Filling Two Needs with One Deed: Combo Pricing Plans for Computing-Intensive Multimedia Applications

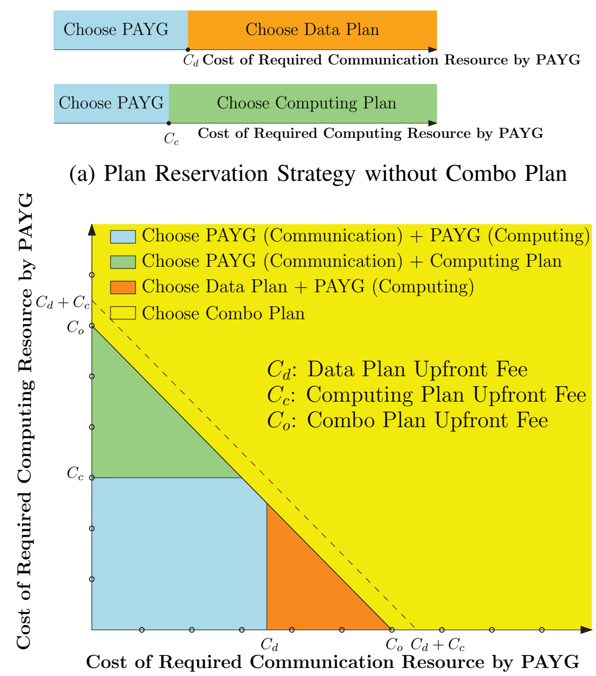
Shizhe Zang, Wei Bao, *Member, IEEE,* Phee Lep Yeoh, *Member, IEEE,* Branka Vucetic, *Fellow, IEEE,* and Yonghui Li, *Fellow, IEEE*

***Abstract*—In this paper, we examine new plan pricing schemes for multimedia applications that offload computing-intensive tasks to computing servers incurring both communication and computing costs. Pricing schemes offered to users include: 1) a pay-as-you-go payment for usage of communication or computing resources, 2) an upfront data (resp. computing) plan for unlimited usage of communication (resp. computing) resources during a period, and 3) an upfront combo plan for unlimited usage of both communication and computing resources during a period. We aim to solve an online plan reservation problem: the amount of resources needed by a task is only known when it arrives, i.e., the future is unknown. However, even if the resource usage of future tasks is known in advance, the plan reservation problem is NP-hard and thus challenging. To tackle this problem, we propose a randomized online reservation (ROR) scheme to reserve plans probabilistically, where the probability is determined by the recent usage of resources. The performance gap (competitive ratio) between our proposed scheme and optimal solution is analyzed and derived in closed-form, and this gap is proved to be the minimum among all online algorithms which do not know the usage of future tasks. Trace-driven simulations verify the cost advantage of ROR and characterize how different prices of plans influence users' plan reservation strategies.**

***Index Terms*—Multimedia Computing, Competitive Analysis, Online Algorithm, Cost Management.**

I. Introduction

With the rapid development of advanced multimedia applications such as augmented reality (AR) [1] and self-driving cars [2], the resultant avalanche of multimedia data needs to be not only delivered by low-latency communication links, but also processed promptly by high-capacity computing facilities. For example, in self-driving cars, on-board cameras continuously record and transmit detailed information of the surrounding traffic environment to a computing server for video processing. Thanks to the recent advance in communication and computing technologies, 4G wireless networks can support up to 10 Mbps links [3] and the data rate can be further expanded to 1 Gbps [4] in upcoming 5G age, whilst an Amazon cloud computing instance (m4.xlarge) can provide up to 4 virtual central processing units (vCPUs) [5]. This is sufficient to realize a wide variety of interactive multimedia applications, such as Pokemon Go (AR game), which requires 1 Mbps data rate and the computing power of 2 vCPUs [6].



