APPENDIX
PERFORMANCE ANALYSIS

RECaching is enhanced by three mechanisms that perform four kinds of decisions for edge caching. We perform competitive analysis for each decision separately [48]. In order to analyze the performance throughout the lifespan of a data object, we divide its lifespan into phases according the caching decisions of RECaching enhanced by the mechanisms. Then the performance of RECaching for each phase is analyzed. For the caching decision given by Algorithm 2, the analysis is detailed in Proposition 1.

Proposition 1. *The caching decision of Algorithm 2 is* $(1+\max\{\epsilon,\beta\})$*-competitive, where* $\epsilon,\beta\in[0,1]$.

Proof. For RECaching, if it observes g_t^e requests and performs an action $a_t=1$, then an object is decided to be cached at the edge for the next T_t^e time slots. In order to calculate the cost incurred by this action, or in other words, the cost of the next T_t^e time slots, we first denote the number of requests during T_t^e time slots as $g^e_{[T_t^e]}$. According to cost definitions of storage, bandwidth and replica creation detailed in Section III, the caching cost for this decision is

$$\begin{aligned} C^{REC}_{[T_t^e]}(O) \leq & C^e_{[T_t^e]}(O) = P_s^c O T_t^e + P_s^e O T_t^e + P_b^e O g^e_{[T_t^e]} \\ & + \min_{e'\in\mathcal{N}_t^e}\left\{C_t^{c\to e}(O), C_t^{e'\to e}(O)\right\}, \end{aligned} \tag{51}$$

where $C^e_{[T_t^e]}(O)$ is calculated as Eqn. (24). We use '$\leq$' in the calculation of $C^{REC}_{[T_t^e]}(O)$ because the actual lifespan of caching the object at the edge is less than or equal to T_t^e due to the proposed lazy caching mechanism (i.e., Algorithm 5).

For the offline optimal decision, we discuss its cost by case based on the comparison between g_t^e that is used to calculate the caching lifespan T_t^e and $g^e_{[T_t^e]}$ that is the actual number of requests during the caching lifespan. For the case of $g^e_{[T_t^e]}\geq g_t^e$, we have $C^c_{[T_t^e]}(O)\geq C^e_{[T_t^e]}(O)$ for $g^e_{[T_t^e]}$ requests, which means that the offline optimal decision is to cache the object. The cost of the offline optimal decision is

$$\begin{aligned} C^{OPT}_{[T_t^e]}(O) \geq & P_s^c O T_t^e + P_s^e O + P_b^e O g^e_{[T_t^e]} \\ & + \min_{e'\in\mathcal{N}_t^e}\left\{C_t^{c\to e}(O), C_t^{e'\to e}(O)\right\}, \end{aligned} \tag{52}$$

We use $\geq$ in the calculation of $C^{OPT}_{[T_t^e]}(O)$ because the minimum cost for T_t^e time slots occurs when all requests are served at the edge in one time slot. Thus, the competitive ratio in this case is

$$\frac{C^{REC}_{[T_t^e]}(O)}{C^{OPT}_{[T_t^e]}(O)} \leq \frac{P_s^c O T_t^e + P_s^e O T_t^e + P_b^e O g^e_{[T_t^e]} + \dots}{P_s^c O T_t^e + P_s^e O + P_b^e O g^e_{[T_t^e]} + \dots} \tag{53a}$$

$$\leq \frac{P_s^c T_t^e + P_s^e T_t^e + P_b^e g^e_{[T_t^e]}}{P_s^c T_t^e + P_s^e + P_b^e g^e_{[T_t^e]}} \tag{53b}$$

$$\leq \frac{P_s^c T_t^e + P_s^e T_t^e}{P_s^c T_t^e} \tag{53c}$$

$$= \frac{P_s^c + \epsilon P_s^c}{P_s^c} = 1+\epsilon. \tag{53d}$$

In Eqn. (53a), we use '$\dots$' in the numerator and denominator to indicate the replica creation cost. Eqn. (53b) holds by omitting the replica creation cost in both numerator and denominator with the fact $C^{REC}_{[T_t^e]}(O)\geq C^{OPT}_{[T_t^e]}(O)$. Eqn. (53c) holds due to $P_s^e + P_b^e g^e_{[T_t^e]} > P_b^e g^e_{[T_t^e]}$. We finally obtain Eqn .(53d) by substituting Eqns. (49) for P_s^e.

