IMS Network Deployment Cost Optimization Based on Flow-Based Traffic Model

Jie Xiao, Changcheng Huang and James Yan
Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada
{jiexiao, huang}@sce.carleton.ca, jim.yan@sympatico.ca

Abstract—The IP Multimedia Subsystem (IMS) is envisioned as the next-generation IP-based multimedia system that integrates data, speech, and video network services over both wireless and wireline networks. Modeling and design of IMS network has been an important area to both researchers and network providers. Our interest is in the area of developing efficient design models and optimization methods for IMS networks. In this paper, we focus on optimizing the cost of SIP server deployment in an IMS network. To reflect the traffic loads on the servers, a flow-based model is used to characterize the SIP traffic. Formulated as a linear programming problem, the cost optimization involves mapping a logical IMS core network topology into a physical network topology. Three potential mapping strategies are proposed. Each strategy's specific constraints are incorporated into the mathematical formulation of the problem. A numerical example of each strategy is presented, and the discussion on the formulations is provided.

***Index Terms*— IP Multimedia Subsystem, CSCF Server, SIP signaling Traffic, Cost Optimization, Mapping Strategy**

1. INTRODUCTION

IMS is envisioned as the next generation IP-based multimedia communication system that integrates data, speech, and video network technology and covers wireless and wireline networks. The IMS [1][1], as a new core network domain, was first introduced by the Third Generation Partnership Project (3GPP) in two phases (release 5 and release 6) [2] for Universal Mobile Telecommunications System (UMTS) networks. 3GPP2 further defined an IP multimedia framework, which finally harmonized with the IMS.

Figure 1 illustrates a simplified IMS core network architecture. IMS based networks consist of distinct Call/Session Control Function (CSCF) servers and Home Subscriber Server (HSS). There are three types of CSCF servers, the Proxy-, Interrogating- and Server-Call/Session Control Function servers (P-CSCF, I-CSCF, S-CSCF, respectively [1]). This paper only focuses on the network entities illustrated.

Based on IP technology, IMS provides a multimedia session control service that allows mobile users to access new multimedia and multisession applications as well as to establish synchronous multimedia sessions across fixed and mobile terminals **Error! Reference source not found.**[3][4]. Since the service creation interfaces are standardized by IMS, they allow for the development of new multimedia and multi-session applications. IMS offers this session control to the applications by Session Initiation Protocol (SIP) [5][6].

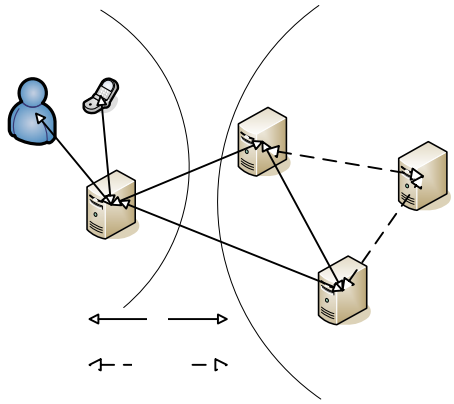

Figure 1: Simplified IMS Core Network Architecture

SIP, as an application layer control protocol, is defined by Internet Engineering Task Force (IETF) [6]. SIP lies at the core of IMS architecture and plays the role of session establishment, modification, and termination between two or more end users. Each CSCF server is actually a specialized SIP server.

Modeling and design of IMS network have always been an important area to both researchers and network providers. Our interest is in the area of developing efficient design models and optimization methods for IMS network. In our research, we focus on IMS network deployment cost optimization to produce a good network design potentially capable of securing considerable saving. In this paper, the mathematical modeling and the application of efficient optimization methods are applied. We, as designers, have to make selective use of various available theoretical models and different approximations, such as physical node capacity and physical link bandwidth. Also we consider various practical constraints in specific models.

In a real IMS network, the logical IMS network can be mapped to different IMS physical networks with different mapping strategies, where the logical CSCF servers are mapped into the physical node(s). In this paper, three potential mapping strategies are proposed. Then, each mapping strategy can be formulated as cost optimization issue, in a linear programming problem. Each strategy's specific constraints are incorporated into the mathematical formulation of the problem. In order to better understand how to formulate the cost optimization in a specific strategy, an example for each mapping strategy is provided, including the detailed procedures. The discussion on the formulations is provided at the end.

The challenge for studying IMS is the complexity of its signaling procedures. There are numerous signaling procedures

defined in IMS [3]. Each user agent (UA) may trigger a specific signaling procedure at a specific time depending on its specific call scenario at that moment. Estimating server loads based on individual signaling procedures is therefore not scalable. We proposed a flow-based traffic model that allows predicting the loads of IMS CSCF servers in a scalable way by utilizing the characteristics of IMS messages in [8]. A flow is an aggregation of signaling messages that traverse the same path in an IMS network. In [8] we demonstrated that all signaling procedures can be aggregated into 17 flows, as shown in Table 1. The flow concept has significantly simplified the process to estimate the loads of various CSCF servers while the correlation structure across the loads of these servers is still captured. For more details about the flow concept, readers are referred to [8]. In this paper, we assume that all signaling procedures have been aggregated into 17 flows that traverse various CSCF servers. We therefore focus on the issue of mapping CSCF servers to physical server nodes.