(a) Plan Reservation Strategy without Combo Plan

(b) Plan Reservation Strategy with Combo Plan

Fig. 1: Single Period Optimal Offline Plan Reservation

These advanced multimedia applications inspire a great market opportunity for multimedia service providers to design attractive plans offering joint communication and computing resources. Nevertheless, similar to early 3G mobile plans with separate data usage and voice calls charges [7], we envision that users will initially be charged for their communication and computing resources separately. Under separate schemes, users typically purchase the communication and computing resources in a pay-as-you-go (PAYG) fashion, or pay an upfront fee for unlimited usage of reserved communication or computing resources in a period (data or computing plan reservations) [8, 9]. However, such separate pricing schemes mismatch multimedia applications which offload computing-intensive tasks to computing servers and require joint provisioning of communication and computing resources. This will motivate combo plans that bundle the sale of both commu-

S. Zang, P. L. Yeoh, B. Vucetic, and Y. Li are with the School of Electrical and Information Engineering, Faculty of Engineering and Information Technologies, The University of Sydney, Sydney, NSW, 2006, Australia (e-mail: szan1435@uni.sydney.edu.au; phee.yeoh@sydney.edu.au; branka.vucetic@sydney.edu.au; yonghui.li@sydney.edu.au)

W. Bao is with the School of Computer Science, Faculty of Engineering and Information Technologies, The University of Sydney, Sydney, NSW, 2006, Australia (e-mail: wei.bao@sydney.edu.au)

nication and computing resources to fill two needs with one deed, which will offer incentives for both service providers and users [10]. The combo plan is more efficient to the service providers, as it reduces their marketing and distribution costs [11]. Due to its cost efficiency, the combo plan can also offer cost savings to users compared with separate data and computing plans, which is more attractive.

The impact of the combo plan on the multimedia services is illustrated in Fig. 1 by considering the plan reservation for a single period in an offline manner, where users know the resource usage as well as arrival time of future offloading tasks in advance. Without the combo plan, the plan reservation strategies for required communication and computing resources can be considered separately in Fig. 1a. Users will reserve data plans (orange region in Fig. 1a) if the cost of required communication resources by PAYG is greater than the data plan upfront fee C_d, and will reserve computing plans (green region in Fig. 1a) if the cost of required computing resource by PAYG is greater than the computing plan upfront fee C_c. The attractiveness of the combo plan is demonstrated in Fig. 1b where users need to consider both communication and computing resources together. If the cost of required computing and communication resources by PAYG is greater than the combo plan upfront fee C_o, they will reserve a combo plan to save money (yellow region in Fig. 1b).

In reality, it is not straightforward to determine the plan reservation strategy based on Fig. 1b. An obstacle for users to reserve plans is that they have very limited information about the *future* task information which includes the communication and computing resource usage and the arrival time of the task. Due to the bursty feature of multimedia traffic [12, 13] and the unpredictability of real world workload [14], the task information is only known when the task arrives. Therefore, users cannot locate the region in Fig. 1b to choose the optimal plan for cost minimization. For each type of resource, simply choosing PAYG costs more than the plan when the future usage is high [5], while reserving a plan wastes the upfront fee when the future usage is low or even zero. When both communication and computing resources are required, users need to estimate whether a combo plan will save money compared with the combination of PAYG and an individual data or computing plan. Furthermore, when the plan reservation strategy is made for multiple periods, the scheduling of plan reservation depends on the choice of the reserved plan above, and makes the whole plan reservation problem even more complicated.

In this paper, we investigate the impact of combo plan pricing on multimedia services. Specifically, we address the key question: *How do users reserve plans for cost minimization without future information?* When the future task information is known a priori, the plan reservation problem is NP-hard (Section III). To tackle the problem without the future task information, we formulate an online optimization problem where we make no assumptions on the future. We solve the online optimization problem by adopting the competitive analysis method [15] and proposing a novel randomized online reservation (ROR) scheme. By comparing the combo plan price with prices of individual data and computing plans, the ROR scheme will categorize the combo plan into different cases. Under each case, the ROR scheme design specific probability mass functions to reserve plans probablistically based on the recent usage of communication and computing resources. We prove that our proposed ROR scheme achieves a constant competitive ratio[1] for any multimedia traffic patterns and the *minimum* competitive ratio among all online algorithms. By considering all users adopt the ROR scheme due to its cost advantage, we derive a threshold value which will be beneficial for the design of combo plan pricing. All users will select the combo plan over individual data and computing plans when the price of combo plan is less than or equal to the threshold value. The contributions of our paper are summarized as follows:

- We propose a practical randomized online reservation scheme for users to minimize the plan reservation cost by characterizing the user plan reservation strategy under different combo plan prices, which does not require any future information and matches the bursty and unpredictable nature of the multimedia traffic.
- We conduct theoretical analysis of online joint resource reservation problem, and prove that the proposed randomized online reservation scheme achieves the minimum competitive ratio among all online algorithms in all plan pricing cases.
- The derived threshold value for the price of combo plan ensures that the combo plan is more attractive compared with individual plans, and will be beneficial for the design of combo plan pricing.

We verify the cost minimization advantage of our ROR scheme over online benchmark schemes using trace-driven simulations. We illustrate how the user reserves plans with different combo plan prices in the ROR scheme. By investigating the impact of the reservation period, we conclude that a lengthy reservation period will increase the cost of the plan reservation and force users to select PAYG for services.

II. Related Work

In this section, we review existing research works on service provider pricing design and user resource reservation strategies in communication and computing markets, as well as bundling strategies and their applications.

A. Pricing and Resource Reservation for Communication

In [16] and [17], market-driven pricing models adopted by communication network service providers were classified as static [18, 19] and dynamic [20–22] pricing schemes. By considering resource reservation strategies of users, dynamic pricing schemes have been considered for network congestion control [20, 21] and resource control [22]. Multimedia specific pricing models have been developed in [23–28]. A dynamic and congestion-sensitive pricing structure was proposed in [23] for real-time multimedia services. In [24], a quality-of-video

[1]The competitive ratio here is the ratio between the cost of our proposed ROR scheme and the cost of optimal offline scheme which solves the plan reservation problem optimally with all future information to be known.

oriented pricing scheme was proposed. Video streaming based rate control and resource allocation was considered in [25–27]. In [28], a pricing-based online algorithm was proposed to allocate communication and computing resources for real-time multimedia services by minimizing the cost of the service provider. The problem was formulated as one of lowest cost subgraph packing, and the proposed algorithm achieved a constant competitive ratio.

Pricing plan designs were considered in [29], where users chose smart phone plans based on their cost, service, and flexibility. Secondary users' strategies for revenue maximization were analyzed in [30], where users required spectrum in a non-cooperative game framework for cognitive radio spectrum sharing. In an auction game framework for bandwidth allocation [31], users' strategies consisted of required amount of bandwidth and the expected price of the bandwidth and aimed for utility maximization. With both upfront plan reservation and PAYG payment options, a spectrum acquisition framework for mobile virtual network operators was investigated in [32] assuming a given distribution of traffic statistics and future PAYG price prediction. This framework is not applicable for computing-intensive multimedia services where future data traffic statistics are unpredictable.

B. Pricing and Resource Reservation for Computing

In cloud computing, static pricing schemes with fine granularity were investigated in [33]. Dynamic pricing mechanisms were proposed in [34] based on market demand, and in [35] based on customer's historical records. In [36], a parameter-based cloud resource pricing model was proposed by considering the Black-Scholes-Merton model and service provider investments. Game theory based pricing schemes for cloud computing have been considered in [37, 38].

In terms of computing instance reservations, an offline heuristic greedy algorithm and a 3-competitive rolling-horizon based online algorithm were proposed in [39] with PAYG and upfront plan pricing. By considering the price and demand uncertainties, a stochastic programming model was formulated in [40] for long-term cloud resource provisioning with multiple stages. In [41], an online multi-instance reservation framework was proposed for IaaS clouds without considering the future information. The proposed strategy was equivalent to solving the Bahncard problem when the user demand was only one instance at a time, and both proposed deterministic as well as randomized online reservation algorithms had optimal competitive ratios. However, this strategy cannot be directly applied to multimedia services as the communication resource reservation and combo plan are not considered. In our previous work [42], a deterministic online plan reservation algorithm was proposed for the edge-computing services which jointly considered the communication and computing resources. The proposed ROR scheme is a probability distribution of previous deterministic online plan reservation algorithms. Therefore, our proposed ROR scheme is more advanced and offers better competitive ratio as it uses a more general probabilistic approach to reserve plans.

