2.2 建模

2.2.1 定义变量

a. **用户集** U:
$U = \{u_1, u_2, \cdots, u_n\}$，表示所有用户，不同用户可能处于不同的地理位置，位置可用二维坐标 (x_i, y_i) 表示。

b. **服务集** S:
$S = \{s_1, s_2, \cdots, s_m\}$，表示所有可可供用户连接的边缘和云服务器的集合，包含边缘节点 S_{edge} 和云端节点 S_{cloud}，位置可用二维坐标 (x_j, y_j) 表示。

- 每个服务器上可能部署了一个或多个服务实例。如果没有用户连接到某个服务器，该服务器可以不部署任何服务实例，处于“非活跃”状态。
  服务实例的数量受到服务器资源限制的约束，例如 CPU、内存和带宽。

c. **用户优先级** Q_i:
每个用户 u_i 会根据其重要性（如付费用户、普通用户等）赋予一个优先级值 Q_i，用户优先级越高，表明该用户越重要。

d. **用户优先级权重** L_i:
优先级为 Q_i 的用户对应的权重为 L_i。 Q_i 越大，L_i 越大。
其中，

- L_i 为用户资源分配系数。（权重越高的用户在有相同数据请求的情况下，获得的服务器资源更多，即有更强的计算能力）

$$L_i = Q_i$$

e. **请求数据大小** D_i:
请求数据大小 D_i 是用户 u_i 发往服务器（云或边缘）的数据量。

f. **连接变量** x_{ij}:
用二进制变量 x_{ij} 表示用户 u_i 是否连接到边缘或云服务器 s_j 。

$$x_{ij} = \begin{cases} 1 & \text{用户} u_i \text{连接到服务器} s_j \\ 0 & \text{用户} u_i \text{没有连接到服务器} s_j \end{cases}$$

- 满足每个用户连接到唯一的服务器（约束4）

g. **计算资源分配** R_i
用 R_i 表示用户 u_i连接到服务器 s_j时，服务器需要分配的资源来处理请求。资源需求集合 R_i 可定义为:

$$R_i = \{r_i^{cpu}, r_i^{mem}, r_i^b, \cdots\}$$

其中:

- r_i^{cpu} 表示用户 u_i连接到服务器 s_j时所需的 CPU资源量, $r_i^{cpu} = f_{cpu}(D_i) \cdot L_i$;
- r_i^{mem} 表示用户 u_i连接到服务器 s_j时所需的 内存 资源量，$r_i^{mem} = f_{mem}(D_i) \cdot L_i$;
- r_i^b 表示用户 u_i连接到服务器 s_j时所需的 带宽 资源，$r_i^b = f_b(D_i) \cdot L_i$。

L_i 是用户的资源分配优先级系数，通过用户优先级调整服务器分配给用户的资源量；优先级较高的用户（具有较大的 L_i），将获得更多的资源，即会有更短的响应时间。

h. **响应时间** t_{ij}：
用 t_{ij} 表示用户 u_i 连接到服务器 s_j 的响应时间，由两部分组成，$t_{trans_{ij}}$ 和 $t_{proc_{ij}}$ 。

- **传输延迟** $t_{trans_{ij}}$：
  表示用户 u_i 到服务器 s_j 的传输延迟，根据服务器是**边缘节点**还是**云节点**有所不同：
  - **边缘节点的传输延迟** $t^e_{trans_{ij}}$：

$$t^e_{trans_{ij}} = t^e_{d_{ij}} + t^e_{b_{ij}}$$

其中：

$$t^e_{d_{ij}} = \frac{d^e_{ij}}{v_e}$$

为**物理传输延迟**。d^e_{ij} 表示用户 u_i 到边缘节点 s_j 的距离，v_e 为边缘节点的网络传播速度。

$$t^e_{b_{ij}} = \frac{D_i}{b_e(t)}$$

为**带宽延迟**。D_i 是用户请求数据的大小，$b_e(t)$ 为边缘节点可用带宽，可能随时间变化。

- **云节点的传输延迟** $t^c_{trans_{ij}}$：

$$t^c_{trans_{ij}} = t^c_{d_{ij}} + t^c_{b_{ij}}$$

其中：

$$t^c_{d_{ij}} = \frac{d^c_{ij}}{v_c}$$

为**物理传输延迟**。d^c_{ij} 表示用户 u_i 到云节点 s_j 的距离，v_c 为边缘节点的网络传播速度，因为其物理距离更远，所以通常比边缘节点传播速度低。