Table 1: Summary of 17 flows

Flows	Flow Path
1	→ P → S → --- → S → P →
2	→ S → P →
3	→ P → S →
4	→ I → S → P →---→ P → S → I →
5	→ S → P →---→ P → S →
6	→ P → S → I →
7	--- S ---
8	→ I → S →---→ S → I →
9	→ S → I →
10	--- P ---
11	→ P → I → S → I → P →
12	→ I → S → I →
13	→ I → S → P →
14	→ I →S →
15	→ HSS → I
16	→ HSS → S
17	→ S → HSS

Most of IMS-related research work currently has concentrated on IMS architecture and SIP protocol development [4], network performance evaluation under varying network parameters [3][9], and the Quality of Service (QoS) issue [10]. To our best knowledge, there is no such reference that provides the formulation on IMS network deployment cost optimization problem utilizing the flow-based traffic model.

2. MAPPING STRATEGY

The IMS servers (P/S/I-CSCF and HSS) illustrated in Figure 1 are all logical entities. In a real network, all logical servers need to be implemented on physical node(s). How to map logical servers located in a logical IMS core network topology to physical node(s) located in physical IMS network topology is not standardized, but is of great interest to the network providers. Network providers may choose different mapping strategies to achieve their own objectives. On the other hand, an industry-leading network provider may want a mapping strategy that provides high reliability and high expandability. Moreover, each mapping strategy has its own advantages and disadvantages.

Network providers select a mapping strategy with the best performance results according to their needs and actual network conditions, including the number of users, the capacity of physical nodes, the budget plan, and so forth. This requires the providers to consider both advantages and disadvantages of each mapping strategy, in order to determine the one that is satisfied by themselves and their users. In the following sections, we start with a generic mapping strategy, and then focus on two special mapping strategies, which can be widely used in the public domain.

2.1 Generic Mapping Strategy

The Generic Mapping Strategy is a method that allows for the mapping of a logical CSCF server into any physical node. The upper part of Figure 2 illustrates a logical IMS core network topology, which is the way that the IMS messages pass through the network from one logical server to the next without regard to the physical interconnection of the physical nodes. The loads of each logical server can be predicted by applying the flow-based traffic model [8]. However, in the physical structure of the IMS network, also called physical IMS network topology, which is depicted in the lower part of Figure 2, the loads of each physical node can be estimated according to the different mapping strategies, each determining how the logical servers are mapped into the physical node(s). .

Figure 2 shows the Generic Strategy of mapping logical servers in the IMS core network into physical nodes interconnected through a network. In this case, any physical node can host one or more logical server(s). On the other hand, two or more physical nodes can host one or more identical logical servers. Any two or more physical nodes can be identical, which means that they can host the same logical servers. The Generic Mapping Strategy includes all possible mapping ways.

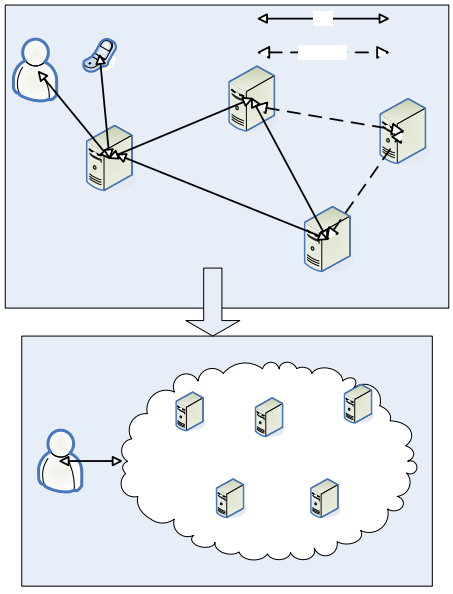

Figure 2: A Generic Mapping Strategy

Although the mapping strategy is called generic, certain constraint is necessary to minimize the search for the optimal solution. The constraint is related to the flow concept. All flows listed in Table 1 can be classified into two types: round trip flow and single trip flow. The round trip flow is defined as a flow that traverses the involved logical servers twice: one in the forward direction and one in the reverse direction. The single trip flow passes the involved logical servers only once. The selection of physical nodes for performing both directions in a round trip flow should be identical. This is because the physical nodes chosen in the forwarding direction may hold some information regarding the end users. It will minimize the information to be duplicated on different physical nodes by choosing the reverse path to be the same physical nodes except the order is reversed.

2.1.1 Notations for Network Modeling

As you will see, a good mathematical notation can represent a specific design problem in a compact and unambiguous way. It helps us to understand the formulation better.

Physical Node, Logical Server, and Flow

The four different logical servers in logical IMS network are labeled with the generic label v, where $v = 1, 2, 3, 4$, and the physical nodes are denoted as y, where $y = 1, 2, \cdots, Y$, and Y is the number of physical nodes in the network. A flow is denoted as f, where $f = 1, 2, \cdots, 17$. A direct physical link connects its physical end nodes directly.

Flow Demand, Physical Path

Flow demand volume is denoted as h_f, and it represents the traffic volume (number of messages) in a given unit of time. For flow *f*, the total number of available physical paths is denoted by P_f, and they are labeled with p from the first physical path to the total number of physical paths, i.e. $p = 1, 2, \cdots, P_f$. This sequence is called the list of candidate physical paths. Each physical path p connects the physical end nodes of flow *f*, and it is described as the set of physical links of which the physical path is composed of. In this paper, we assume the candidate paths for a flow are known to the carrier. A carrier may decide the candidate paths based on its own policy.

