutsrated in Figure 1, since the location of the P-CSCF, as an entry point to the IMS, contributes to a large part of SIP signaling delay [4]. In this case, the MVNO, thus, has its own P-CSCF hosted in a VM (Virtual Machine) while using the other components of the MNO. This scenario allows the MVNO to access to the control part. Based on this scenario, We propose to optimize the location of these virtual machines (VMs) into the suitable MNO physical servers ensuring a minimum of latency for the MVNO subscribers. We can similarly formulate the problem of optimizing the other IMS components. Our optimization task is a decision problem. We have to select the suitable P-CSCF where to route the subscribers connected to a common access point considered in this study as a user group.

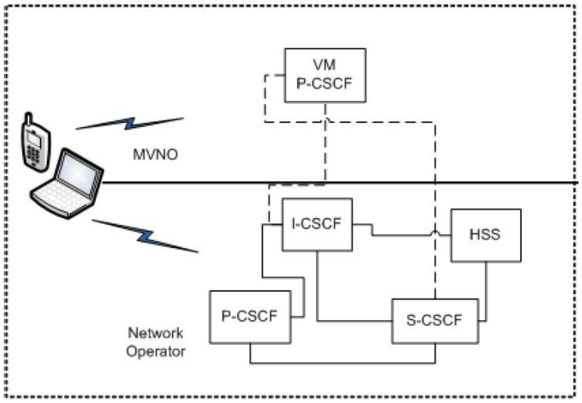

Fig. 1. Adopted scenario: virtualization of the P-CSCF.

We can classify subscribers based on three common criteria defined by most of the network operators. The first classification is conducted according to the quality of service (QoS) and services chosen by the subscribers. The second classification is done according to the profile of the users (business class, residential class, traveling class..). Finally, the last classification is based on the subscriber geographic location. In our study, we used this last classification which has the advantage to group the users who have the same network delay (latency). Subscribers attached to the same access point generally share the same geographic location. The network delay of subscriber groups is measured between the access point and the P-CSCF. Therefore, all the subscribers within the same group have the same measured network delay. The goal of the study is to determine the best VM locations among the available physical MNO servers that minimize the user group latencies. The remainder of this paper is structured as follows. Section 2 analyses the different related works. Section 3 outlines the formulation of the virtual machine location problem. Section 4 studies the computational complexity of our problem. Section 5 presents the results of our problem resolution using LPsolve as a solving engine. Section 6 analyzes the sensitivity of our location problem. Section 7 proposes an heuristic algorithm and evaluates its performance. Finally, section 8 concludes our study and presents the direction of future work.

II. Related Work

What distinguishes our work from other studies is our interest in the performance improvement of the SIP signaling delay while saving physical resource and deployment time. Minimizing the network cost is crucial for the wireless networks to achieve the QoS objectives. [5] and [6] optimize the assignment of cells to switches. They formulate the problem as an integer-programming and propose some heuristics to solve it. In our work, we optimize both the positioning of VMs in potential locations and the assignment of users to these VMs hosting P-CSCF servers. [7] focus on SIP session setup delay and propose optimizing it using an adaptive retransmission timer. It evaluates SIP session setup performances with various protocols. [8] also concentrates on IMS and SIP session setup delay by presenting a model for cross-layer performance. In our study, we exploit virtualization to improve the signaling delay by placing virtualized P-CSCFs the nearest to the users.

Different works concentrates on the VM to PM (Physical Machine) mapping. In [9], the authors propose a joint virtual machine placement to minimize the traffic costs in data center. In [10], the authors propose energy efficient resource management system for virtualized Cloud data centers that reduces operational costs and provides required Quality of Service by consolidating the different VMs. Authors in [11] conduct a survey of research in energy-efficient cloud computing in data center and detais the different algorithm used to place VMs while minimizing energy. In [12] authors propose a dynamic VM to PM mapping algorithm, through migration and consolidation, that reduces the amount of PM capacity required to support a specific rate of SLA violations. Moreover, at a fixed amount of PM capacity, it reduces the rate of SLA violations. However, in order to benefit from this performance, it assumes the presence of variable workloads that are easily forecasted. Authors in [13] propose another dynamic VM to PM mapping algorithm based on an autonomic controller that manages this mapping in accordance to user specified policies such as power consumption reduction. This autonomic controller help bypass the computational difficulty of manually finding and optimizing the mapping problem solutions. More importantly, it easily expresses the optimization objectives and constraints directly from specified user policies. In another research work in [14], the VM to PM mapping problem is solved by taking advantage of the resource allocation parameters that are provided as features of common hypervisors such as VMWARE ESX and XEN. Usually, three parameters are specified for each VM : its minimum guaranteed allocation, its maximum allocation and its weight in its contention over the shared resources. Based on this parameters, authors propose a smooth mechanism for mapping VMs to PMs while taking power-performance tradeoffs in consideration.

In [15], the placement of virtual machines across multiple clouds is studied with the objective of minimizing users' budgets. It considers the two available billing plans : reservation based and on demand.

In a related research work, the consolidation and migration of VMs put more stress on the network that have to be virtualized to support more bandwidth between servers using virtual path splitting. In this context, the network mapping problem is studied in [16].

In [17], it is the data transfer time consumption between VMs that is considered in the placement optimization problem. In the same direction, in [18], authors optimize the VM placement by considering the traffic pattern between VMs. It places VMs with high traffic in between within a short network distance. In previous works, minimizing the network delay between VM clients and VMs, rather than between VMs, was not considered explicitly in the VM placement problem.

In this paper we formulate the VM to server mapping problem as an optimization problem with the objective of minimizing the latency while all capacity and VM installation cost constraints are satisfied. We show that the problem is NP complete, we solve it using Branch and Bound method and greedy heuristic and we conduct a sensitivity analysis of its solution. To the best of our knowledge, this was never attempted before on this type of VM to PM mapping problems for IMS and delay sensitive services.