In the other case of $g^e_{[T_t^e]} < g_t^e$, we have $C^c_{[T_t^e]}(O) < C^e_{[T_t^e]}(O)$ for $g^e_{[T_t^e]}$ requests. Thus, the offline optimal decision is not to cache the object at the edge, which incurs a cost of storing the object and serving requests in the cloud only:

$$C^{OPT}_{[T_t^e]}(O) = C^c_{[T_t^e]}(O) \tag{54}$$

where $C^c_{[T_t^e]}(O)$ is calculated as Eqn. (23). For RECaching, we first discuss the competitive ratio when $g^e_{[T_t^e]}=0$. According to the proposed lazy caching mechanism, an object is cached at the edge when it receives an additional request after the caching decision. Because the number of requests $g^e_{[T_t^e]}$ during the decided caching lifespan is 0, the object is not cached at the edge in fact. Thus, in this case, the cost of RECaching only includes the cost of storage for T_t^e time slots in the cloud, i.e.,

$$C^{REC}_{[T_t^e]}(O) = P_s^c O T_t^e. \tag{55}$$

Obviously, the cost of the offline optimal decision is

$$C^{OPT}_{[T_t^e]}(O) = P_s^c O T_t^e, \tag{56}$$

because there is no request, there is no QoS penalty. Thus, the competitive ratio of $g^e_{[T_t^e]}=0$ is

$$\frac{C^{REC}_{[T_t^e]}(O)}{C^{OPT}_{[T_t^e]}(O)} = 1. \tag{57}$$

Then we analyze the competitive ratio when $0<g^e_{[T_t^e]}<g_t^e$. In this case, the offline optimal decision is not to cache the object at the edge, which incurs a cost of

$$\begin{aligned} C^{OPT}_{[T_t^e]}(O) = & C^c_{[T_t^e]}(O) = P_s^c O T_t^e + P_b^c O g^e_{[T_t^e]} \\ & + \omega^e\left|\sum_{t=t+1}^{t+T_t^e} J_t^e(C_t^e(O)-C_t^c(O))\right|. \end{aligned} \tag{58}$$

Then the performance is analyzed as

$$\frac{C^{REC}_{[T_t^e]}(O)}{C^{OPT}_{[T_t^e]}(O)} \leq \frac{P_s^c O T_t^e + P_s^e O T_t^e + P_b^e O g^e_{[T_t^e]} + \dots}{P_s^c O T_t^e + P_b^c O g^e_{[T_t^e]} + \dots} \tag{59a}$$

$$\leq \frac{P_s^c O T_t^e + P_s^e O T_t^e + P_b^e O g^e_{[T_t^e]} + \dots}{P_s^c O T_t^e + P_b^c O g^e_{[T_t^e]}} \tag{59b}$$

$$= 1 + \frac{P_s^e T_t^e + P_b^e g^e_{[T_t^e]} - P_b^c g^e_{[T_t^e]} + \dots}{P_s^c T_t^e + P_b^c g^e_{[T_t^e]}} \tag{59c}$$

$$\leq 1 + \frac{P_s^e T_t^e + P_b^e g^e_{[T_t^e]}}{P_s^c T_t^e + P_b^c g^e_{[T_t^e]}} \tag{59d}$$

$$= 1 + \frac{\epsilon P_s^c T_t^e + \beta P_b^c g^e_{[T_t^e]}}{P_s^c T_t^e + P_b^c g^e_{[T_t^e]}} \tag{59e}$$

$$\leq 1 + \frac{\max\{\epsilon,\beta\}\left(P_s^c T_t^e + P_b^c g^e_{[T_t^e]}\right)}{P_s^c T_t^e + P_b^c g^e_{[T_t^e]}} \tag{59f}$$

$$= 1+\max\{\epsilon,\beta\}. \tag{59g}$$

In Eqn. (59a), we use '$\dots$' in the numerator and denominator to indicate the replica creation cost and the QoS penalty cost, respectively. Eqn. (59b) holds because the QoS penalty cost is always greater than or equal to 0, which means that the fraction will not be smaller when ignoring the QoS penalty cost from the denominator. In Eqn. (59c), we isolate a constant and divide the numerator and denominator by O. Eqn. (59d) holds with the fact as follows

$$-P_b^c g_{[T_t^e]}^e + \min_{e' \in \mathcal{N}_t^e} \{P_b^c, P_b^{e'} u(I_t^{e'})\} \leq 0.$$

In Eqn. (59e), we use P_s^c and P_b^c to represent P_s^e and P_b^e with Eqns. (49) and (50), respectively.

To sum up, the competitive ratio for the caching decision of Algorithm 2 is $1 + \max\{\epsilon, \beta\}$ by comparing Eqns. (53), (57) and (59). □

For the caching decision given by the compensation strategy for RECaching, i.e., Algorithm 3, we discuss its competitive analysis in Proposition 2.