Figure 3 depicts an example of mapping the logical servers into four physical nodes, which are hosting the corresponding logical servers, as shown in the bracket. A list of physical paths that can carry flow 3 (flow path: →P→S→; f = 3) is drawn in the lower part of Figure 3. Table 2 lists the candidate physical paths for flow 3 under the network topology in Figure 3. Moreover, for physical path 5, physical node #3 can handle flow 3 alone.

Now, flow demand volume is assigned to the available physical paths. The loads assigned to physical path p, a candidate physical path of flow *f* are denoted by w_{fp}, as shown in Table 2. Since the demand volume of flow *f* needs to be realized by the traffic on all the candidate physical paths, we can write the following equation:

$$w_{31} + w_{32} + w_{33} + w_{34} + w_{35} + w_{36} = h_3 \quad \textbf{(1)}$$

It leads to the demand constraint, which can be written in a general form as follows:

$$\sum_{p=1}^{P_f} w_{fp} = h_f \quad \textbf{(2)}$$

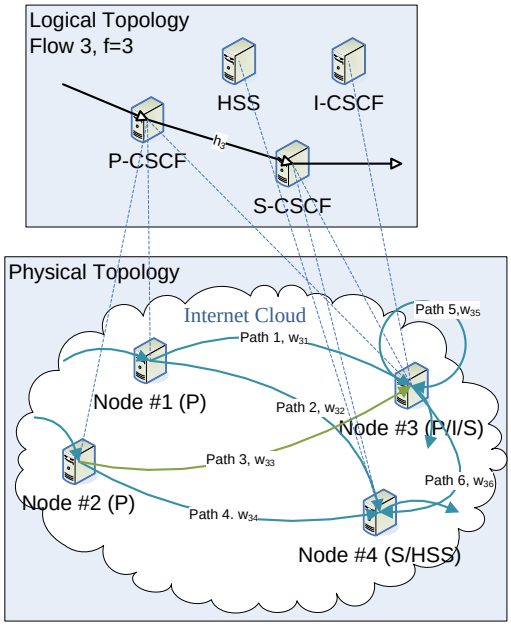

Figure 3: An Example of Mapping, 6 Available Physical Paths for Flow 3 (Flow Path: →P→S→)

Table 2: A List of Candidate Physical Paths for Flow 3, under the Network Topology in Figure 3

p	**Candidate physical paths for flow 3 (flow path: →P→S→)**	w_{fp}
1	Physical node #1 → #3	w_{31}
2	Physical node #1 → #4	w_{32}
3	Physical node #2 → #3	w_{33}
4	Physical node #2 → #4	w_{34}
5	Physical node #3	w_{35}
6	Physical node #3 → #4	w_{36}

Indicator

Two indicators are defined for formulating the design problem. The first indicator, denoted by α_{fpyv}, indicates the relationship among physical node *y*, logical server *v*, physical path *p*, and flow *f*. It is defined as:

$$\alpha_{fpyv} = \begin{cases} 1, \text{if physical node y hosts logical server v along} \\ \quad \text{physical path p of flow f} \\ 0, \text{otherwise.} \end{cases} \quad \textbf{(3)}$$

α_{fpvy} is constant and obtained from the analysis performed on network topology assuming that the carrier knows all candidate paths for each flow based on its policy.

The second indicator determines the number of times that logical server *v* is involved in flow *f*, and it is denoted by β_{fv}.

As discussed, there are two types of flow, which are the round trip flow and the single trip flow as summarized in Table 3. The path of each flow is provided in Table 1. The logical servers along a round trip flow are involved twice. And, the logical servers along a single trip flow are involved once. However, flow 11 and flow 12 are special cases due to the flow traverses logical S-CSCF server only once although they look like a round trip flow. β_{fv} is written as follows:

$$\beta_{fv} = \begin{cases} 2, \text{if logical server v is involved in flow f, and flow f is a round trip} \\ 1, \text{if logical server v is involved in flow f, and flow f is a single trip} \\ 2, \text{if logical server v is P - CSCF, for flow 11} \\ 2, \text{if logical server v is I - CSCF, for flow 11} \\ 1, \text{if logical server v is S - CSCF, for flow 11} \\ 2, \text{if logical server v is I - CSCF, for flow 12} \\ 1, \text{if logical server v is S - CSCF, for flow 12} \\ 0, \text{otherwise.} \end{cases} \quad \textbf{(4)}$$

Table 3: The Types of Flows

Types of flow	Flow
Single trip flow	#2, #3, #6, #7, #9, #10, #13, #14, #15, #16, #17
Round trip flow	#1, #4, #5, #8
Special flow	#11, #12

Load, Rate, Capacity

The cost of a physical node mainly depends on its capacity. Thus, the loads of physical node are predicted in order to determine the best choice for the capacity of a physical node.

The loads on each physical node indicate the number of actual IMS messages processed in a given unit of time, and we denote these loads as l_y for physical node *y*. As we can see in Figure 3, the loads of a physical node are calculated as the summary of the loads of the flows that traverse the node.

$$l_y = \sum_f \left[\sum_v \beta_{fv} \left(\sum_p \alpha_{fpvy} w_{fp} \right) \right] \quad \textbf{(5)}$$

In specific, this equation can be divided into three steps.

