The draft for the cost minimization of a federated cloud participant

***Abstract*—We plan to consider the assignment of virtual machines under the dynamic request arrival scenario. The objective is to minimize the total cost.**

I. System Model

We study the problem that how an individual cloud provider in the federated clouds minimize its cost.

A. Cloud federation and job model

We consider a cloud federation with J cloud providers in total. N^j is the number of servers in cloud $j, 1 \leq j \leq J$. Each cloud in the federation provides M types of instances(i.e., virtual machines). Each server in cloud j can host H_m^j type m instances, $1 \leq m \leq M$. The price for type m instance of cloud j is p_m^j per unit time. The price for cloud j transferring data out or in is p_{do}^j or p_{di}^j per unit volume.

We consider a time slotted system. Let $t = 0, 1, 2, 3..., T$ be the time slots. We model users' requests for a bundle of instances for a fixed time size as jobs. One job type is denoted by a $(m+1)$-tuple $(s_1^k, ..., s_M^k, w^k)$. $s_m^k, 1 \leq m \leq M$ is the required number of type m instances by job type k. w^k is the required service time of job type k. Let $\mathcal{K}$ be the set of all job types. Let $N_m^j(t)$ be the number of servers configured to provide type m instances at time t. Each server in cloud j can host n_m^j type m instances.

Let $\mathcal{A}_k^j(t)$ be the set of type k jobs arriving at time slot t in cloud j, $|\mathcal{A}_k^j(t)| = A_k^j(t)$. Let $\mu_k^j(t)$ denote the number of type k jobs at cloud j newly served at time slot t. Then, the number of type k jobs at cloud j newly served at time slot $t - w, 0 \leq w \leq w^k - 1$ is $\mu_k^j(t-w), 0 \leq w \leq w^k - 1$.

We use a queue Q_k^j to denote the workload of job type k at cloud j. The dynamic of Q_k^j is as follows:

$$Q_k^j(t+1) = \max\{Q_k^j(t) - \sum_{w=0}^{w^k-1} \mu_k^j(t-w), 0\} + w^k \cdot A_k^j(t), \\ 1 \leq j \leq J, 1 \leq k \leq K. \tag{1}$$

Hence, there are $\mu_k^j(t-w)$ type k jobs running in the federation newly served at time slot $t - w$. Indexing the $\mu_k^j(t-w)$ jobs from 1 to $\mu_k^j(t-w)$. Let $\mathcal{L}$ be the set of all jobs being served in the federation at one time slot. A job l in $\mathcal{L}$ can be denoted by a 4-tuple (j, k, t, h), which means the job is the h-th type k job from cloud j that receive service at time slot t. With the denotation for a job, we use a 2-tuple to denote an instance s of job l (m, a), which means the instance is the a-th one of job l's type m instances. Let $\mathcal{S}_l$ be the set of all instances of job l.

Each job is either served or dropped (subject to a penalty) before the maximum response delay d.

Let $I_{l,s}^i$ be the indicator whether the instance $s \in \mathcal{S}_l$ in cloud i or not.

$$I_{l,s}^i = 1 \text{if instance } s \text{ is in cloud } i \\ I_{l,s}^i = 0 \text{if not.} \tag{2}$$

The number of type m instances that is purchased by cloud j from cloud i is:

$$r_m^{ji} = \sum_{l:j_l=j} \sum_{s:m_s=m} I_{l,s}^i, 1 \leq i \leq J \tag{3}$$

The total type m instances run in cloud i satisfies the following constraint:

$$\sum_{j=1}^{J} r_m^{ji} \leq N_m^i \cdot n_m, \\ 1 \leq i \leq J, 1 \leq m \leq M. \tag{4}$$

The traffic from cloud j to cloud i,i.e., tranferring into cloud i, is

$$T_{ji} = \sum_{l \in \mathcal{L}} \sum_{s_1 != s_2} I_{l,s_1}^j \cdot I_{l,s_2}^i T_{s_1,s_2} \tag{5}$$

The migration of instance s in a job l can be calculated by

$$G_{l,s} = \frac{\sum_{i=1}^{J} |[I_{l,s}^i(t) - I_{l,s}^i(t-1)]|}{2}. \tag{6}$$

If $G_{l,s} = 0$, there is no migration, else, instance s in a job l is migrated. We can use this to indicate whether there is a migration or not.

B. Problem definition

We formulate the cost minimization problem for one cloud provider in the federation.

Let us consider the cost at time slot t of cloud j in the federation.

The operational cost includes the electricity cost, is related to the number of on servers:

$$C_o = \beta^j \cdot \sum_{m=1}^{M} N_m^j$$

The network cost related to transferring data out and in:

$$C_d = p_{do}^j \cdot \sum_{i!=j} T_{ji} + p_{di}^j \cdot \sum_{i!=j} T_{ij}$$

The migration cost related to the migration times:

$$C_m = \alpha_j \cdot \sum_{l \in \mathcal{L}} \sum_{s} G_{l,s}$$

The cost of running instances in other cloud providers, which is the total cost of running instances in other providers minus the total revenue by running instances from other providers:

$$C_s = \sum_{i!=j} \sum_{m=1}^{M} p_m^i \cdot r_m^{ji} - \sum_{i!=j} \sum_{m=1}^{M} p_m^j \cdot r_m^{ij}$$

Hence, the total cost at time slot t is:

$$C = C_o + C_d + C_m + C_s$$

The time-averaged cost of cloud j for running instances is:

$$\overline{C} = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} C(t) \tag{7}$$

The cost minimization problem at cloud j can be formulated as follows:

$$\min \overline{C}$$
$$\text{constraints } 1 - 6.$$

The variables include $\mu_k^j(t)$, i.e., the number of new type k jobs served at time t, $I_{l,s}^j$, i.e., the instance s assignment indicator, N_m^j, i.e., the instance provisioning)