Proposition 2. *The caching decision of Algorithm 3 is* $\max\{1+\epsilon, 2-\beta\}$*-competitive.*

Proof. Algorithm 3 is triggered to cache an object for $T_t^{e_{com}}$ time slots when the thresholds h_T and h_g are reached in time slot t. This proof starts by counting the number of requests from time slot t. We use t' to denote the time slot when the number of requests is equal to h_g, $t' > t$. The cost performance of the caching decision can be analyzed by comparing t' and $t + T_t^{e_{com}}$.

If $t' \leq t + T_t^{e_{com}}$, which means that the caching decision is cost-effective because the number of requests during the caching lifespan $T_t^{e_{com}}$ is greater than or equal to the threshold h_g triggering the caching decision, the cost of caching the object at the edge for $T_t^{e_{com}}$ time slots is

$$\begin{aligned} C_{[T_t^{e_{com}}]}^{REC}(O) \leq C_{[T_t^{e_{com}}]}^e(O) = P_s^c O T_t^{e_{com}} + P_s^e O T_t^{e_{com}} \\ + P_b^e O g_{[T_t^{e_{com}}]}^e + \min_{e' \in \mathcal{N}_t^e} \left\{ C_t^{c \to e}(O), C_t^{e' \to e}(O) \right\}. \end{aligned} \tag{60}$$

We use '$\leq$' in the above equation because the actual lifespan of caching the object at the edge is less than or equal to $T_t^{e_{com}}$ due to the proposed lazy caching mechanism (i.e., Algorithm 5). Similar to Eqn. (52), the cost of optimal decision is

$$\begin{aligned} C_{[T_t^{e_{com}}]}^{OPT}(O) \geq & P_s^c O T_t^{e_{com}} + P_s^e O + P_b^e O g_{[T_t^{e_{com}}]}^e \\ & + \min_{e' \in \mathcal{N}_t^e} \left\{ C_t^{c \to e}(O), C_t^{e' \to e}(O) \right\}, \end{aligned} \tag{61}$$

Thus, similar to Eqn. (53), the competitive ratio in this case is

$$\frac{C_{[T_t^{e_{com}}]}^{REC}(O)}{C_{[T_t^{e_{com}}]}^{OPT}(O)} \leq 1 + \epsilon. \tag{62}$$

If $t' > t + T_t^{e_{com}}$, which means that the caching lifespan is not cost-effective because the number of requests during the caching lifespan is less than h_g, the cost between time slots t and t' is

$$\begin{aligned} C_{[t'-t]}^{REC}(O) = & C_{[T_t^{e_{com}}]}^e(O) + C_{[t'-t-T_t^{e_{com}}]}^c(O) \\ \leq & (P_s^c + P_s^e) O T_t^{e_{com}} + P_s^c O (t' - t - T_t^{e_{com}}) \\ & + P_b^c O h_g + \min_{e' \in \mathcal{N}_t^e} \left\{ C_t^{c \to e}(O), C_t^{e' \to e}(O) \right\} \\ & + \omega^e \left| \sum_{t=t+1}^{t'} J_t^e (C_t^e(O) - C_t^c(O)) \right|. \end{aligned} \tag{63}$$

We use '$\leq$' in the above equation because we do not know which of h_g requests are served by the cached replica at the edge and assume that h_g requests are all served by the cloud. The offline optimal decision is not to cache the object at the edge that will incur additional unnecessary cost. Its cost is calculated as

$$\begin{aligned} C_{[t'-t]}^{OPT}(O) = C^c{}_{[t'-t]}(O) = P_s^c O (t' - t) + P_b^c O h_g \\ + \omega^e \left| \sum_{t=t+1}^{t'} J_t^e (C_t^e(O) - C_t^c(O)) \right|. \end{aligned} \tag{64}$$

Thus, the competitive ratio in this case can be obtained by simplifying Eqn. (65). By omitting the QoS penalty item in both numerator and denominator, dividing the numerator and denominator by O and re-arranging items in the numerator, we obtain Eqn. (66a) as follows

$$\begin{aligned} \frac{C_{[t'-t]}^{REC}(O)}{C_{[t'-t]}^{OPT}(O)} & \leq \frac{P_s^e T_t^{e_{com}} + P_s^c(t'-t) + P_b^c h_g + \dots}{P_s^c(t'-t) + P_b^c h_g} && \text{(66a)} \\ & = 1 + \frac{P_s^e T_t^{e_{com}} + \min_{e' \in \mathcal{N}_t^e} \{P_b^c, P_b^{e'} u(I_t^{e'})\}}{P_s^c(t'-t) + P_b^c h_g} && \text{(66b)} \\ & = 1 + \frac{P_b^c h_g - P_b^e h_g}{P_s^c(t'-t) + P_b^c h_g} && \text{(66c)} \\ & \leq 1 + \frac{P_b^c h_g - P_b^e h_g}{P_b^c h_g} = 1 + \frac{P_b^c - \beta P_b^c}{P_b^c} && \text{(66d)} \\ & = 2 - \beta. && \text{(66e)} \end{aligned}$$