Step a: $\left(\sum_p \alpha_{fpvy} w_{fp} \right)$, accumulates the loads allocated to each available physical path *p* of the flow *f*, with a given logical server *v* and physical node *y*.

Step b: $\left[\sum_v \beta_{fv} \left(\sum_p \alpha_{fpvy} w_{fp} \right) \right]$, represents the total loads on physical node *y*, for flow *f*.

Step c: $\sum_f \left[\sum_v \beta_{fv} \left(\sum_p \alpha_{fpvy} w_{fp} \right) \right]$, sums up the loads for all the involved flows.

Our goal is to find the capacity of a physical node that can satisfy the loads requirement. Let c_y represents the processing capability of a physical node. c_y can be in various units that are related to the cost of the server. For example, c_y can be the number of certain CPUs the server carries.

Since in IMS network, each logical server has different message functions to be processed, we need to decide how much capacity is required to process messages for each type of logical server. This capacity coefficient is denoted as κ_v for logical server *v*. κ_v is in the unit of time-capacity product. For example, κ_v =2 can mean that a message requires two time units for a logical server with a single CPU or one time unit for a logical server with two CPUs. Here we assume the overhead associated with multiple CPUs is negligible to simplify the analysis. However our analysis can be generalized to include those overheads.

It should be noted that messages processed by the same logical server do not necessarily take the same amount processing time due to their different types. To simplify our analysis, we take κ_v as the statistical mean of the capacity required by all types of messages processed by the logical server.

Equation (5 calculates the loads of physical node *y*, which represent a sum of the loads on logical server *v* that is hosted in a physical node *y*. Hereby, the capacity of physical node *y* should be greater than the accumulation of the loads of logical server *v* times κ_v, for all possible logical server *v* that the physical node *y* hosts. This is a second set of constraints that can be generally written as:

$$\sum_f \left[\sum_v \kappa_v \beta_{fv} \left(\sum_p \alpha_{fpvy} w_{fp} \right) \right] \le c_y \quad \textbf{(6)}$$

Furthermore, in the design problem, we aim to minimize the physical node capacity cost. Therefore, a rate ε_y is introduced and it represents the cost per unit processing capability for physical node *y*.

2.1.2 Formulation of Cost Optimization Problem

IMS network deployment cost optimization issue is considered with a set of given flow demand volumes. A complete version of formulating the cost optimization issue as a linear programming problem for a Generic Mapping Strategy is provided below. When the physical network construction is given, the formulation can be formed.

- Indices:

- f =1, 2,…, 17, flow
- v = 1, 2, 3, 4, logical server
- p =1, 2, … P_f, candidate physical path for flow f
- y=1, 2, …Y, physical node

- Constants:

- α_{fpyv}: =1, if physical node *y* hosts logical server *v* along physical path *p* that is one available physical path of flow *f*; 0, otherwise.
- β_{fv} : number of times that logical server *v* is involved in flow *f*.
- h_f : flow demand volume for flow *f*.
- κ_v : capacity coefficient in time-capacity product unit for logical server *v*.
- ε_y : cost coefficient per unit processing capacity for physical node *y*.

- Variables:

- w_{fp} : loads allocated to physical path *p* of flow *f*.

- **Objective**: Minimize total network physical nodes cost.

$$F = \sum_y \varepsilon_y \cdot c_y \quad (7)$$

- **Constraints:**

- Demand Constraints:

$$\sum_p w_{fp} = h_f \quad (8)$$

- Capacity Constraints:

$$\sum_f \left[\sum_v \kappa_v \beta_{fv} \left(\sum_p \alpha_{fpvy} w_{fp} \right) \right] \le c_y \quad (9)$$

- Constraints on variables,

$$w_{fp} \ge 0 \text{ (continuous, non - negative)} \quad (10)$$

$$c_y \ge 0 \text{ (continuous, non - negative)} \quad (11)$$

According to Equation 8 and 9, the cost optimization issue can be formulated as a linear programming problem. In the optimal solution of this problem, all constraints presented in Equation 9 are binding, i.e., the physical node loads are equal to the physical nodes capacities; however, the capacity of the physical node may not come in continuous value. To reduce the unused capacity, we can set c_y to integer. Then the problem becomes an integer programming problem which is more difficult to solve.

2.1.3 Example

An example to show the formulation of the cost optimization problem is provided. The simple network is shown in Figure 4, with 5 physical nodes connected through a network. Each physical node is hosting one or more logical server(s), as shown in the bracket after the name of physical node. The assumptions are made as follows:

1. There are 3 flows involved; they are flow 11, 15, 16, i.e. $f = 11,15,16$. The corresponding physical paths are extracted from Figure 4 for each flow.
2. It is easy to see that the number of potential paths that can be used as candidate paths is large. To simplify the example, we assume only the paths listed in Table 4 are the candidate paths.