III. Virtual Machine Location Problem

A very interesting advantage of server virtualization is the capability of consolidating geographically distributed physical servers belonging to different local infrastructure providers and to share them between different service providers. Thus, it will be possible to run VMs on physical servers that are virtualized and available at places geographically close to their clients [19]. The benefits should be significant for delay-sensitive services such as multimedia telephony signaling. In this paper, we deal with the problem of mapping VMs to physical servers for delay sensitive services which can be modeled as an optimization problem. In the problem's objective, we minimize the round trip time (RTT) delays between VMs and their correspondant clients. With the capacity constraint consideration, minimizing the RTT in the objective, leading to better consolidation and less power consumption.

A. Assumptions

1) Network Delay: In the rest of the paper, we refer the delay to be optimized in the objective functions as the network delay between the clients and their serving VMs. It depends on the number of hops, the speed and length of links along the network path between clients and their VMs. We consider an M/M/1 queuing model at the different network nodes. The network delay D_{jk} is expressed as the sum of the queuing, transmission and propagation delays of the different intermediate nodes and links between the group of clients j, having the same RTT, and their served VM k. We consider a user group as a set of subscribers attached to the same access point. D_{jk} can be formulated as:

$$D_{jk} = \sum_{n=1}^{R} T_{Q_n} V_{jkn} + \sum_{l=1}^{B} (T_{t_l} + T_{p_l}) H_{jkl} \tag{1}$$

where V_{jkn} is a three dimensional binary routing matrix that indicates the nodes constituting the path between the client group's access j and its serving VM k; H_{jkl} is a similar matrix that indicates the links constituting this path. T_{Q_n}, T_{t_l} and T_{p_l} represents respectively the queuing, transmission and propagation delay

2) Installation cost: In this study, we consider the installation cost as the installation time C_{ik} of the virtual machine i over the physical equipment k. Let us define $\nu = \{\nu_1, ...\nu_i, ..., \nu_M\}$ as the set of virtual machines to be created for the MVNO. Each virtual machine is characterized by its installation time which is expressed as follows [1]:

$$C_{ik} = t_{A_{ik}} + t_{C_{ik}} \tag{2}$$

where $t_{A_{ik}}$ and $t_{C_{ik}}$ are defined as :

- $t_{A_{ik}}$ (virtual application time) is the time that the application packages containing the software modules take to be executed in a virtual machine.
- $t_{C_{ik}}$ (creation time) is the time taken to create a virtual machine over a physical equipment.

B. Problem Formulation

Let us consider a mobile communication network formed by N P-CSCF physical nodes belonging to different MNOs. A set of M virtual machines which constitute the MVNO P-CSCF components that must be created and placed in the physical MNO nodes in order to allow the MVNO to have control over his machines and manage his traffic. Moreover, let us consider a set of K subscriber groups. Our decision problem is to select the suitable P-CSCF where to route the subscribers connected to a common access point considered in this study as a user group. We start from the premise that the existing P-CSCF physical nodes must be used to locate the P-CSCF virtual machines, since it saves cost and allow the MVNOs to have more control on their offers. Thus, our virtual machine location problem (VMLP) consists of selecting a maximum of M nodes out of the N which form the MNO physical nodes, in order to locate in them the M MVNO virtual machines. We can locate more than a VM in the same physical node.

1) Input Definition: The inputs of the problem can be delineated as

- M: is the number of MVNO P-CSCF virtual machines to be placed in the network.
- N: is the number of P-CSCF physical equipment belonging to the MNOs.
- K: is the number of MVNO user groups.
- C: is a $M \times N$ dimensional installation cost matrix where $C_{ik} \geq 0$ is the installation time of a VM i over a physical equipement k.
- D: is a $K \times N$ dimensional network delay matrix where $D_{jk} \geq 0$ gives the delay of connecting group j to the P-CSCF physical equipment k.
- W: is a group weight vector, $W = [w_1, ..., w_K]$. It represents the traffic generated by the users of the group j
- ρ: is a VM traffic capacity vector, $\rho = [\rho_1, ..., \rho_M]$. It is the maximum of traffic that can be supported by the VM k to have an acceptable delay.
- α: is a server traffic capacity vector, $\alpha = [\alpha_1, ..., \alpha_N]$. It is the maximum of traffic that can be supported by the server i to have an acceptable delay.
- γ: is the maximum allowable VM installation cost vector over the different physical equipment, $\gamma = [\gamma_1, ..., \gamma_N]$

2) Variable Definition: The variables of the problem can be defined as follows

- Y: is a $M \times N$ dimensional binary VM location matrix such that its elements are given by: $y_{ik} = \begin{cases} 1 \text{ , if the VM } i \text{ is created in the physical equipment } k \\ 0 \text{ , otherwise.} \end{cases}$
- X: is a $K \times N$ dimensional binary matrix such that its elements are defined as follow: $x_{jk} = \begin{cases} 1 \text{ , if the group } j \text{ is connected to the machine } k \\ 0 \text{ , otherwise.} \end{cases}$

3) Location Problem Formulation: This problem can be expressed as a mixed integer-programming (MIP) problem [20] [6] with the objective of minimizing the network delay of MVNO subscribers. Based on the inputs and variables, our problem (P) can be formulated mathematically as follows:

$$minZ = \sum_{j=1}^{K} \sum_{k=1}^{N} D_{jk} x_{jk}$$

$$Subject\ to: \sum_{i=1}^{M} C_{ik} y_{ik} \leq \gamma_k,\ k = 1, 2.....N \tag{3}$$

$$\sum_{k=1}^{N} x_{jk} = 1,\ j = 1, 2.....K \tag{4}$$

$$\sum_{k=1}^{N} y_{ik} = 1,\ i = 1, 2.....M \tag{5}$$

$$x_{jk} \leq \sum_{i=1}^{M} y_{ik},\ j = 1, 2.....K,\ k = 1, 2, ...N \tag{6}$$

$$\sum_{i=1}^{M} \rho_i y_{ik} \leq \alpha_k,\ k = 1, 2, ...N \tag{7}$$

$$\sum_{j=1}^{K} w_j x_{jk} - \sum_{i=1}^{M} \rho_i y_{ik} \leq 0,\ k = 1, 2.....N \tag{8}$$