In Eqn. (66a), '...' indicates the replica creation cost. Then, we isolate a constant and replace $T_t^{e_{com}}$ with Eqn. (41) in Eqns. (66b) and (66c), respectively. Eqn. (66d) holds because we omit $P_s^c(t'-t)$ in the denominator which is greater than 0 and replace P_b^e with P_b^c using Eqn. (50).

Finally, by comparing Eqns. (62) and (66), we obtain that the competitive ratio is $\max\{1+\epsilon, 2-\beta\}$. □

Next, for the caching decision given by the prolonged caching mechanism for RECaching, i.e., Algorithm 4, the cost performance is analyzed in Proposition 3.

Proposition 3. *The caching decision of Algorithm 4 is* $(1 + \max\{\epsilon, \beta\})$*-competitive, where* $\epsilon, \beta \in [0, 1]$.

Proof. This proof is similar to that of Proposition 1, so we present a sketch here. If an object that has been cached at the edge is decided to continue to be cached for another $T_t^{e_{pro}}$

$$\frac{(P_s^c+P_s^e)OT_t^{e_{com}}+P_s^cO(t'-t-T_t^{e_{com}})+P_b^cOh_g+\min_{e'\in\mathcal{N}_t^e}\{P_b^cO,P_b^{e'}Ou(I_t^{e'})\}+\omega^e\left|\sum_{t=t+1}^{t'}J_t^e(C_t^e(O)-C_t^c(O))\right|}{P_s^cO(t'-t)+P_b^cOh_g+\omega^e\left|\sum_{t=t+1}^{t'}J_t^e(C_t^e(O)-C_t^c(O))\right|} \tag{65}$$

time slots by Algorithm 4 when it receives g_t requests in time slot t, the incurred cost for $T_t^{e_{pro}}$ time slots is

$$C_{[T_t^{e_{pro}}]}^{REC}(O)=P_s^cOT_t^{e_{pro}}+P_s^eOT_t^{e_{pro}}+P_b^eOg_{[T_t^{e_{pro}}]}^e, \tag{67}$$

where $g_{[T_t^{e_{pro}}]}^e$ denotes the number of requests during the cached $T_t^{e_{pro}}$ time slots. For the offline optimal decision, similar to the proof of Proposition 1, we have two cases. If $g_{[T_t^{e_{pro}}]}^e \geq g_t$, the offline optimal decision is to cache the object. Otherwise, it is not to cache the object. Therefore, its cost $C_{[T_t^{e_{pro}}]}^{OPT}(O)$ has a lower bound of

$$P_s^cOT_t^{e_{pro}}+\begin{cases} P_s^eO+P_b^eOg_{[T_t^{e_{pro}}]}^e, & g_{[T_t^{e_{pro}}]}^e \geq g_t, \\ P_b^cOg_{[T_t^{e_{pro}}]}^e, & g_{[T_t^{e_{pro}}]}^e < g_t. \end{cases} \tag{68}$$

When $g_{[T_t^{e_{pro}}]}^e \geq g_t$, the competitive ratio is calculated as

$$\frac{C_{[T_t^e]}^{REC}(O)}{C_{[T_t^e]}^{OPT}(O)} \leq \frac{P_s^cOT_t^e+P_s^eOT_t^e+P_b^eOg_{[T_t^e]}^e}{P_s^cOT_t^e+P_s^eO+P_b^eOg_{[T_t^e]}^e} \tag{69a}$$

$$\leq \frac{P_s^cT_t^e+P_s^eT_t^e}{P_s^cT_t^e} \tag{69b}$$

$$= \frac{P_s^c+\epsilon P_s^c}{P_s^c}=1+\epsilon. \tag{69c}$$

When $g_{[T_t^{e_{pro}}]}^e < g_t$, the competitive ratio is calculated as

$$\frac{C_{[T_t^{e_{pro}}]}^{REC}(O)}{C_{[T_t^{e_{pro}}]}^{OPT}(O)}=\frac{P_s^cOT_t^{e_{pro}}+P_s^eOT_t^{e_{pro}}+P_b^eOg_{[T_t^{e_{pro}}]}^e}{P_s^cOT_t^{e_{pro}}+P_b^cOg_{[T_t^{e_{pro}}]}^e} \tag{70a}$$