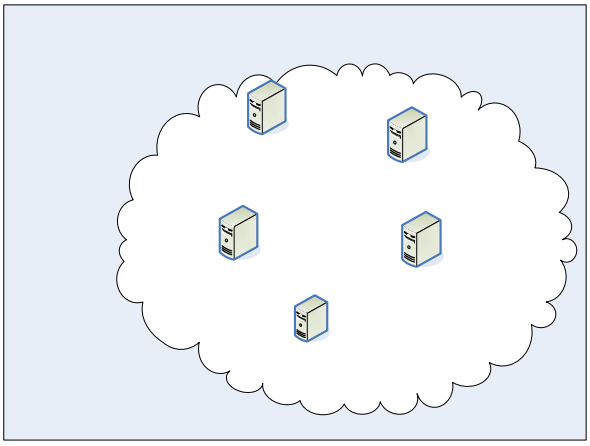

Figure 4: An Example of Formulating Cost Optimization Problem, with 3 Flows Involved, for a Generic Mapping Strategy

Table 4: The Candidate Physical Paths, Reference to Figure 4

p	The choice of physical nodes				Candidate physical paths for 3 flows
	P	I	S	HSS	**Flow 11 (→P→I→S→I→P→)**
1	#1	#4	#2		→#1(P) → #4 (I) → #2 (S) → #4 (I) →#1(P) →
2	#1	#4	#3		→#1(P) →#4(I) → #3 (S) → #4(I) →#1(P) →
					Flow 15 (→HSS→I→)
1		#4		#5	→#5 (HSS) → #4 (I) →
					Flow 16 (→HSS→S→)
1			#2	#5	→#5 (HSS) → #2 (S)→
2			#3	#5	→#5 (HSS) →#3 (S) →

The given information is provided as follows:

a) Flow demand volume,

$$h_f = \begin{cases} 400, \text{f} = 11 \\ 250, \text{f} = 15 \\ 200, \text{f} = 16 \end{cases} \quad (12)$$

b) Capacity coefficient for logical server *v*,

$$\kappa_v = \begin{cases} 2, \text{v} = 1 \\ 1, \text{v} = 2 \\ 5, \text{v} = 3 \\ 2, \text{v} = 4 \end{cases} \quad (13)$$

c) Cost coefficient for physical node *y*,

$$\varepsilon_y = \begin{cases} 5, \text{y} = 1 \\ 10, \text{y} = 2 \\ 10, \text{y} = 3 \\ 5, \text{y} = 4 \\ 5, \text{y} = 5 \end{cases} \quad (14)$$

The objective function can be written as:

$$\textit{Minimize} \quad F = 5c_1 + 10c_2 + 10c_3 + 5c_4 + 5c_5 \quad (15)$$

Subject to:

a) Demand Constraints:

$$w_{11,1} + w_{11,2} = h_{11} = 400 \quad (16)$$

$$w_{15,1} = h_{15} = 250 \quad (17)$$

$$w_{16,1} + w_{16,2} = h_{16} = 200 \quad (18)$$

b) Capacity Constraints:

$$2 \cdot 2 \cdot (w_{11,1} + w_{11,2}) \le c_1 \quad (19)$$

$$5 \cdot (w_{11,1}) + 5 \cdot (w_{16,1}) \le c_2 \quad (20)$$

$$5 \cdot (w_{11,2}) + 5 \cdot (w_{16,2}) \le c_3 \quad (21)$$

$$2 \cdot (w_{11,1} + w_{11,2}) + (w_{15,1}) \le c_4 \quad (22)$$

$$2 \cdot (w_{15,1}) + 2 \cdot (w_{16,1} + w_{16,2}) \le c_5 \quad (23)$$

c) Constraints on variables:

$$w_{fp} \ge 0, \text{ for} \begin{cases} p = 1, 2, \text{for f} = 11 \\ p = 1, \text{for f} = 15 \\ p = 1, 2, \text{for f} = 16 \end{cases} \quad (24)$$

2.2 *Customized Mapping Strategy 1*

Four logical servers (P/I/S-CSCF, HSS) that we concentrated on play different roles in IMS system. When using Customized Mapping Strategy 1 to map these logical servers to the physical nodes, each physical node only hosts one logical server, and it focuses on performing one type of tasks. The tasks on each physical node are clearly demarcated. This mapping strategy is desired when the loads of two or more logical servers exceed the capacity of a physical node. This is often the case for the network providers with a large number of users.

Overall, there are three advantages by using this mapping strategy. First, it is easier to create a backup physical node and upgrade capacity for the future. Second, this strategy brings a small impact to the system when the failure occurs on the physical nodes, because each physical node only takes care of one type of tasks. Third, it is easy to implement and maintain the physical nodes. The main disadvantage of this mapping strategy is cost. The remaining capacity of the physical nodes which host one type of logical server can not be allocated to other logical servers and therefore will be wasted.

Figure 5 depicts this mapping strategy, where each physical node hosts only one logical server. Multiple physical nodes, which host the same type of logical server, can then perform a load balance.