$$t^c_{b_{ij}} = \frac{D_i}{b_c(t)}$$

为**带宽延迟**。D_i 是用户请求数据的大小，$b_c(t)$ 为云节点可用带宽，可能随时间变化。

- **处理时间** $t_{proc_{ij}}$：
  $t_{proc_{ij}}$ 是服务器处理请求的时间，取决于**请求数据大小**和服务器的**处理速率**。处理时间计算也根据**边缘节点和云节点**的不同资源情况有所区别。

- **边缘节点的处理时间** $t^e_{proc_{ij}}$：

$$t^e_{proc_{ij}} = \frac{D_i}{P^e_{ij}}$$

其中，P^e_{ij} 表示边缘节点 s_j 对用户 u_i 的**数据请求的处理能力**，与边缘节点 s_j 分配给用户 u_i 的资源量有关：

- **云节点的处理时间** $t^c_{proc_{ij}}$：

$$t^c_{proc_{ij}} = \frac{D_i}{P^c_{ij}}$$

其中，P^c_{ij} 表示云节点 s_j 对用户 u_i 的**数据请求的处理能力**，与云节点 s_j 分配给用户 u_i 的资源量有关。

- 注：P^c_j 和 P^e_j 分别为边缘节点和云节点的**全部计算能力**，由于云节点的资源比边缘节点丰富，则应满足：

$$P^c_j > P^e_j$$

- **综合响应时间计算**：
  用户 u_i 连接到服务器 s_j 的响应时间 t_{ij} 为：
  - **若** s_j **位于边缘节点**：

$$t_{ij} = t^e_{trans_{ij}} + t^e_{proc_{ij}} = \left(t^e_{d_{ij}} + t^e_{b_{ij}}\right) + t^e_{proc_{ij}} = \left(\frac{d^e_{ij}}{v_e} + \frac{D_i}{b_e(t)}\right) + \frac{D_i}{P^e_{ij}}$$

  - **若** s_j **位于云节点**：

$$t_{ij} = t^c_{trans_{ij}} + t^c_{proc_{ij}} = \left(t^c_{d_{ij}} + t^c_{b_{ij}}\right) + t^c_{proc_{ij}} = \left(\frac{d^c_{ij}}{v_c} + \frac{D_i}{b_c(t)}\right) + \frac{D_i}{P^c_{ij}}$$

- **平均响应时间计算**：

$$\bar{T} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} \cdot \left(\begin{cases} \left(\frac{d^e_{ij}}{v_e} + \frac{D_i}{b_e(t)}\right) + \frac{D_i}{P^e_{ij}}, & s_j \in S_{edge} \\ \left(\frac{d^c_{ij}}{v_c} + \frac{D_i}{b_c(t)}\right) + \frac{D_i}{P^c_{ij}}, & s_j \in S_{cloud} \end{cases}\right)$$

i. **部署实例成本** c_j：
用 c_j 表示在服务器 s_j 部署服务的成本，由于云节点和边缘节点的差异性，二者的部署成本也有所不同。

- **边缘节点部署成本** c^e_j：

$$c^e_j = c^e_{fixed_j} + c^e_{usage_j}$$

*

- **固定成本** $c^e_{fixed_j}$：用于租赁边缘服务器（每个边缘节点的固定租赁费用）；
- **资源使用成本** $c^e_{usage_j}$：是边缘节点在实际运行中产生的资源消耗成本，如 CPU、内存和带宽消耗，按资源占用进行计算：

$$c^e_{usage_j} = \sum_{i=1}^{n} x_{ij} \cdot (r^{cpu}_i \cdot p^e_{cpu} + r^{mem}_i \cdot p^e_{mem} + r^b_i \cdot p^e_b)$$

其中，p^e_{cpu}、p^e_{mem}、p^e_b 分别表示边缘节点上单位 CPU、内存和带宽资源的单价；x_{ij} 表示用户 u_i是否连接到服务器 s_j。