$$=1+\frac{P_s^eT_t^{e_{pro}}+P_b^eg_{[T_t^{e_{pro}}]}^e-P_b^cg_{[T_t^{e_{pro}}]}^e}{P_s^cT_t^{e_{pro}}+P_b^cg_{[T_t^{e_{pro}}]}^e} \tag{70b}$$

$$\leq 1+\frac{P_s^eT_t^{e_{pro}}+P_b^eg_{[T_t^{e_{pro}}]}^e}{P_s^cT_t^{e_{pro}}+P_b^cg_{[T_t^{e_{pro}}]}^e} \tag{70c}$$

$$=1+\frac{\epsilon P_s^cT_t^{e_{pro}}+\beta P_b^cg_{[T_t^{e_{pro}}]}^e}{P_s^cT_t^{e_{pro}}+P_b^cg_{[T_t^{e_{pro}}]}^e} \tag{70d}$$

$$=1+\max\{\epsilon,\beta\}. \tag{70e}$$

Eqn. (70c) holds due to $P_b^cg_{[T_t^{e_{pro}}]}^e \geq 0$. In conclusion, the competitive ratio of the caching decision of Algorithm 4 is $1+\max\{\epsilon,\beta\}$. □

Finally, Proposition 4 analyzes the performance of uncached periods of RECahcing enhanced by the mechanisms.

Proposition 4. *The uncached periods of RECaching improved by the mechanisms is $(1+\epsilon)$-competitive.*

Proof. If an object is not cached at the edge, it means that the compensation mechanism is not triggered. That is, the number of requests during the past h_T consecutive time slots without cached replicas does not reach the threshold h_g. Thus, the incurred cost of RECaching has an upper bound,

$$C_{[h_T]}^{REC}(O) \leq P_s^cOh_T+P_b^cOh_g + \omega^e\left|\sum_{h_T}J_t^e(C_t^e(O)-C_t^c(O))\right|. \tag{71}$$

For the offline optimal decision, if it has a knowledge of requests during the uncached time slots in advance, it can decide to whether to cache the object at the edge or not with the goal of cost optimization. We discuss the cost of the offline optimal decision by case. If the optimal decision is not to cache the object, the incurred cost is

$$C_{[h_T]}^{OPT}(O) < P_s^cOh_T+P_b^cOh_g + \omega^e\left|\sum_{h_T}J_t^e(C_t^e(O)-C_t^c(O))\right|. \tag{72}$$

Obviously, the competitive for the decision of uncached periods in this case is 1. If the optimal decision is to cache the object at the edge, the incurred cost has a lower bound of

$$C_{[h_T]}^{OPT}(O) \geq P_s^cOh_T+P_s^eO+P_b^eOh_g + \min_{e'\in\mathcal{N}_t^e}\{P_b^cO,P_b^{e'}Ou(I_t^{e'})\}. \tag{73}$$

In this case, the competitive ratio is calculated as

$$\frac{C_{[h_T]}^{REC}(O)}{C_{[h_T]}^{OPT}(O)} \leq \frac{P_s^ch_T+P_b^ch_g+\sum_{h_T}C_t^e(1)-\sum_{h_T}C_t^c(1)}{P_s^ch_T+P_s^e+P_b^eh_g+\min_{e'\in\mathcal{N}_t^e}\{P_b^c,P_b^{e'}u(I_t^{e'})\}} \tag{74a}$$

$$\leq \frac{P_s^ch_T+P_s^eh_T+P_b^eh_g+\min_{e'\in\mathcal{N}_t^e}\{P_b^c,P_b^{e'}u(I_t^{e'})\}}{P_s^ch_T+P_s^e+P_b^eh_g+\min_{e'\in\mathcal{N}_t^e}\{P_b^c,P_b^{e'}u(I_t^{e'})\}} \tag{74b}$$

$$\leq \frac{P_s^ch_T+P_s^eh_T+P_b^eh_g}{P_s^ch_T+P_s^e+P_b^eh_g} \tag{74c}$$

$$\leq \frac{P_s^ch_T+P_s^eh_T}{P_s^ch_T}=\frac{P_s^c+\epsilon P_s^c}{P_s^c}=1+\epsilon. \tag{74d}$$

In conclusion, the cost performance for not caching can be bounded by the competitive ratio $1+\epsilon$. □