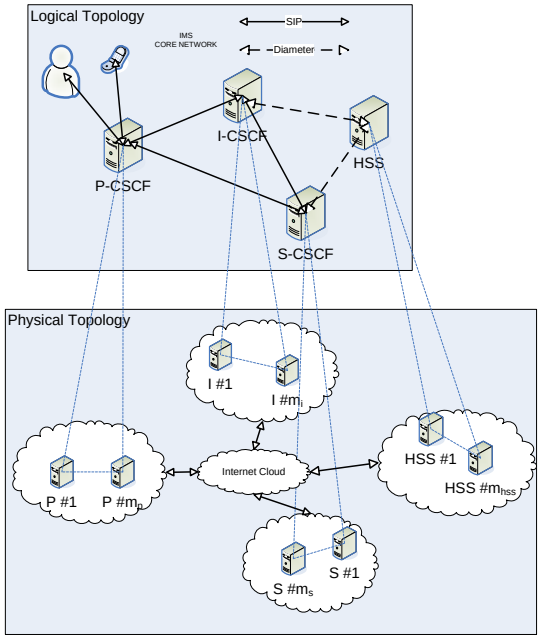

Figure 5: Customized Mapping Strategy 1

2.2.1 *Formulation of Cost Optimization Problem*

In this mapping strategy, it is easy to see that the physical nodes that support one logical server play loads balance among themselves. Because there is no sharing of residual capacities among different logical servers, the physical nodes that support different types of logical servers can be optimized separately. This can significantly simplify the optimization process because the number of variables to be optimized only depends on the number of physical nodes hosting the same logical server rather than the number of candidate paths. The optimization problem described in the last section can be reformulated as follows.

- Indices:

- f =1, 2,…, 17, flow
- v = 1, 2, 3, 4, logical server
- y_v=1, 2, …Y_v, physical node that hosts logical server v

- Constants:

- α_{fv} : =1, if flow f traverses logical server v; =0, otherwise.
- β_{fv} : number of times that logical server v is involved in flow f.
- h_f : flow demand volume for flow f.
- κ_v : capacity coefficient in time-capacity product unit for logical server v.
- ε_{vy} : cost coefficient per unit processing capacity for physical node y that hosts logical server v.

- Variables:

- w_{fvy_v} : loads allocated to physical node y of flow f for logical server v.

- Objective: Minimize total physical nodes cost for logical server v.

$$F_v = \sum_{y_v} \varepsilon_{vy_v} c_{vy_v} \quad \textbf{(25)}$$

- Constraints:

- Demand Constraints:

$$\sum_{y_v} w_{fvy_v} = h_f \alpha_{fv} \quad \textbf{(26)}$$

- Capacity Constraints:

$$\sum_f \kappa_v \beta_{fv} \alpha_{fv} w_{fvy_v} \le c_{vy_v} \quad \textbf{(27)}$$

- Constraints on variables,

$$w_{fvy_v} \ge 0 \text{ (continuous, non - negative)} \quad \textbf{(28)}$$

$$c_{vy_v} \ge 0 \text{ (continuous, non - negative)} \quad \textbf{(29)}$$

2.2.2 *Example*

An example is provided to formulate the cost optimization problem when using mapping strategy 1. The network topology illustrated in Figure 6 is the topology shown in Figure 4 with the relocation of logical servers in 5 physical nodes. The assumptions and given information remain the same as provided in Section 2.1.3. In this case, only Nodes #2 and #3 need to be optimized for logical server S-CSCF.

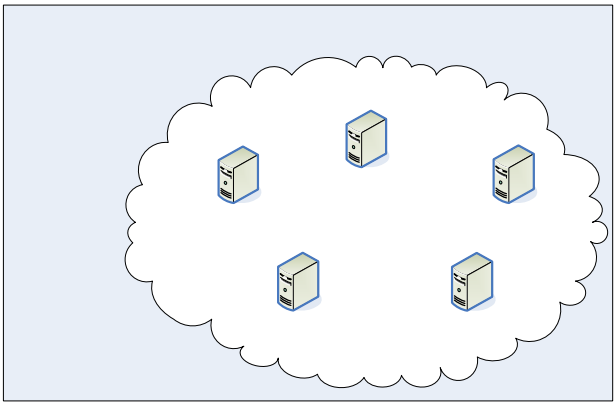

Figure 6: An Example of Formulating the Network Cost Optimization Problem, with 3 Flows Involved, for Customized Mapping Strategy 1

The objective function can be written as:

$$Minimize \quad F_3 = 10c_{3,2} + 10c_{3,3} \tag{30}$$

Subject to:

- Demand Constraints:

$$w_{11,3,2} + w_{11,3,3} = h_{11} = 400 \tag{31}$$

$$w_{16,3,2} + w_{16,3,3} = h_{16} = 200 \tag{32}$$

- Capacity Constraints:

$$5 \cdot (w_{11,3,2} + w_{16,3,3}) \le c_{3,2} \tag{33}$$

$$5 \cdot (w_{11,3,2} + w_{16,3,2}) \le c_{3,3} \tag{34}$$

- Constraints on variables:

$$w_{11,3,2} \ge 0, w_{11,3,3} \ge 0 \tag{35}$$

$$w_{16,3,2} \ge 0, w_{16,3,3} \ge 0 \tag{36}$$

In this example, each physical node hosts only one logical server. It reduces the complexity of the problem formulation. Moreover, it carries a small number of variables so that it requires the less computation time to solve the problem, compared with the generic mapping strategy.

2.3 Customized Mapping Strategy 2

While the Customized Mapping Strategy 1 discussed in the last section fits large carriers with numerous users, the mapping strategy discussed in this section fits small carriers who try to pact different logical servers into the same physical nodes to save footprint and cost. The Customized Mapping Strategy 2 is illustrated in Figure 7. In this strategy, physical nodes are divided into different groups. One logical server can be hosted by the physical nodes located in one group only. One group of physical nodes can host one or more than one logical servers. Furthermore, we assume that a message that traverses the logical servers that belong to the same group will be processed by one physical node only in the group. This constraint can reduce the traveling time within a group. In the extreme case, if each group hosts only one logical server, this becomes the customized mapping strategy 1.

