Cost Efficient Scheduling for Delay-sensitive Tasks in Edge Computing System

Yongchao Zhang*, Xin Chen*, Ying Chen*, Zhuo Li*, Jiwei Huang†

问题的提出：边缘计算系统的一个关键问题是如何在完成卸载任务的同时降低系统成本。

目标：任务调度问题 ->优化问题：在满足所有任务延迟要求的同时使系统成本最小化

解决方案：

系统模型：

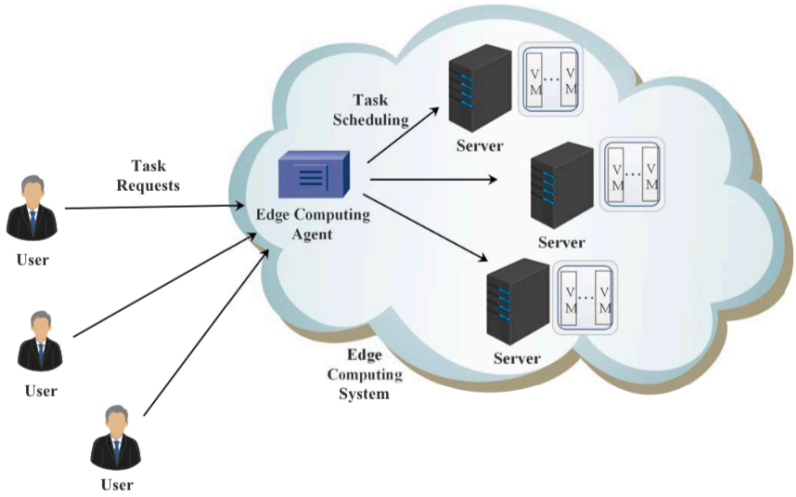

Fig. 1. Edge Computing Architecture.

边缘计算系统由边缘计算代理（ECA）和若干异构边缘服务器组成。ECA有所有可用资源的信息，并且与每个部署的服务器进行通信。每个服务器运行几个虚拟机，代表用户处理卸载的任务。通过将计算任务卸载到边缘计算系统实现更好的体验质量。ECA将根据其资源需求选择可用的服务器对从用户卸载的每个计算任务进行处理。

Problem Formulation ：

1) Task Model and Server Model:

a group of delay-sensitive tasks $T = \{t1, t2, \cdots, tn\}$

Each task $ti \in T$ is indicated by $ti = \{di, wi, \delta i, si\}$

在边缘计算系统中有m个heterogeneous edge servers 。

Each server $ej \in E$ is denoted by $ej = \{Bj,Vj,Rj,Sj\}$.

bij表示为当ti被调度到ej进行处理时的带宽需求。xij来表示任务ti是否被调度到服务器ej

$$x_{ij} = \begin{cases} 1, & \text{task } t_i \text{ is scheduled to the server } e_j \\ 0, & \text{otherwise.} \end{cases}$$

2)Optimization Problem::

在边缘计算系统中，ECA将每个卸载的任务调度到边缘服务器。当服务器被分配任务时，它将处于ON状态，否则处于OFF状态。优化问题表示为：

$$\min_{x_{ij}} \sum_{j=1}^{m} y_j C_j$$

其中Cj表示当服务器ej处于ON状态时边缘计算系统支付的成本。

$$y_j = \begin{cases} 1, & \text{server } e_j \text{ is in ON state} \\ 0, & \text{otherwise.} \end{cases}$$

约束条件：

$$\sum_{j=1}^{m} x_{ij} = 1 \quad \forall i \in 1, 2, \cdots, n \qquad (15)$$

$$\sum_{i=1}^{n} x_{ij} s_i \leq S_j \quad \forall j \in 1, 2, \cdots, m \qquad (16)$$

$$\sum_{i=1}^{n} x_{ij} \leq V_j \quad \forall j \in 1, 2, \cdots, m \qquad (17)$$

$$\sum_{i=1}^{n} x_{ij} b_{ij} \leq B_j \quad \forall j \in 1, 2, \cdots, m \qquad (18)$$

对于15 因为一个任务只能由一个服务器处理
对于16 因为调度到服务器ej的每个任务的总存储需求不能超过服务器ej的存储资源
对于17 因为调度到服务器ej的任务总数不能超过服务器ej上的VM数量
对于18 因为对于每个边缘服务器ej，调度到服务器ej的任务所需带宽的总和必须不大于服务器ej的带宽。
该优化问题为NP-hard

算法：TTSCO算法：任务调度算法，称为两阶段任务调度成本优化（TTSCO）
TTSCO算法包括两个阶段。在第一阶段，我们利用改进的BF算法得到了初步的任务调度策略。在第二阶段中，对之前的任务调度方案进行了优化，得到了最终的任务调度策略。