$$x_{jk} \in \{0, 1\};\ y_{ik} \in \{0, 1\} \tag{9}$$

The first constraint makes sure that the VM installation cost over the physical IMS server chosen do not exceed the maximum cost fixed by the MVNO. The second constraint ensures that a user is connected to only one P-CSCF. The third constraint ensures that VM i is created in only one physical equipment. In the fourth constraint, we make sure that if there is no VM created on a physical equipment we can not have a user group connected to it. The fifth constraint guarantees that the sum of the traffic capacities of the VMs created on the physical equipment k does not exceed its capacity. Constraint 6 ensures that the sum of the capacities of the VMs created on the physical machine k cannot be smaller than the sum of the weights of the groups assigned to them. The weight of a group is the traffic generated by its subscribers in number of requests per second in busy hour. Constraint 7 precises that the variables are binary.

IV. Computational complexity

The time complexity estimates the magnitude of the number of steps to solve an instance of the problem. In the following proposition, we use a proof by restriction [21] [20] to show that our problem belongs to a notably difficult class of problems. The proof by restriction consists on showing that our problem contains a known NP-complete problem as a special case [22] [6].

Proposition 1: Problem (P) is NP-complete.

Proof: Let $s(x_{jk}) = \sum_{j=1}^{K} D_{jk} x_{jk}$ be the cost (latency) of assigning subscribers to P-CSCF server k, $k = 1, 2.....N$. Let $v(x_{jk}) = \sum_{j=1}^{K} w_j x_{jk}$ and $B = \min_k \alpha_k$. We consider the restricted case of P where we have no constraint on the VM installation cost: $C_{ik} = 0$ for all $i \in \{1, 2.....M\}$ and all $k \in \{1, 2.....N\}$. We also relax the problem by omitting constraints (7), (8) and (9). Let U be the set of all the P-CSCF servers and $U' \subset U$ be the subset containing the selected P-CSCF where to locate the VMs. When we combine constraint (10) and (11), we have $\sum_{j=1}^{K} w_j x_{jk} \leq \alpha_k$. Problem P reduces to

Find optimal U' to minimize $\sum_{k \in U'} s(x_{jk})$

Subject to $\sum_{k \in U'} v(x_{jk}) \leq B$

This is the Knapsack problem which is NP-complete [21]. Thus the knapsack problem is a special case of our VM location problem. We can conclude by restriction that our problem is also NP-complete. ■

To solve P, we use the Branch and Bound method (ILP technique) which is an enumeration algorithm exploring all the possible branches of the tree. This method have an exponentiel running time with an increasing number of variables and constraints. To reduce this running time, we propose in section VII a polynomial time heuristic.

V. Computational Experience

We used LPsolve [23] as the ILP solve engine that uses the Branch and Bound method which is the most commonly-used algorithm for solving ILPs. For our experiments, we vary the number of MVNO user groups and evaluate the objective function, the positioning of the VM and groups. We consider a network with 10 P-CSCF servers $(N = 10)$ connected through NSFnet network. The propagation delays differ from one link to another. This set $F = \{2; 3; 4; 5; 6; 8; 10; 12; 14\}$ represents the propagation delays of the different links in milliseconds. The SIP average packet size is 500 bytes. We set $M = 4$ and we vary the number of groups K. Our installation cost matrix is given by:

$$C = \begin{pmatrix} 180 & 200 & 230 & 270 & 320 & 380 & 450 & 385 & 467 & 397 \\ 190 & 210 & 240 & 280 & 330 & 390 & 460 & 404 & 494 & 424 \\ 200 & 220 & 250 & 290 & 340 & 400 & 470 & 414 & 504 & 434 \\ 210 & 230 & 260 & 300 & 350 & 410 & 480 & 424 & 514 & 444 \end{pmatrix}$$

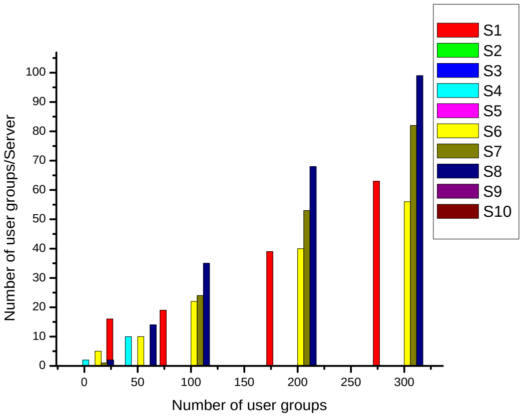

Fig. 2. Subscriber's positioning.

Fig. 2 shows the assignment of the different user groups to the selected P-CSCF servers. We denote by S_i the server i. The capacity utilization of the VMs and physical servers increases with the number of groups. When this utilization exceeds the maximum allowed capacity, the groups with larger latency are assigned to other servers. When the number of groups increases from 10 to 20, the selected servers change: S_4 is replaced by S_1. This new selection provides a better objective function. The solution guarantees a global optimal latency for all the groups but not for each group separately.

VI. Sensitivity Analysis

As the traffic between the access points and P-CSCF servers vary, we concentrate in this section on the impact of network delay change. We also study the effect of VM installation cost change. The integer program is solved in section V. Here, we are interested in the effect of small changes in the right-hand side and objective function coefficients. To gain further insight into the behavior of the model, we conduct a sensitivity analysis over the network delay D_{jk} (objective function coefficients) and maximum allowable installation cost γ_k (right hand side of the first constraint). A good solution should tolerate the deviation in the parameter values from their design values. After obtaining the optimal solution X^* and Y^* of our problem (P), the range of allowable change in the parameters in order to keep the current optimal solution, can be deduced from the sensitivity study. In our approach, we restrict the change to a single parameter at a time. The range of optimality is obtained whereby any further change in the selected parameter value would trigger a change in the optimal solution X^* and Y^*. This will allow us to determine the parameter limit (lower/upper limits).

