APPENDIX

Lemma 1. *(Dual Feasibility Guarantee.) $D^{(IJ)}$ is the objective value achieved by a dual feasible solution.*

Proof. In this proof, the core idea is to find a dual feasible solution of problem (9) with the objective value of $D^{(IJ)}$. Suppose that $(\{\hat{\delta}_{ij}\}_{i,j}, \{\hat{\lambda}_t\}_t, \{\hat{\varphi}_i\}_i)$ is the feasible solution. We next determine the specific values of the above variables ensuring they satisfy all the constraints in problem (9). Let

$$\hat{\lambda}_t = \lambda_t^{(IJ)}, \hat{\varphi}_i = \varphi_i^{(J)}, \forall k, t, i,$$

where $\lambda_t^{(IJ)}$ and $\varphi_i^{(J)}$ are the values of dual variables λ_t and φ_i after processing all requests, respectively. For each user i's j-th request, since we select the execution decision that achieves the maximum $F(ijt)$ (defined in equality (12)), combining with λ_t (defined in equality (10)) and φ_i (defined in equality (11)) are monotonically increasing functions, i.e.,

$$\lambda_t^{(ij)} \leq \lambda_t^{(IJ)}, \varphi_i^{(j)} \leq \varphi_i^{(J)}, \forall i, t,$$

then Constraint (9a) is satisfied. Additionally, The definition of λ_t and φ_i ensures

$$\lambda_t^{(ij)} \geq 0, \varphi_i^j \geq 0, \forall t, i, j.$$

Furthermore, the quality (14) indicates $\delta_{ij} \geq 0, \forall i, j$. Thus, Constraint (9b) also holds. In summary, D^I is the objective value achieved by the designed dual feasible solution

$$\tilde{\delta}_{ij} = \delta_{ij}, \hat{\lambda}_t = \lambda_t^{(IJ)}, \hat{\varphi}_i = \varphi_i^{(J)}, \forall i, j, t.$$

□

Lemma 2. *When $\alpha = \ln\theta$, for each user i's j-request, if it is accepted and scheduled to execute at time slot t, then $\lambda_t^{(ij-1)}(z_t^{(ij)} - z_t^{(ij-1)}) \geq \frac{1}{\alpha} C(\lambda_t^{(ij)} - \lambda_t^{(ij-1)})$.*

Proof. To prove this lemma, we only need

$$\alpha \geq \frac{C(\lambda_t^{(ij)} - \lambda_t^{(ij-1)})}{\lambda_t^{(ij-1)}(z_t^{(ij)} - z_t^{(ij-1)})}. \tag{18}$$

We use $\lambda_t^{(ij-1)}$ to denote the value of the variable λ_t before processing the j-th request of the i-th user. Similarly, we also use $z_t^{(ij-1)}$ to denote the value of the variable z_t before processing the j-th request of the i-th user. To obtain the value of α, we assume $r_{ij} \ll C, \forall i, j$, then $r_{ij} = z_t^{(ij)} - z_t^{(ij-1)}$ can be expressed as dz. Similarly, $\lambda_t^{(ij)} - \lambda_t^{(ij-1)}$ can be expressed as $d\lambda_t$. Then, the inequality (18) can be rewritten as

$$\alpha \geq \frac{Cd\lambda_t}{\lambda_t dz}. \tag{19}$$

The derivative of the function λ_t with respect to the variable z is

$$d\lambda_t = \frac{1}{e}\frac{1}{C}\theta^{\frac{z}{C}} \ln\theta dz. \tag{20}$$

Next, we substitute $d\lambda_t$ and λ_t in inequality (19) with equality (20) and (10), respectively. Then we obtain

$$\alpha \geq \ln\theta. \tag{21}$$

Therefore, when $\alpha = \ln\theta$, Lemma 2 holds. □

Lemma 3. *(Relationship between Almost-Feasible problem and dual problem.) $\tilde{P}^{(IJ)} = \frac{1}{\alpha+\beta} D^{(IJ)}$.*

Proof. We first define two types of conditions for each request to assist in the proof, including Almost-Feasible Condition and Feasible Condition:

Definition 2. ***Almost-Feasible Condition***: $F(ijt^*) > 0$

Definition 3. ***Feasible Condition***: $(F(ijt^*) > 0) \wedge (R(ijt) + r_{ij} \leq C, \forall t) \wedge (B(ij) + e_p \leq B_i, \forall i)$

In Algorithm 1, line 5 checks Almost-Feasible Condition, while line 5 and line 7 together check Feasible Condition. We refer to the solution generated by the Almost-Feasible Condition as the Almost-Feasible Primal Solution. Likewise, the solution generated by the Feasible Condition is called the Feasible Primal Solution. An Almost-Feasible Primal Solution can be easily transformed into a Feasible Primal Solution by simply not executing requests that satisfy line 5 but not line 7 in Algorithm 1. Let $\tilde{P}^{(ij)}$ denote the objective function value of problem (8) achieved by the Almost-Feasible Primal Solution after processing user i's j-th request. Denote $D^{(ij)}$ as the objective function value of dual problem (9) after processing user i's j-th request.

In this lemma, we will specify the relationship between $\tilde{P}^{(IJ)}$ and $D^{(IJ)}$. Let S_a be the set of requests that satisfy the Almost-Feasible Condition, i.e., $F(ijt^*) > 0, \forall i \in S_a$. Let S_r be the set of requests that the Almost-Feasible Condition directly rejects. For request $(i, j) \in S_r$, $\tilde{P}^{(ij)} - \tilde{P}^{(ij-1)} = 0$, $D^{(ij)} - D^{(ij-1)} = 0$. For request $i \in S_a$, we have

$$\tilde{P}^{(ij)} - \tilde{P}^{(ij-1)} = u_{ij}.$$

We also have

$$\begin{aligned} D_1^{(ij)} - D_1^{(ij-1)} = \delta_{ij} + C(\lambda_{t^*}^{(ij)} - \lambda_{t^*}^{(ij-1)}) \\ + B_i(\varphi_i^{(ij)} - \varphi_i^{(ij-1)}). \end{aligned} \tag{22}$$

From (11) we obtain

$$\varphi_i^{(ij)} - \varphi_i^{(ij-1)} = \varphi_i^{(ij-1)} \frac{e_p}{B_i} + \beta \frac{u_{ij}}{B_i}. \tag{23}$$

We next use the above result to substitute $\varphi_i^{(ij)} - \varphi_i^{(ij-1)}$ in (22), and obtain

$$\begin{aligned} D_1^{(ij)} - D_1^{(ij-1)} = \delta_{ij} + C(\lambda_{t^*}^{(ij)} - \lambda_{t^*}^{(ij-1)}) \\ + e_p\varphi_i^{(ij-1)} + \beta u_{ij}. \end{aligned} \tag{24}$$

From Lemma 2 we have

$$\lambda_t^{(ij-1)}(z_t^{(ij)} - z_t^{(ij-1)}) \geq \frac{1}{\alpha} C(\lambda_t^{(ij)} - \lambda_t^{(ij-1)}). \tag{25}$$

Then we have

$$\begin{aligned} D_1^{(ij)} - D_1^{(ij-1)} &\leq \delta_{ij} + \alpha(z_t^{(ij)} - z_t^{(ij-1)})\lambda_{t^*}^{(ij-1)} \\ &\quad + e_p\varphi_i^{(ij-1)} + \beta u_{ij}. \\ &\leq \alpha(\delta_{ij} + e_p\varphi_i^{(ij-1)} \end{aligned} \tag{26}$$

$$+ (z_t^{(ij)} - z_t^{(ij-1)})\lambda_{t^*}^{(ij-1)}) + \beta u_{ij} \quad (27)$$

$$\leq \alpha(\delta_{ij} + e_p\varphi_i^{(ij-1)} + r_{ij}\lambda_{t^*}^{(ij-1)}) + \beta u_{ij}$$

$$= \alpha u_{ij} + \beta u_{ij} \quad (28)$$

$$= (\alpha+\beta)(\tilde{P}^{(ij)} - \tilde{P}^{(ij-1)}). \quad (29)$$

Therefore, $\tilde{P}^{(IJ)} = \frac{1}{\alpha+\beta}D^{(IJ)}$ and hence Lemma 3 holds. □

