Federated Cloud Pricing(Draft)

When there is no federation among clouds, it is difficult to switch to other cloud service providers after it is involved in the current cloud service provider. The data is locked-in. The federation cloud makes it easily for users to switch among different cloud service providers.

The cost of a cloud provider changes with the price of power dynamically. There are papers discussing how to make use of the dynamic power price in different places to reduce the cost of clouds. The federation cloud also promotes this reduction.

The question is how the federation cloud should price its users. The pricing policy should achieve the maximum profit for the federation cloud. Meanwhile, to make an individual cloud provider willing to participate in the federation, the profit expectation should be increased after the federation.

1 Problem Formulation

We assume the federation cloud has M distributed data centers, denoted by $\mathcal{D} = D_1, \ldots, D_M$. Each data center D_i has N_i homogeneous servers.

1.1 The request model

Consider the system operates in slotted time $t \in 0, 1, 2, \ldots$.

The requests for cloud services may request different VM configurations. The requests for the same type of VM configurations are the same type of requests. We first consider one type request. Users may request the same type of VMs for different time durations. We model a request as the request for one machine one time slot. We differentiate the service for the same type of requests according to the responsive time. Let S be the number of different service level agreements. Let $\mathcal{A}_s(t)$ denote the potential requests for SLA s that arrive at the beginning of time slot t, and let $A_s(t) = |\mathcal{A}_s(t)|$. We denote the price the federation cloud offers for VMs with SLA s by $p_s, 1 \leq s \leq S$. We assume potential requests' valuation of the service satisfies the cumulative distribution function (CDF) $F_s(v)$, v is potential requests' valuation. Hence, $(1 - F_s(p_s))$ portion of potential requests will request service. The actual request arrival rates are $[1 - F_s(p_s)] \cdot A_s(t)$ and queued in a buffer s in the federation cloud for service.

1.2 Server operation model

The federation cloud can select a data center $D_i, 1 \leq i \leq M$ from M distributed data centers to serve a request. Hence, different data centers may have different service rate. Let $N_i(t)$ be the active servers at time slot t at data center i. Each server can run μ VMs. Hence, at each time slot, the federation cloud control the number of activated servers $N(t) = (N_1(t), \ldots, N_M(t))$ to adjust each data center's service rate. Each time slot, the served requests are $\eta \cdot (N_1(t), \ldots, N_M(t))$. Let $(\mu_1, \mu_2, \ldots, \mu_S$ denote the service rate for different queues.

1.3 The profit model

When data center D_i runs $N_i(t)$ servers, it consumes the power of $P_i(N_i(t))$. Let $c(t) = (c_1(t), \ldots, c_M(t))$ denote the price of power at different data centers. The total power cost at time slot t is $CP(t) =$

$\sum_{i=1}^{M} c_i(t) \cdot P_i(N_i(t))$. The revenue at time slot t is $R(t) = \sum_{s=1}^{S} \sum_{s=1}^{S} [1 - F_s(p_s)] A_s(t) \cdot p_s$. The profit at time slot t is $P(t) = R(t) - CP(t) = \sum_{s=1}^{S} \sum_{s=1}^{S} [1 - F_s(p_s)] A_s(t) \cdot p_s - \sum_{i=1}^{M} c_i(t) \cdot P_i(N_i(t))$.

1.4 The profit maximization problem

Let $Q(t) = (Q_1(t), Q_2(t), \ldots, Q_S(t))$ be the vector denoting the requests queued at the federation cloud at time slot t. The request queue for SLA level s dynamics as follows:

$$Q_s(t+1) = \max[Q_s(t) - \mu_s(t), 0] + [1 - F_s(p_s)] \cdot A_s(t)$$

The time average profit of the federation cloud is:

$$\bar{P} \triangleq \lim_{t \to \infty} \sup \frac{1}{t} \sum_{\tau=0}^{t-1} E\{P(t)\}.$$

The profit maximization problem is for the federation cloud to choose appropriate price $p_s(t)$, the number of activated servers $N_i(t)$.