A. Change in D_{jk}

We assume a change in the network delay value D_{jk} and denote the new value by $D'_{jk} = D_{jk} + \Delta D_{jk}$. Therefore, our objective function becomes:

$$min\ Z = \sum_{j=1}^{K}\sum_{k=1}^{N} D'_{jk}x_{jk} = \sum_{j=1}^{K}\sum_{k=1}^{N}(D_{jk} + \Delta D_{jk})x_{jk} \quad (10)$$

We keep the constraints unchanged. We note (P') our new problem.

1) Change limits: As for the original ILP problem (P), we use the Branch and Bound method to solve the new problem (P') and find the ΔD_{jk} limits while keeping the original optimal solution X^* and Y^* unchanged. We begin the Branch and Bound method by a LP relaxation of the problem (P'). In this relaxation, we replace the constraints $x_{jk} \in \{0, 1\}$ and $y_{ik} \in \{0, 1\}$ by $0 \le x_{jk} \le 1$ and $0 \le y_{ik} \le 1$.

Definition 1: Given $\zeta_j = \{D_{j1}, D_{j2}, ..., D_{jk}, ...D_{jN}\}$ the set of network delays for a group j with the N different P-CSCF physical servers, we define ζ'_j as the increasing ordered set of ζ_j such that $D^r_{jk} \in \zeta'_j$ is the r^{th} element and $D^r_{jk} \le D^{r+1}_{jk'}$ where k and $k' \in \{1, 2, ..., N\}$. ζ'_j contains only the network delay between the user group's access points and the selected P-CSCF servers in the original optimization (P) (corresponding to $y_{ik} = 1$).

Theorem 1: We suppose that X^* is the optimal solution of (P') and $D^r_{jk} \in \zeta'_j$ is the corresponding network delay of x^*_{jk} where $x^*_{jk} = 1$.

(i) If $0 \le D^r_{jk} + \Delta D^r_{jk} \le D^{r+1}_{jk'}$ then the optimal solution does not change.

(ii) If $D^r_{jk} + \Delta D^r_{jk} > D^{r+1}_{jk'}$ and $x_{jk'}$ satisfies all the constraints then the optimum changes and the values become $x^*_{jk'} = 1$ and $x^*_{jk} = 0$.

Proof: (i) If $0 \le D^r_{jk} + \Delta D^r_{jk} \le D^{r+1}_{jk'}$ then D'_{jk} still represents the smallest delay of the group j with the selected P-CSCF servers. There are no better solutions than the actual one. Therefore, the optimal one does not change.

(ii) If $D^r_{jk} + \Delta D^r_{jk} > D^{r+1}_{jk'}$ then D'_{jk} is no more the optimal solution. There is a better delay in the set of selected servers. If the capacity of the server k' allows the addition of the group j then the new selected server is the k' (representing the next smallest delay of the group j in the ordered list of selected servers). ■

2) Results: In this section, we consider as previous M=4, N=10 and fix K=20. The installation cost matrix doesn't change and the network delay matrix (in milliseconds) for the first 5 groups is given by:

$$D = \begin{pmatrix} 43\ 26\ 77\ 61\ 51\ 50\ 75\ 57\ 106\ 64 \\ 65\ 49\ 51\ 52\ 15\ 66\ 47\ 41\ 70\ 53 \\ 38\ 53\ 25\ 51\ 123\ 35\ 62\ 52\ 73\ 49 \\ 47\ 73\ 67\ 40\ 63\ 15\ 43\ 69\ 42\ 57 \\ 46\ 72\ 66\ 39\ 62\ 14\ 42\ 68\ 41\ 56 \\ ... \end{pmatrix}$$

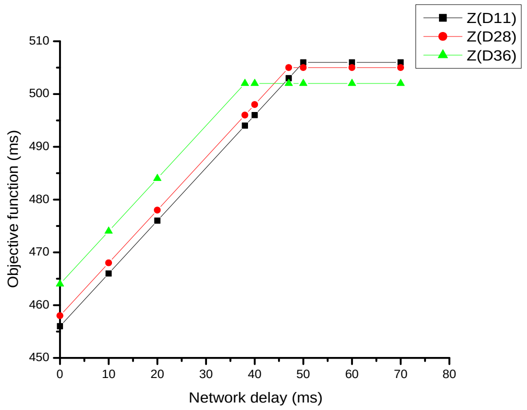

Fig. 3. Impact of D_{jk} change on the objective function value (case of D_{11}, D_{28} and D_{36}).

We use LPsolve to evaluate the sensitivity of the different objective function coefficients by relaxing the problem. We evaluate the range (limit) of each objective function coefficient separately without varying the other coefficients. This means that as this coefficient varies in this range the solution does not change. The values for the variables and the constraints will remain unchanged as long as the objective coefficient stays in this range. Obviously, the objective function value will vary and also the sensitivity information of the other variables. However, the solution will remain unchanged as we change a single coefficient in the objective. Fig 3 illustrates the sensitivity analysis of D_{11}, D_{28}, D_{36} on their respective objective function values $Z(D_{11})$, $Z(D_{28})$ and $Z(D_{36})$. It shows the impact of these coefficient changes on the objective function value. First, when D_{11} varies from 0 to 50, the optimum doesn't change ($x_{11} = 1, x_{28} = 1, x_{36} = 1, x_{46} = 1, x_{56} = 1, x_{67} = 1, x_{77} = 1, x_{86} = 1, x_{98} = 1, x_{107} = 1, x_{118} = 1, x_{121} = 1, x_{137} = 1, x_{147} = 1, x_{151} = 1, x_{168} = 1, x_{177} = 1, x_{186} = 1, x_{196} = 1, x_{208} = 1, y_{31} = 1, y_{16} = 1, y_{27} = 1, y_{48} = 1$ and all the other variables are equal to zero). Only the objective function increases. But when D_{11} exceeds 50, the optimum changes especially the value of x_{11} which becomes equal to zero. It's replaced by x_{16} which becomes equal to one. All the other variables remain the same. The first group becomes assigned to server S_6 instead of S_1. The objective function takes a constant value of 506. The value of 50, which constitutes the upper limit of D'_{11}, represents D_{16} the next delay value in the ζ'_1 set after D_{11}. It's the minimum network delay between user group one and the active servers selected to locate the VMs (corresponding to $y_{ik} = 1$). Then, when we vary D_{28} from 0 to 47 (corresponding to the D_{27} value), the optimum doesn't change. Only the objective function varies. When D_{28} exceeds 47, the optimum changes. x_{28} becomes equal to 0 and x_{27}=1. The objective function provides a constant value of 505. The second group becomes assigned to S_7 instead of S_8. Finally, when D_{36} varies from 0 to 38 (corresponding to the D_{31} value), the optimum remains the same and only the objective function varies. When D_{36} exceeds 38, the optimum changes:

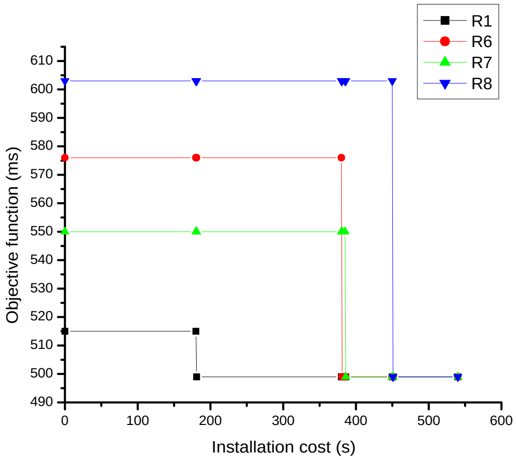

Fig. 4. Impact of γ_k change on the objective function value (case of γ_1, γ_6, γ_7 and γ_8).

x_{36}=0 and x_{31}=1. The objective function becomes constant (502). The third group becomes assigned to S_1 instead of S_6. We continued to evaluate the sensitivity of all the other objective function coefficients for different values of K, N and M and we arrived to the same conclusion at each time.

B. Change in γ_k

We assume changes in the maximum allowable VM installation cost values γ_k and denote the new values by $\gamma'_k = \gamma_k + \Delta\gamma_k$. Our first constraint becomes:

$$\sum_{i=1}^{M} C_{ik} y_{ik} \leq \gamma'_k = \gamma_k + \Delta\gamma_k, \;\; k = 1, 2.....N \qquad (11)$$

We keep the other constraints and objective function unchanged. We note (P'') our new problem.

1) Change limits: we use also the Branch and Bound method to solve the new problem.

Definition 2: We define by Z^*, the optimal solution of (P) and $Z'(\gamma'_k)$ the optimal solution of (P'') corresponding to the right hand side of constraint (1) $\gamma'_k = \gamma_k + \Delta\gamma_k$.

Theorem 2: Given $\psi_k = \{C_{1k}, C_{2k}, .., C_{Mk}\}$ the set of installation cost of the different VMs over the physical server k, we note by $C_k^{min} = min\{C_{ik} : C_{ik} \in \psi_k, i = 1, 2, .., M\}$. If $\gamma'_k < C_k^{min}$ then Z^* becomes $Z'(\gamma'_k)$ and the optimum X^* and Y^* changes.

Proof: If $\gamma'_k < C_k^{min}$ then no VM can be installed on the server k. This latter can't be selected so the optimum changes and becomes the solution of $Z'(\gamma'_k)$. ■

2) Results: In this section, we evaluate the sensitivity of the constraint (1) (for k =1, 6, 7 and 8) also called dual value, which specifies how much the objective function will vary if the constraint value is incremented in a specific range. Fig 4 illustrates the impact of γ_k change on the objective function value. We note the constraint (1) by R_1 for $k = 1$, R_6 for $k = 6$, R_7 for $k = 7$ and $R8$ for $k = 8$. We choose to study these cases because they are the only values of constraint (1) that have an impact on the optimum and the objective function. We start with the same optimum as for section A (optimum of the original problem (P)). When $\gamma_1 \geq 200$, the optimum doesn't change. But, when $190 \leq \gamma_1 < 200$, only changes in Y^* occurs: y_{31} becomes equal to zero and y_{21} to one whitout change in the objective function. Consequently, y_{26} and y_{36} change also. For $180 \leq \gamma_1 < 190$, also, only Y^* changes (y_{21}=0, y_{11}=1, y_{17}=0 and y_{27}=1) which results in the installation of VM_1 instead of the VM_2 in the first server because the installation costs of VM_2, 3 and 4 in server S_1 exceed the maximum installation cost (γ_1) allowed by the MVNO. For $0 \leq \gamma_1 < 180$, the objective function increases from 499 to 515 and S_1 isn't used anymore because it does no more satisfy contraint (1). It's replaced by S_4. All the optimum changes (X^* and Y^*) and becomes ($x_{16} = 1, x_{28} = 1, x_{36} = 1, x_{46} = 1, x_{56} = 1, x_{64} = 1, x_{77} = 1, x_{86} = 1, x_{98} = 1, x_{104} = 1, x_{118} = 1, x_{124} = 1, x_{137} = 1, x_{144} = 1, x_{154} = 1, x_{168} = 1, x_{177} = 1, x_{186} = 1, x_{196} = 1, x_{208} = 1, y_{14} = 1, y_{36} = 1, y_{27} = 1, y_{48} = 1$ and all the other variables are equal to zero). The value of 180 represents the value of C_{11}=min$\{c_{i1} : C_{i1} \in \psi_1, i = 1, 2.., 4\}$. For constraint R_6, when $0 \leq \gamma_6 < 380$, the objective function increases from 499 to 576 and the optimum changes. For constraint R_7, when $0 \leq \gamma_7 < 450$, the objective function increases from 499 to 550. Finally, for R_8 and for $0 \leq \gamma_8 < 385$, the objective function increases from 499 to 603.