TABLE I: Summary of Notations

Notation	Parameter
$d(i), c(i), t(i)$	Data volume, computing workload, and arrival time of task i
Δp	Pricing granularity
$C_d\Delta p$, $C_c\Delta p$, $C_o\Delta p$	Data plan, computing plan, and combo plan upfront fees
T	Plan reservation period
$x_d(i)$, $x_c(i)$	Binary decisions to satisfy the data volume and computing workload of task i by PAYG
$y_d(i)$, $y_c(i)$, $y_o(i)$	Binary decisions to reserve the data plan, computing plan, and combo plan at $t(i)$
$\mathbf{S}$	Multimedia task sequence
θ_d, θ_c	Typical data cost and typical computing cost
$\gamma_o^{(1)}, \gamma_o^{(2)}$	Thresholds to reserve combo plans
γ_d, γ_c	Thresholds to reserve data and computing plans

C. Bundling Strategy

The strategic motivation for bundling sale was analyzed in [11]. In [43], optimal bundling product pricing strategies were proposed for each segmented user groups based on their willingness to pay (maximum price the user group will buy the product). A probabilistic approach for pure products, bundle products, and mixed products pricing designs was considered in [44] based on the statistics of user demand and willingness to pay. The effect of bundle products on the user's willingness to pay was investigated by surveys in [45]. In [46], the impact of the bundle product pricing on user behaviour was analyzed by surveys. In this paper, due to the bursty and unpredictabe nature of the multimedia traffic, we perform the bundled product (combo plan) pricing analysis by modeling the user behaviour without assuming the user's demand and willingness to pay.

III. Problem Formulation

In this section, we formulate the online plan reservation problem for multimedia services with unknown future task information. In our system model, we consider that computing-intensive multimedia tasks on the user side are transmitted via communication links to computing servers for processing. Although the computing capability of modern mobile devices has been improved, offloading computing intensive tasks to the server can reduce the energy consumption and offer better performance compared with local processing. The important notations in this paper are summarized as Table I.

In our system, a user can have more than one device (self-driving car, AR kit, etc) and different types of computing-intensive multimedia tasks (object tracking, video processing, etc) to offload. We consider the scenario where different multimedia tasks labeled by $i \in \{1, 2, 3, \ldots\}$ arrive from devices of the user, and are served by the computing servers. Each multimedia task i is characterized by a 3-tuple $\big(d(i), c(i), t(i)\big)$, where $d(i)$ is the data volume in MB required by task i, $c(i)$

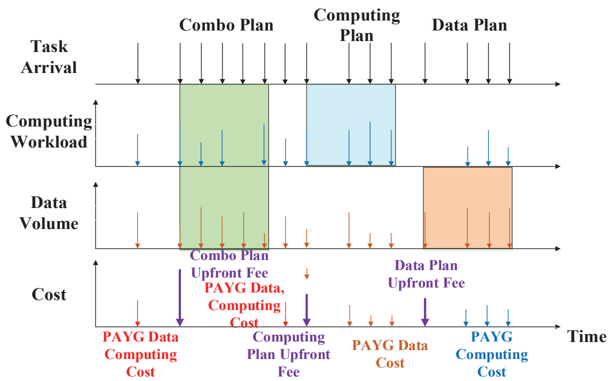


Fig. 2: Task Description and Payment Options

is the computing workload in vCPU·hours to serve task i, and $t(i)$ is the arrival time instance of task i. Without loss of generality, we assume $t(1) < t(2) < \ldots$, and $t(i)$ can be any positive real number.

As multimedia traffic is bursty [12, 13] and workload of real applications is unpredictable [14], we consider that the user does not have information about future multimedia task arrivals. In other words, the user only knows $\big(d(i), c(i), t(i)\big)$ at $t(i)$, i.e., when the ith task arrives. This is formally referred as an online system [15]. In other words, we do not have any assumption on the distribution of multimedia task arrivals. Furthermore, we consider that each multimedia task should be processed immediately by paying the service provider from the pricing options outlined below, as most real-time multimedia applications (AR and self-driving car) have low-latency requirements.