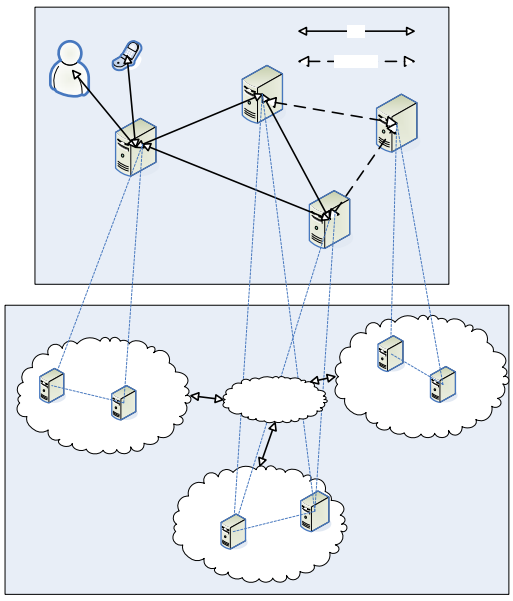

Figure 7: Customized Mapping Strategy 2

The Customized Mapping Strategy 2 allows a physical node to host two or more logical servers. For many network providers, the capacity of one physical node may be more than the loads of one logical server. This mapping strategy is more practical than the previous strategy for this type of carriers.

In general, there are two main advantages of using this mapping strategy over the first one. First, mapping more than one logical server to a physical node can utilize the existing physical node capacity, thus reduce costs. Second, this method can reduce the messages' traveling time.

The main disadvantage of this mapping strategy over the previous one is the complexity involved. When a physical node hosts more than one logical server, the physical node has to handle more types of tasks which may interfere with each other. The maintenance cost will clearly be higher.

2.3.1 Formulation of Cost Optimization Problem

In this mapping strategy, the group concept is introduced. A group of physical nodes is denoted as g, where $g = 1,2,3,...,G$ and G is the total number of groups we have in the network topology. In a group, the physical nodes host the same logical servers. Every physical node y belongs to one group. Because the residual capacities of the physical nodes can only be shared among the physical nodes in the same group, the optimization can be decomposed into the optimization of each group of physical nodes. It can be formulated as the follows.

- Indices:

- f=1, 2,…, 17, flow
- v = 1, 2, 3, 4, logical server
- g=*1,2,3,…,G*, group number
- y_g=1, 2, …Y_g, physical node located in group g.

- Constants:

- α_{fg} : =1, if flow f traverses group g.
- δ_{fvg} : =1, if flow f traverses logical server v mapped to group g.
- β_{fv} : number of times that logical server v is involved in flow f.
- h_f : flow demand volume for flow f.
- κ_v : capacity coefficient in time-capacity product unit for each logical server v.
- ε_{gy} : cost coefficient per unit processing capacity for physical node y in group g.

- Variables:

- w_{fgy_g} : loads of flow f allocated to physical node y in group g.

- Objective: Minimize total physical nodes cost for group g.

$$F_g = \sum_{y_v} \varepsilon_{gy_g} c_{gy_g} \tag{37}$$

- Constraints:

- Demand Constraints:

$$\sum_{y_v} w_{fgy_g} = h_f \alpha_{fg} \tag{38}$$

- Capacity Constraints:

$$\sum_f \sum_v \kappa_v \beta_{fv} \delta_{fvg} w_{fgy_g} \le c_{gy_g} \quad (39)$$

- Constraints on variables,

$$w_{fgy_g} \ge 0 \text{ (continuous, non - negative)} \quad (40)$$

$$c_{gy_g} \ge 0 \text{ (continuous, non - negative)} \quad (41)$$

2.3.2 Example

The network topology illustrated in Figure 8 is provided to show the formulation of cost optimization problem using the Mapping Strategy 2. This network topology is the topology discussed in Section 2.1.3 with a relocation of the logical servers in the 5 physical nodes. The assumptions and given information remain the same as provided in Section 2.1.3. Clearly only Groups #1 and #2 can be optimized.

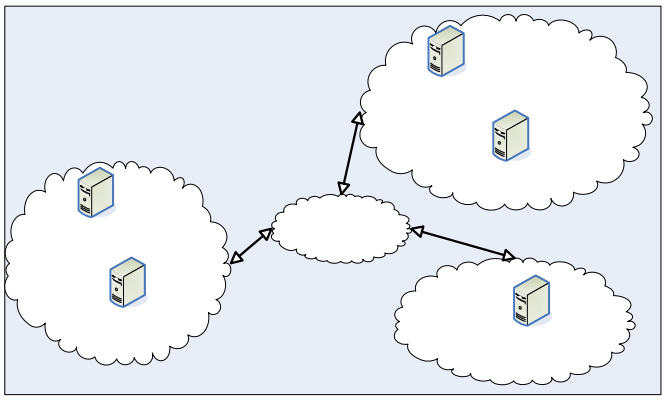

Figure 8: An Example of Formulating Network Cost Optimization problem, with Flows Involved, for Customized Mapping Strategy 2