VII. Proposed Heuristic

In the previous sections, we used the Branch and Bound method, which is an implicit enumeration providing bounds through linear programming relaxations. This method is efficient and reaches the optimal solution but its resolution time tends to be exponential. We need to find a good feasible solution quickly. The idea behind our heuristic is to reach a solution near the optimal in less time. We include dynamic programming in our resolution if changes occur in the latencies based on the sensitivity analysis conducted in section VI. Due to the complexity of the problem, exact methods cannot deal with problems with more than a few number of users. Consequently, we propose an heuristic based on the greedy algorithm. We denote by G_i the users belonging to the same group i. The goal is to reduce the subscriber latency by assigning (routing) them to their nearest P-CSCF in terms of network delays. Let denote by S the solution of the selected set of P-CSCF servers where to locate the MVNO virtual servers. $S \subset U$ where U is the set of all the N P-CSCF servers belonging to the different MNOs.

A. Algorithm

- Start with an empty set of P-CSCF servers where to locate VMs: set $S^0 = \oslash$ and $t = 0$.
- Choose the P-CSCF server k to add to $S^t = S^{t-1} \cup \{k\}$ whose additional cost $\sum_{j=1}^{K} min_{k \in S^t} D_{jk}$ is minimum.
- If the previous solution S^{t-1} is feasible and the cost has not decreased, stop with the solution $S = S^{t-1}$.
- Otherwise if the number of selected P-CSCF servers $|S| = M$ or $t = N$ (we tried all the servers), stop with $S = S^t$.

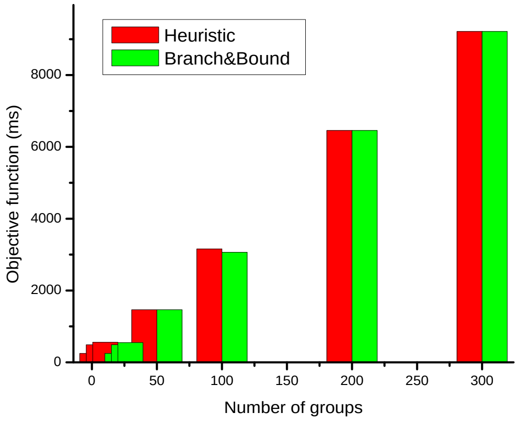

Fig. 5. Heuristic and Branch and Bound model objective function.

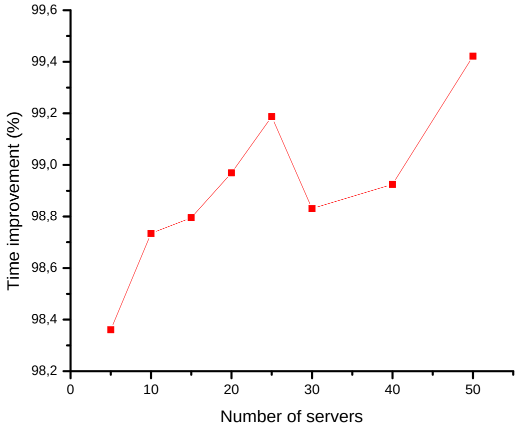

Fig. 6. Time improvement.

- Otherwise, $t \leftarrow t+1$ and return to step 2.
- If changes occur in the latency value of group G_j, we denote by D'_{jk} the new value: if $D'_{jk} > D_{jk'}$ and $x_{jk} = 1$ where $D_{jk'}$ is the latency with the next P-CSCF server in S, then set $x_{jk} = 0$, $x_{jk'} = 1$ and the new objective function becomes $Z_{new} = Z_{old} + (D_{jk'} - D_{jk})$.

This heuristic has a polynomial running time complexity. In fact, it stops after no more than M iterations. Recall that M is the number of VMs to be placed in the IMS network. The maximum number of selected servers is obtained when we place one VM per P-CSCF. In each iteration, we compute $(N-i+1)$ costs to find the minimum one where i is the iteration number. The number of maximum operation done by this algorithm to find a solution is $N+(N-1)+(N-2).....+(N-i+1) = \sum_{i=1}^{M}(N-i+1) = O(NM)$.

B. Results

To compare the results of the heuristic with those of the Branch and Bound model, we consider the same network as in section V. To evaluate the scalability of our heuristic, we conduct experiences with important number of user groups and larger topologies (large N and M). Fig 5 illustrates the heuristic and Branch and Bound model results by varying the number of user groups. The heuristic produces solutions near to those of the Branch and Bound in less time. We remark differences between the heuristic and Branch and Bound results for the values $K = 20$ and $K = 100$. For $K = 20$, the Branch and Bound model reaches a minimum of 550 ms for the objective function and the heuristic produces a solution of 559 ms which is very close to the Branch and Bound model solution. For $K = 100$, the heuristic reaches a different but closest solution to the Branch and Bound model, 3157 ms instead of 3067 ms.

We conduct experiences with important number of user groups and larger physical server to evaluate the scalability of our heuristic. For larger topologies and user group numbers, the time to find the optimal solution for the Branch and Bound model becomes intractable. It reaches 521895 ms for $K = 5000$ while the heuristic CPU time vary between 1 ms and 17 ms as a maximum value for $K = 10000$.

We consider the time improvement as a metric for measuring the heuristic performance. It's defined as $\frac{Branch\ and\ Bound\ model\ time - Heuristic\ time}{Branch\ and\ Bound\ model\ time}$. Fig 6 shows the time improvement function of the number of physical servers. The improvement varies between 94 % and 99 %. It reaches 99,4 % for a number of servers $N \geq 40$. The heuristic computation yield generally to a solution closest to optimal and in a little time.

