MOERA: Mobility-agnostic Online Resource Allocation for Edge Computing

Lin Wang, Lei Jiao, Jun Li, Julien Gedeon, and Max Mühlhäuser

一. 问题的提出：

边缘计算的一个大挑战是在用户移动性强加于高动态的情况下，边缘资源的有效分配和适应。

二. 目标：边缘云系统中运行的移动应用程序的有效的在线动态资源分配

三. 解决方案：

3.1. 模型：边缘云系统

一个随着时间而变化的h时隙系统，一个边缘云为n的边缘计算系统，用s=s1，…，s n表示，由一个城域网（WAN）互连。每个边缘云s（s ∈ S）的最大容量为Cs. 两个边缘云s1和s2之间的网络延迟表示为d(s1,s2)，边缘云覆盖一个小的地理区域，系统中的任何用户都只能从最近的边缘云接收coverage 。

3.2 问题的定义：

3.2.1 用户和工作量：

一组m个用户（以U=u1，…，um表示）在移动。在时隙t∈T，用户u∈U 连接到覆盖用户附近的边缘云S * u,t 上的接入点，并将计算任务卸载到边缘云，系统中总共会产生λu的工作量。用Xs,u,t表示在时间t时为边缘云s中的用户u分配的资源量。系统为每个用户分配的资源总量不应小于用户的工作负载，即

$$\sum_{s\in S_u} x_{s,u,t} \geq \lambda_u,$$

lu,t 表示用户U在时隙T中的位置 d(lu,t , s * u,t) 时隙T中用户U的访问延迟

3.2.2 成本：

系统的性能具有四种一般成本类型：Operation cost. 、Service quality cost 、Reconfiguration cost 和 Migration cost。前两种成本属于静态成本范畴，在每个时间段内独立产生；后两种成本属于动态成本范畴，仅对跨连续时间段的决策转换收取费用。

Operation cost.：虚拟机的使用，包括硬件资源（如CPU和内存）、硬件或软件的定期维护开销、能耗，甚至碳排放。用a s,t ， t>0表示时隙t中边缘云的“operation price”，即每单位工作量的成本。The total operation cost：

$$E_O = \sum_{t\in T}\sum_{s\in S} a_{s,t} \sum_{u\in U} x_{s,u,t},$$

Service quality cost：与用户和其工作负载之间的网络延迟成比例，对于给定的边缘云S和用户U，服务质量成本的特征是用户的访问延迟d(lu,t,s*u,t) 和 访问边缘云与承载用户U工作负载的每个边缘云之间的延迟的加权和，因此，系统中的总服务质量成本可以表示为

$$E_Q = \sum_{t\in T}\sum_{u\in U}\left(d(l_{u,t}, s^*_{u,t}) + \sum_{s\in S_u}\frac{x_{s,u,t}}{\lambda_u} d(s^*_{u,t}, s)\right)$$

Reconfiguration cost： 假设重新配置成本与增加的工作量成正比总重配置成本计算为

$$E_R = \sum_{t\in T}\sum_{s\in S} c_s \left(\sum_{u\in U} x_{s,u,t} - \sum_{u\in U} x_{s,u,t-1}\right)^+$$

Migration cost：

$$E_M = \sum_{t\in T}\sum_{s\in S} b_s^{out} w_{s,t}^{out} + b_s^{in} w_{s,t}^{in}.$$

the total cost of the system： 系统成本总和为所有成本的加权和，权重隐式的包含在每个成本中了
因此总成本为：

$$E = E_O + E_R + E_Q + E_M.$$

3.3 问题抽象化：

$$\begin{aligned}
\min \quad & P_0 = \overbrace{E_O + E_Q}^{\text{static}} + \overbrace{E_R + E_M}^{\text{dynamic}} \\
\text{s.t.} \quad & \textstyle\sum_{s\in S_u} x_{s,u,t} \geq \lambda_u,\ \forall u,\ \forall t, && \text{(7a)} \\
& \textstyle\sum_{u\in U} x_{s,u,t} \leq C_s,\ \forall s,\ \forall t, && \text{(7b)} \\
& x_{s,u,t} \geq 0,\ \forall s,\ \forall u,\ \forall t. && \text{(7c)}
\end{aligned}$$

（7a）确保为每个用户分配足够的资源
（7b）确保不违反每个边缘云的容量约束

四. 解决算法：MOERA —— Mobility-agnostic Online Edge Resource Allocation

首先，MOERA进行了一个 gap-preserving transformation ， 简化原始问题。

然后，MOERA基于在每个时间段内求解具有精心设计的对数目标的子问题，并且所有子问题的解最终将构成一个针对原始资源分配问题的可行解。

1.a gap-preserving transformation

P0中的迁移成本是双向计算的，所以对P0目标中的迁移成本进行了转换，简化处理。由此生成以下新的表达式p1：

其中：

$$\min \quad P_1 = E_O + E_R + E_Q + \textstyle\sum_{t\in T}\sum_{s\in S} b_s w_{s,t}^{in}$$

$$b_s \triangleq b_s^{out} + b_s^{in}$$

使 $\quad w_{s,t}^{in} = \textstyle\sum_{u\in U} w_{s,u,t}$

所以P1为 $\quad P_1 = E_O + E_R + E_Q + \displaystyle\sum_{t\in T}\sum_{s\in S}\sum_{u\in U} b_s w_{s,u,t}.$

2. 求解子问题， for the current time slot t.

$$\begin{aligned}
\min \quad & \overline{P_2}(t) = \textstyle\sum_{s\in S}\sum_{u\in U} a_{s,t} x_{s,u,t} \\
& + \textstyle\sum_{u\in U}\left(d(l_{u,t}, s_{u,t}^*) + \sum_{s\in S_u} \frac{x_{s,u,t}}{\lambda_u} d(s_{u,t}^*, s)\right) \\
& + \textstyle\sum_{s\in S} \frac{c_s}{\eta_s}\left((x_{s,t}+\varepsilon_1)\ln\frac{x_{s,t}+\varepsilon_1}{x_{s,t-1}^*+\varepsilon_1} - x_{s,t}\right) \\
& + \textstyle\sum_{s\in S}\sum_{u\in U} \frac{b_s}{\tau_{s,u}}\left((x_{s,u,t}+\varepsilon_2)\ln\frac{x_{s,u,t}+\varepsilon_2}{x_{s,u,t-1}^*+\varepsilon_2} - x_{s,u,t}\right) \\
\text{s.t.} \quad & \textstyle\sum_{s\in S_u} x_{s,u,t} \geq \lambda_u\ \forall u, && \text{(11a)} \\
& \textstyle\sum_{k\in S\setminus s}\sum_{u\in U} x_{k,u,t} \geq \sum_{u\in U}\lambda_u - C_s,\ \forall s, && \text{(11b)} \\
& x_{s,u,t} \geq 0,\ \forall s,\ \forall u, && \text{(11c)}
\end{aligned}$$

P2为凸函数，约束为线性的
所以在每个时隙t ∈ T中P2(t)的最优解X * s,u,t 构成了对P1的可行解。

实验：

指标

Single-objective vs. multi-objective.

Different workload distributions. —— 在不同工作负载场景下的性能

Different mobility levels . —— 移动的用户数和总用户数之间的比率测量

Synthetic mobility patterns. —— 验证算法的通用性

Algorithm parameters. —— 动态成本权重与目标中静态成本权重（表示为μ）之比的影响

Proportion of edge clouds.

Comparison with Prediction-based Approaches

Running Time