- **云节点部署成本** c^c_j（云节点通常基于资源使用量付费，可以根据资源消耗量和处理请求数量来计算成本）：

$$c_j^c = c_{usage_j}^c + c_{net_j}^c$$

*

- **云资源使用成本** $c_{usage_j}^c$ ： 按需计算，包括 CPU、内存和带宽使用量。

$$c_{usage_j}^c = \sum_{i=1}^{n} x_{ij} \cdot (r_i^{cpu} \cdot p_{cpu}^c + r_i^{mem} \cdot p_{mem}^c + r_i^b \cdot p_b^c)$$

其中，p_{cpu}^c 、p_{mem}^c、p_b^c 分别表示云节点上单位 CPU、内存和带宽资源的单价；x_{ij} 表示用户 u_i是否连接到服务器 s_j。

- **网络流量成本** $c_{net_j}^c$ ： 用户从不同区域访问云，产生的额外网络传输费用，根据流量数据量计算。

$$c_{net_j}^c = \sum_{i=1}^{n} x_{ij} \cdot D_i \cdot p_{net}^c$$

其中，p_{net}^c 为云平台的流量单价，D_i 为用户 u_i 的数据请求的大小。

- **综合部署成本**
  总部署成本可表示为所有边缘和云节点成本的总和：

$$C_{total} = \sum_{s_j \in S_{edge}} c_j^e + \sum_{s_j \in S_{cloud}} c_j^c$$

2.2.2 目标函数（优化公平性）

- **公平性目标函数**:

$$f = \min \left[\lambda_1 \cdot \left(\sum_i \left| \frac{T_i}{T_{i+1}} - r \right| \right) + \lambda_2 \cdot \sum_i (1 - F_{\text{Jain}_i}) \right]$$

T_i 是优先级 i 的平均响应时间，r 是相邻优先级的期望响应时间比（1.2-1.5）

其中:

- 第一项是响应时间比例偏差指数，衡量**不同优先级之间**的公平性。
- 第二项是Jain公平性指数，衡量**同一优先级内部**的公平性。
- λ_1 和 λ_2 是权重系数，调整它们可以平衡这两项指标的重要性。

通过调整这两个系数，在优化过程中控制优先级之间的公平性和同一优先级内部的公平性之间的权衡。

2.2.3 约束条件

约束1：不同优先级用户的平均响应时间约束

为每个优先级类别的用户设定一个不同的最大允许的平均响应时:

$$\frac{1}{|U_{Q_i}|} \sum_{u_j \in U_{Q_i}} \sum_{s_k \in S} x_{jk} \cdot t_{jk} \leq T_{Q_i}^{max}, \quad \forall Q_i$$

$T_{Q_i}^{max}$ 是优先级类别 Q_i 的最大响应时间上限。这个约束确保每个优先级类别的用户响应时间不会超过设定的上限。

约束2：部署成本

- 用 C_{edge} 表示边缘节点的总预算，用 C_{cloud} 表示云端的总预算。两者的总和不得超过服务提供商的整体预算 C_{max} 。

$$C_{edge} + C_{cloud} = \sum_{s_j \in S_{edge}} c_j^e + \sum_{s_j \in S_{cloud}} c_j^c \leq C_{max}$$

其中，S_{edge} 为所有边缘节点集合，S_{cloud} 为云端节点集合，c_j 表示在节点 s_j 上部署的成本。

约束3：边缘节点计算资源限制

- 每个边缘节点 s_j 的资源消耗不超过其最大可用资源。

$$\sum_{i=1}^{n} x_{ij} \cdot r_i^{cpu} \leq R_j^{cpu\_max}, \forall s_j \in S$$

$$\sum_{i=1}^{n} x_{ij} \cdot r_i^{mem} \leq R_j^{mem\_max}, \forall s_j \in S$$

$$\sum_{i=1}^{n} x_{ij} \cdot r_i^{b} \leq R_j^{b\_max}, \forall s_j \in S$$

其中，$R_j^{cpu\_max}$、$R_j^{mem\_max}$、$R_j^{b\_max}$ 分别为边缘节点 s_j 的 CPU 、内存和带宽的最大可用资源。