Fig 7 illustrates the heuristic solution. It proposes an optimal repartition of the user groups by selecting the suitable P-CSCF among the available while minimizing the total latency of the different groups. For a given MVNO having 8 user groups, the optimal servers where the groups have to be placed are S_1, S_4, S_7 and S_8. This repartition ensures latencies of 13 ms for G_4, 14 ms for G_5, 39 ms for G_8 (these 3 groups are connected to S_1), 23 ms for G_1, 24 ms for G_6 and G_7 (these 3 groups are connected to S_8), 14 ms for G_3 connected to S_7 and 27 ms for G_2 connected to S_4. When the number of MVNO groups increases to 9, the optimum changes. The server S_2 is selected instead of S_4. The latency of G_2 increases due to this change but the overall QoS (latency) of the 9 groups becomes better. By adding new groups to the MVNO, the optimal solution could change from the oldest one even though the latencies didn't change. The new optimum (new group repartition) ensures a better global QoS for all the MVNO groups and not necessary the best latency for each one. The Fig 8 shows also the utilization increase of the VMs and physical P-CSCF servers by the increase of the MVNO group numbers.

For our next experiment, we set $K = 20$, $N = 10$ and vary M. Fig 8 demonstrates that when we increase the number of VMs to be placed among the available physical servers, the group latencies are reduced considerably. In fact, the objective function decreases from 914 ms to 391 ms. This decrease reaches a limit for a certain number of VMs (8 VMs in this experiment). There is no need to create more VMs to improve the QoS, i.e. the SIP signaling delay in this case, unless a need

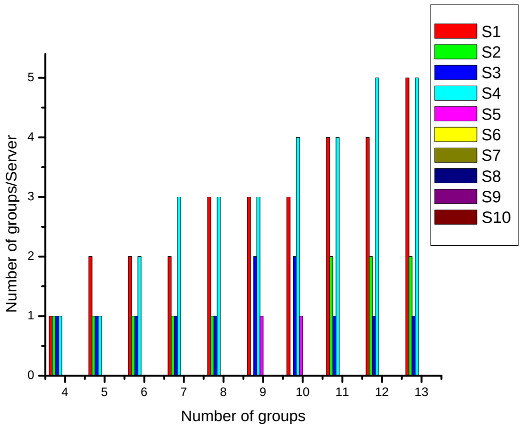

Fig. 7. User group repartition.

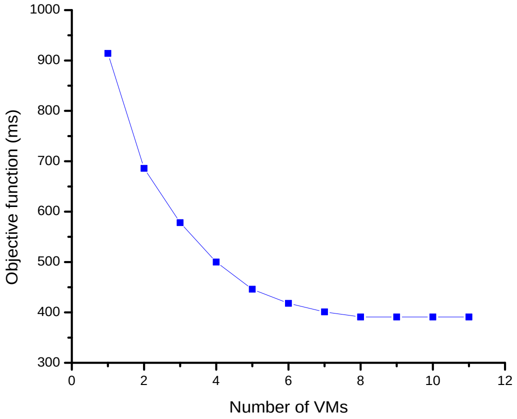

Fig. 8. Impact of VM number variation.

or constraint of capacity push the MVNO to do so. In fact, if the number of users per group increases considereably and exceeds the VM capacity, the creation of new VM in the same physical server is needed. A compromise could be done by the MVNO between the number of VMs to be created according to the needed capacity and an acceptable QoS for his different groups.

Fig 9 illustrates the variation of the latencies (objective function) as a function of the number of physical servers where the VMs could be placed. We set $K = 20$, $M = 4$ and vary N. This figure shows that a bigger set of available P-CSCF physical server from which to choose the VM positionning contributes in the decrease of latency. In fact, when the number of servers increases from 1 to 7, the objective function decreases by nearly 50%. This decrease has a limit. From a certain value (7 servers in this case), even though we increase the number of available phyical servers the latencies remain the same. There is no need to make available a bigger set of servers if they are farther than the actual ones toward the user groups.

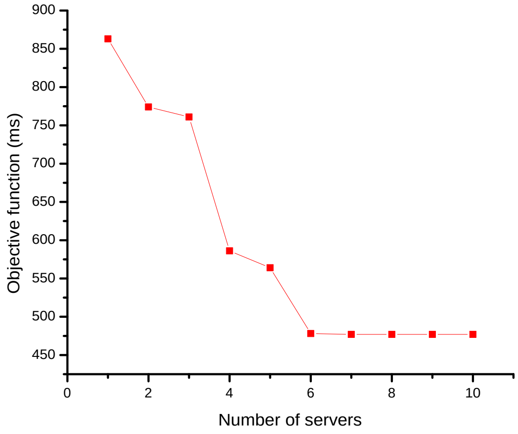

Fig. 9. Latencies function of the number of P-CSCF servers.

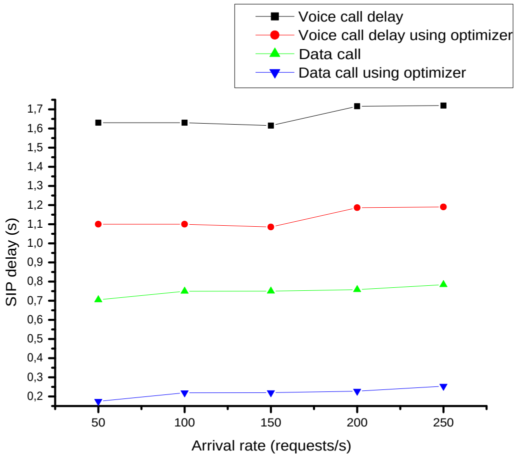

Fig. 10. SIP extablishement delay.

The VM location optimization impacts the total SIP delay via the decrease in the network delay between the user and the P-CSCF. The SIP processing delay doesn't change. To evaluate the performance improvement in the total SIP delay, we determine the session set up delay (voice and data) before and after the network delay optimization based on the formulation in [4]. We use OpenIMSCore as an open source implementation of IMS components. The mobile source and destination are connected and they communicate using UCT IMS Client. We generate an important quantity of SIP arrival rates using a traffic generator and analyze the packets with wireshark. We use Xen to create the different VMs. We will assume that the two subscribers (source and destination) use the same domain and that the MVNO needs to multiply each component by three to serve all of its subscribers. Fig 10 illustrates the performance improvement in SIP session delay when the VMs are placed geographically close to their users. We note that an improvement in the latency reduces the SIP delay by 0,53 s.