The objective function can be written as:

$$Minimize \quad F_1 = 5c_{1,1} + 10c_{1,2} \quad (42)$$

$$Minimize \quad F_2 = 10c_{2,3} + 5c_{2,4} \quad (43)$$

Subject to

- Demand Constraints:

$$w_{11,1,1} + w_{11,1,2} = h_{11} = 400 \quad (44)$$

$$w_{11,2,3} + w_{11,2,4} = h_{11} = 400 \quad (45)$$

$$w_{15,2,3} + w_{15,2,4} = h_{15} = 250 \quad (46)$$

$$w_{16,2,3} + w_{16,2,4} = h_{16} = 200 \quad (47)$$

- Capacity Constraints:

$$2 \cdot 2w_{11,1,1} \le c_{1,1} \quad (48)$$

$$2 \cdot 2w_{11,1,2} \le c_{1,2} \quad (49)$$

$$2w_{11,2,3} + 5w_{11,2,3} + w_{15,2,3} + 5w_{16,2,3} \le c_{2,3} \quad (50)$$

$$2w_{11,2,4} + 5w_{11,2,4} + w_{15,2,4} + 5w_{16,2,4} \le c_{2,4} \quad (51)$$

- Constraints on variables:

$$w_{11,1,1} \ge 0, w_{11,1,2} \ge 0, w_{11,2,3} \ge 0, w_{11,2,4} \ge 0, w_{15,2,3} \ge 0, w_{15,2,4} \ge 0, w_{16,2,3} \ge 0, w_{16,2,4} \ge 0 \quad (52)$$

$$w_{15,2,3} \ge 0, w_{15,2,4} \ge 0 \quad (53)$$

$$w_{16,2,3} \ge 0, w_{16,2,4} \ge 0 \quad (54)$$

3. CONCLUSION

There are three potential mapping strategies introduced in this paper. One of them is the Generic Mapping Strategy that includes all the possible mapping ways. Network providers can decide the type and the number of logical servers hosted in the physical nodes according to their needs. Therefore, once the physical network topology is given by the network provider, cost optimization problem can be formulated. Since a physical node can host one or more logical servers, the complexity of the optimization formulation increases quickly due to the dramatic increase of potential candidate paths.

However, when new constraints are introduced into the formulation, the computation time can be reduced significantly. This has been shown in the two customized mapping strategies. The Customized Mapping Strategy 1 only allows a physical node host one logical server. The overall optimization problem can then be decomposed into the optimization problem for each logical server. The complexity only depends on the number of physical nodes that support the specific logical server. This mapping strategy can be applied to networks owned by large carriers. The Customized Mapping Strategy 2 is, on the other hand, designed for small carriers. In this strategy, physical nodes are divided into groups. Its complexity then depends on the number of physical nodes in a group. While the optimization complexity is reduced, it also allows multiple logical servers mapped to the same physical node. This will also make server utilization more efficient.

Last but not the least, the formulations of the mapping strategies proposed in this paper are all based on the novel flow concept we proposed in [8]. Without the flow concept, it's not possible to formulate the problems in a scalable way.

REFERENCES

[1] G. Camarillo and M. A. Garcia-Martin, "*The 3G IP Multimedia Subsystem (IMS): Merging the Internet and the Cellular worlds*," Second Edition, Wiley, 2005.

[2] J. C. Chen and T. Zhang, "*IP-based Next-Generation Wireless Network*," Wiley, 2004.

[3] G. T. "*Digital Cellular Telecommunications System (Phase 2+); Universal Mobile Telecommunications System (UMTS); IP Multimedia Subsystem (IMS); Stage 2*," version 7.5.0, ETSI TS123228, 2006.

[4] S. Pandey, V. Jain, D. Das, V. Planat and R. Periannan, "Performance Study of IMS Signaling Plane," *Proceeding of International Conference on IP Multimedia Subsystem Architecture and Applications*, pp.1-5, December 2007.

[5] V. Koukoulidis and M. Shah, "The IP Multimedia Domain in Wireless Networks: Concepts, Architecture, Protocols and Applications," *Proceeding of IEEE 6th International Symposium on Multimedia Software Engineering*, pp. 484-490, Miami, December 2004.

[6] M. Handley, H. Schulzrinne, E. Schooler and J. D. Rosenberg, "*SIP: Session Initiation Protocol*," IETF RFC 2543, March 1999.

[7] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnson, J. Peterson, R. Spar M. Handley and E. Schooler, "*The session initiation protocol (SIP)*," IETF RFC 3261, 2001.

[8] J. Xiao, C. Huang and J. Yan, "A Flow-based Traffic Model for SIP Messages in IMS," *IEEE GLOBECOM*, Hawaii USA, November 2009.

[9] A. A Kist, E. A. Kist and R. J. Harris, "SIP Signaling Delay in 3GPP," *Proceeding of the 6th International Symposium on Communications Interworking of IFIP interworking 2002*, pp. 13-16, Fremantle, October 2002.

[10] J. Hwang, N. Kim, S. Kang and J. Koh, "A Framework for IMS Interworking Networks with Quality of Service Guarantee," *Proceeding of the 7th International Conference on Networking*, pp. 454-459, Cancun, April 2008.

Physical Topology
Flow 11, f=11