Lemma 4. *When $\theta \geq e\max_{i,j}\{\frac{u_{ij}}{r_{ij}}\}$, if user i's j-th request makes the allocated resource exceed the capacity C at time slot t, then no future request would be scheduled to execute at time slot t.*

Proof. If user i's j-th request makes the allocated resource exceed the capacity C at time slot t, that is,

$$\sum_{(i,j)\in N_t} z_{ij} \geq C, \quad (30)$$

where N_t is the set of requests admitted before user i's j-th request. Then, from the definition of λ_t in equality (10) we obtain

$$\lambda_t^{(ij)} \geq \frac{1}{e}\theta \geq \max_{i,j}\{\frac{u_{ij}}{r_{ij}}\}. \quad (31)$$

Combining it with the definition of $F(ijt)$ in (12), for any future user $\hat{i}$'s $\hat{j}$-th request, we have $F(\hat{i}\hat{j}t) \leq 0$, thus no future request would be scheduled to execute at time slot t. □

Lemma 5. *Assume $\frac{u_{ij}}{e_p} \geq 1$ and $\beta \geq \max_{i,j}\{\frac{u_{ij}}{e_p}\}$, if user i's j-th request makes the total expenses exceed the budget B_i, then no future request of user i would be accepted to execute.*

Proof. As we can scale the units of r_{ij} and e_p, then easily satisfy the assumption $\frac{u_{ij}}{e_p} \geq 1$, i.e., $u_{ij} \geq e_p$.

From (11) we have

$$\varphi_i^{(j)} + \beta = \varphi_i^{(j-1)}(1 + \frac{e_p}{B_i}) + \beta(\frac{u_{ij}}{B_i}) + \beta \quad (32)$$

$$= \varphi_i^{(j-1)}(1 + \frac{e_p}{B_i}) + \beta(1 + \frac{u_{ij}}{B_i}) \quad (33)$$

$$\geq (\varphi_i^{(ij-1)} + \beta)(1 + \frac{e_p}{B_i}), \quad (34)$$

where the last inequality is due to $u_{ij} \geq e_p$. Next, since

$$1 + x \geq 2^x, \forall x \in [0,1], \quad (35)$$

and $0 \leq \frac{e_p}{B_i} \leq 1$, then the inequality (34) can be rewritten as

$$\varphi_i^{(j)} + \beta \geq (\varphi_i^{(j-1)} + \beta)2^{\frac{e_p}{B_i}}. \quad (36)$$

We recursively apply the above inequality until reaching $\varphi_i^{(0)}$ and obtain

$$\varphi_i^{(j)} + \beta \geq (\varphi_i^{(0)} + \beta)2^{\frac{\sum_{j\in W_i} e_p}{B_i}} = \beta \cdot 2^{\frac{\sum_{j\in W_i} e_p}{B_i}}, \quad (37)$$

where W_t is the set of user i's requests processed before user i's j-th request. The equality (37) holds because $\varphi_i^{(0)} = 0$. If user i's j-th request makes the consumed expenses exceed the pre-defined budget B_i, *i.e.*,

$$\sum_{j\in W_i} e_p \geq B_i, \quad (38)$$

then from equality (37) we have

$$\varphi_i^{(j)} + \beta \geq \beta \cdot 2^{\frac{\sum_{j\in W_i} e_p}{B_i}} \geq 2\beta. \quad (39)$$

That is,

$$\varphi_i^{(j)} \geq \beta. \quad (40)$$

Therefore, for any future user $\hat{i}$'s $\hat{j}$-th request, we have $F(\hat{i}\hat{j}t^*) \leq 0$ due to $\beta = \max_{i,j}\{\frac{u_{ij}}{e_p}\}$, thus any future requests submitted by user i would be rejected.