VIII. Conclusion

In this study, we consider the MVNO virtual machine location problem by optimizing the assignment of MVNO user groups into the different MNO P-CSCF servers. We conduct a sensitivity analysis to study the impact of the user group latency change and VM installation cost variation on the optimal solution. Then, we propose an heuristic which minimizes the SIP signaling delay of the MVNO groups by selecting the suitable physical server where to place the P-CSCF VMs. Our heuristic reaches solutions closest to the optimal in a very little time comparing to the Branch and Bound model. The running time improvement exceeds 90% for larger topologies. In fact, in these cases, the Branch and Bound model CPU time becomes intractable. We also present in our results the performance improvement in the SIP signaling delay. We record an improvement of 0,53 s in the control plane (voice and data call delays) due to the latency reduction. We notice also through our experiments that an increase in the number of VMs to be placed by the MVNO reduces by 50% the latency. Our heuristic provides us the number of VMs that have to be created and beyond which we can not improve the latency unless there is a need of capacity increase. As a future work, in order to benefit from the consolidation of virtual machines in fewer physical servers to save energy, we will consider traffic dynamicity and provide an online single reassignment algorithm that performs VM migrations that quickly improve the objective function regarding to workload variations without rerunning the whole heuristic.

References

[1] I. Bedhiaf, O. Cherkaoui, and G. Pujolle, "Third-generation virtualized architecture for the MVNO context," *Annals of Telecommunications*, vol. 64, no. 5, pp. 339–347, 2009.

[2] 3GPP, "Technical Specification Group Services and System Aspects; Network Architecture (Release 5)," *Technical Report TS23.002*, 2002.

[3] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "RFC3261: SIP: session initiation protocol," 2002.

[4] I. Bedhiaf and O. Cherkaoui, "Performance Characterization of Signaling Traffic in IMS," in *Communications (ICC), 2010 IEEE International Conference on*, 2010.

[5] I. Demirkol, C. Ersoy, M. Caglayan, and H. Deliç, "Location area planning and cell-to-switch assignment in cellular networks," *Wireless Communications, IEEE Transactions on*, vol. 3, no. 3, pp. 880–890, 2004.

[6] P. Bhattacharjee, D. Saha, and A. Mukherjee, "An approach for location area planning in a personal communication services network (PCSN)," *Wireless Communications, IEEE Transactions on*, vol. 3, no. 4, pp. 1176–1187, 2004.

[7] H. Fathi, S. Chakraborty, and R. Prasad, "Optimization of SIP session setup delay for VoIP in 3G wireless networks," *Mobile Computing, IEEE Transactions on*, vol. 5, no. 9, pp. 1121–1132, 2006.

[8] A. J. Miguel A. Melnyk and C. D. Polychronopoulos, "A Cross-Layer Analysis of Session Setup Delay in IP Multimedia Subsystem (IMS) With EV-DO Wireless Transmission," *IEEE Transactions on Multimedia*, 2007.

[9] S. H. M. C. M. C. J. W. Jiang, T. Lany, "Joint VM Placement and Routing for Data Center Trafc Engineering," *INFOCOM*, 2012.

[10] R. B. A. Beloglazov, "Energy Efcient Resource Management in Virtualized Cloud Data Centers," *IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, 2012.

[11] J. A. R. B. A. Beloglazov, "Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing," *Future Generation Computer Systems, Elsevier*, 2012.

[12] N. Bobroff, A. Kochut, and K. Beaty, "Dynamic placement of virtual machines for managing SLA violations," in *Integrated Network Management, 2007. IM'07. 10th IFIP/IEEE International Symposium on*, pp. 119–128, IEEE, 2007.

[13] C. Hyser, B. McKee, R. Gardner, and B. Watson, "Autonomic virtual machine placement in the data center," *Hewlett Packard Laboratories, Tech. Rep. HPL-2007-189*, pp. 2007–189, 2007.

[14] M. Cardosa, M. Korupolu, and A. Singh, "Shares and utilities based power consolidation in virtualized server environments," in *Integrated Network Management, 2009. IM'09. IFIP/IEEE International Symposium on*, pp. 327–334, IEEE, 2009.

[15] S. Chaisiri, B. Lee, and D. Niyato, "Optimal virtual machine placement across multiple cloud providers," in *Services Computing Conference, 2009. APSCC 2009. IEEE Asia-Pacific*, pp. 103–110, IEEE, 2010.

[16] M. Yu, Y. Yi, J. Rexford, and M. Chiang, "Rethinking virtual network embedding: Substrate support for path splitting and migration," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 17–29, 2008.

[17] J. Piao and J. Yan, "A Network-aware Virtual Machine Placement and Migration Approach in Cloud Computing," in *2010 Ninth International Conference on Grid and Cloud Computing*, pp. 87–92, IEEE, 2010.

[18] X. Meng, V. Pappas, and L. Zhang, "Improving the scalability of data center networks with traffic-aware virtual machine placement," in *INFOCOM, 2010 Proceedings IEEE*, pp. 1–9, IEEE, 2010.

[19] A. Greenberg, J. Hamilton, D. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," vol. 39, pp. 68–73, ACM, 2008.

[20] H. Pirkul, S. Narasimhan, and P. De, "Locating concentrators for primary and secondary coverage in a computer communications network," *Communications, IEEE Transactions on*, vol. 36, no. 4, pp. 450–458, 2002.

[21] M. Garey and D. Johnson, *Computers and intractability. A guide to the theory of NP-completeness*. WH Freeman and Company, San Francisco, Calif, 1979.

[22] A. Abutaleb and V. Li, "Location update optimization in personal communication systems," *Wireless Networks*, vol. 3, no. 3, pp. 205–216, 1997.

[23] "LPSolve Refrence Guide 2007," 2007.