约束4：用户与服务器的连接

- 每个用户 u_i 必须连接到唯一一个服务器：

$$\sum_{j=1}^{m} x_{ij} = 1, \forall u_i \in U$$

2.2.4 衡量指标：

分别从**不同的维度**衡量公平性，一个是**不同优先级之间**的公平性（通过响应时间比例偏差指数），另一个是**同一优先级用户内部**的公平性（通过Jain公平性指数）。

2.2.4.1 响应时间比例偏差指数

（衡量不同优先级用户之间的响应时间比例是否符合用户设置的公平时间比例）

$$\sum_{i} \left| \frac{T_i}{T_{i+1}} - r \right|$$ （响应时间比例偏差指数）

其中，T_i 是优先级 i 的平均响应时间，r 是相邻优先级的期望响应时间比（1.2-1.5）。

- 这个指标是计算每一对相邻优先级用户响应时间比与期望响应时间比的偏差，并求取它们的绝对值之和。
- **目标**：该指标越小，说明相邻优先级用户之间的响应时间比越符合客户设定的范围（1.2-1.5倍）。如果某一对优先级之间的响应时间比偏离了期望的比例，那么这个偏差就会增加。

2.2.4.2 Jain公平性指数

(**衡量同一优先级用户之间响应时间的公平性，F_{Jain_i} 越趋近于1，说明优先级 i 的用户越公平**)

$$F_{\text{Jain}_i} = \frac{\left(\sum_{i=1}^{n_i} t_{ij}\right)^2}{n_i \cdot \sum_{i=1}^{n_i} \left(t_{ij}\right)^2}$$

- t_{ij} 是响应时间，n_i 是优先级为 i 的用户总数。

注：

a. 高优先级用户在服务器上获取更多资源

计算资源分配 R_i
用 R_i 表示用户 u_i连接到服务器 s_j时，服务器需要分配的资源来处理请求。**资源分配集合** R_i 可定义为：

$$R_i = \{r_i^{cpu}, r_i^{mem}, r_i^b, \cdots\}$$

其中：

$$\begin{cases} r_i^{cpu} \text{ 表示用户 } u_i \text{ 连接到服务器 } s_j \text{ 时服务器分配给该用户的 CPU 资源量, } r_i^{cpu} = f_{cpu}(D_i) \cdot L_i \\ r_i^{mem} \text{ 表示用户 } u_i \text{ 连接到服务器 } s_j \text{ 时服务器分配给该用户的 内存 资源量, } r_i^{mem} = f_{mem}(D_i) \cdot L_i \\ r_i^b \text{ 表示用户 } u_i \text{ 连接到服务器 } s_j \text{ 时服务器分配给该用户的 带宽 资源量, } r_i^b = f_b(D_i) \cdot L_i \end{cases}$$

- D_i 是用户 u_i 的**数据请求量**；
- $f_x(D_i)$ 为用户数据请求量 D_i 与 **CPU、内存以及带宽** 等资源分配量的转换函数；
- L_i 是用户的**资源分配优先级系数**，通过用户优先级调整服务器分配给用户的资源量；**优先级较高的用户**（具有较大的 L_i），将获得**更多**的资源，即获得的**处理能力更强**，可能有**更短**的响应时间。

b. 服务器处理速率与资源分配

用户 u_i 连接到服务器 s_j 时，服务器分配给用户的**资源量**决定了服务器对该用户请求的**处理能力**：

$$P_{ij}^x = \left(\frac{R_i}{\sum_{u_k \in U} R_k}\right) \cdot P_j^x, x_{kj} = 1$$

- P_{ij}^x：表示服务器 s_j 分配给**用户** u_i 的**计算能力**，其中 x 可以是 e 或 c，分别代表边缘服务器和云服务器。
- P_j^x：表示服务器 s_j 的**总处理能力**，对于边缘服务器是 P_j^e，对于云服务器是 P_j^c。（通常情况下，$P_j^e < P_j^c$）
- $\frac{R_i}{\sum_{u_k \in U} R_k}$：表示用户 u_i 在服务器 s_j 上所**分配的资源量** R_i 与该服务器上**所有用户** u_k 分配的**资源总和**的比例。
- 通过比例分配，每个用户获得的**计算能力**是与其**分配的资源量**成**正比**的。