□

Lemma 6. *(Relationship between primal problem and Almost-Feasible problem.) $P^{(IJ)} \geq \frac{1}{\epsilon}\tilde{P}^{(IJ)}$, where $\epsilon = 1 + \frac{2u_{max}r_{max}}{u_{min}r_{min}}$.*

Proof. Recall that $\tilde{P}^{(IJ)}$ and $P^{(IJ)}$ represent the objective value of almost-feasible problem and primal problem, respectively. We have

$$\frac{\tilde{P}^{(IJ)}}{P^{(IJ)}} = \frac{P^{(IJ)} + \sum_t\sum_{(i,j)\in N_t} u_{ij} + \sum_i\sum_{(j\in Q_i)} u_{ij}}{P^{(IJ)}}, \quad (41)$$

where N_t is the set of requests that meet the almost-feasible condition while not satisfying the capacity constraint (8b). We use $(i,j) \in N_t$ to represent that user i's j-th request is an element of set N_t. According to the Lemma 4, there is at most one request in N_t. For simplicity, we denote the quality value of this request as $u_{ij}^{(t)}$. Let Q_i be the set of requests of user i that results in violating the budget constraint (8c). We use $j \in Q_i$ to indicate that user i's j-th request incurs a violation of the budget constraint (8c). According to the Lemma 5, for each user i, there is at most one request in Q_i. Similarly, we denote the quality value of this request as $u_{ij}^{(i)}$ for simplicity. Then, we can rewrite the equality (41) as

$$\frac{\tilde{P}^{(IJ)}}{P^{(IJ)}} = 1 + \frac{\sum_t u_{ij}^{(t)}}{P^{(IJ)}} + \frac{\sum_i u_{ij}^{(i)}}{P^{(IJ)}}. \quad (42)$$

Next we bound $\frac{\sum_t u_{ij}^{(t)}}{P^{(IJ)}}$ and $\frac{\sum_i u_{ij}^{(i)}}{P^{(IJ)}}$ separately. We have

$$\frac{\sum_t u_{ij}^{(t)}}{P^{(IJ)}} = \frac{\sum_t u_{ij}^{(t)}}{\sum_t\sum_{(i,j)\in M_t} u_{ij}} \leq \max_t\{\frac{u_{ij}^{(t)}}{\sum_{(i,j)\in M_t} u_{ij}}\} \quad (43)$$

$$= \max_t\{\frac{\bar{u}_{ij}^{(t)} r_{ij}^{(t)}}{\sum_{(i,j)\in M_t} \bar{u}_{ij} r_{ij}}\} \leq \max_t\{\frac{u_{max}r_{max}}{u_{min}r_{min}}\} \quad (44)$$

$$= \frac{u_{max}r_{max}}{u_{min}r_{min}}. \quad (45)$$

The set M_t in equality (43) represents the set of all admitted requests across the entire time span. In equality (44), $\bar{u}_{ij}^{(t)} = \frac{u_{ij}^{(t)}}{r_{ij}^{(t)}}$ and $\bar{u}_{ij} = \frac{u_{ij}}{r_{ij}}$. For inequality (44), we respectively denote u_{max} and u_{min} as the maximum and minimum quality value among all requests, which are calculated by $u_{max} = \max_{i,j}\{u_{ij}\}, u_{min} = \min_{i,j}\{u_{ij}\}$. We also denote r_{max} and r_{min} as the maximum and minimum required

computation among all requests, and they are calculated by $r_{max} = \max_{i,j}\{r_{ij}\}, r_{min} = \min_{i,j}\{r_{ij}\}$, respectively.

We also have

$$\frac{\sum_i u_{ij}^{(i)}}{P^{(IJ)}} = \frac{\sum_i u_{ij}^{(i)}}{\sum_i \sum_{j\in D_i} u_{ij}} \leq \max_i \{\frac{u_{ij}^{(i)}}{\sum_{j\in D_i} u_{ij}}\} \tag{46}$$

$$= \max_i \{\frac{\bar{u}_{ij}^{(i)} r_{ij}^{(i)}}{\sum_{j\in D_i} \bar{u}_{ij} r_{ij}}\} \leq \max_i \{\frac{u_{max} r_{max}}{u_{min} r_{min}}\} \tag{47}$$

$$= \frac{u_{max} r_{max}}{u_{min} r_{min}}. \tag{48}$$

In equality (46), we use $j \in D_i$ to denote that the user i's j-th request is admitted and executed by the system.

Therefore, combining the quality (42) and the results in equality (44) and (47), we thus obtain

$$\frac{\tilde{P}^{(IJ)}}{P^{(IJ)}} \leq 1 + \frac{2u_{max} r_{max}}{u_{min} r_{min}}. \tag{49}$$

Therefore, Lemma 6 holds. □