A. PAYG and Plan Pricing

Fig. 2 illustrates how the user can satisfy a sequence of multimedia tasks by selecting different pricing options. Each multimedia task is labelled by an arrow and consists of the corresponding computing workload as well as data volume. The coverage of each plan is described by rectangular regions with different colors, and the cost incurred by each payment option is presented in the last coordinate. Detailed description of each payment option is presented as follows.

1) PAYG: PAYG lets the user pay for the multimedia service based on usage (data volume and computing workload) without any time commitments. The PAYG model is well-supported in both cloud computing (on-demand instances) and wireless communications (prepaid data packages).

The PAYG data price is $\$\Delta p$ per M MB and the PAYG computing price is $\$\Delta p$ per N vCPU·hours. Δp is the pricing granularity in terms of money, and the unit of Δp is in dollars. We consider that both data and computing PAYG adopt the same pricing granularity as they are from the same service provider. If less than M MB of data volume or N vCPU·hours of computing workload is used, Δp dollars will still be charged. For example, a multimedia task with a MB of data volume and b vCPU·hours of computing workload will incur a PAYG cost of $\lceil \frac{a}{M} \rceil \Delta p$ + $\lceil \frac{b}{N} \rceil \Delta p$ dollars.

2) Plans: Compared with PAYG, plans allow a user to use a reserved amount of resources for a given time period by prepaying an upfront fee. Separate plans are currently available for both wireless communication (cellular data plans) and cloud computing (reserved instances). For multimedia tasks, we also consider the dedicated combo plan offering both data volume and computing workload. The above three types of plans are defined as follows:

Data plan: An upfront fee $C_d \Delta p$ is paid at time t to reserve a communication link; all data volume through the communication link required from multimedia tasks in $[t, t+T)$ will be free of charge (orange region in Fig. 2).

Computing plan: An upfront fee $C_c \Delta p$ is paid at time t to reserve a computing instance; all computing workload on the computing instance required from multimedia tasks in $[t, t+T)$ will be free of charge (blue region in Fig. 2).

Combo plan: An upfront fee $C_o \Delta p$ is paid at time t to reserve a communication link and a computing instance; all data volume through the communication link and computing workload on the computing instance required from multimedia tasks in $[t, t+T)$ will be free of charge (green region in Fig. 2).

For each plan, T is the reservation period which determines the frequency of a user selecting plans (width of each colored rectangular region in Fig. 2). For the price of combo plan, we assume $\max(C_d, C_c) \le C_o \le C_d + C_c$. Otherwise, it is straightforward to show that reserving combo plans is always better or worse than reserving data and computing plans individually.

B. The Plan Reservation Problem: Offline vs Online

Upon the arrival of a multimedia task i, if the task is already fully covered by the previous plan, the user does not need to pay for task i. Otherwise, the user needs to pay for task i by choosing from the following options: PAYG, data plan, computing plan, and combo plan. Reserving a plan will involve a high upfront reservation cost, but it may save money if there are many multimedia tasks arriving in the upcoming reservation period. To minimize the overall multimedia service cost, the user needs to effectively combine the above payment options without any future knowledge.

For each multimedia task i, let $x_d(i)$ and $x_c(i)$ denote the decisions to satisfy the required data volume and computing workload by PAYG, respectively. Let $y_d(i)$, $y_c(i)$, and $y_o(i)$ denote the decisions to reserve the data plan, computing plan, and combo plan, respectively. For each of the above decision variables $\big(x_d(i),\, x_c(i),\, y_d(i),\, y_c(i),\, y_o(i)\big)$, its value of 1 represents selecting the option while 0 represents not selecting the option. For example, $(1, 0, 0, 1, 0)$ means that the data volume required by task i is satisfied by PAYG, and a computing plan is reserved at $t(i)$. Suppose there are I multimedia tasks in the multimedia task sequence where I is an arbitrary number. We first define the following offline plan reservation problem as (1). In (1), C in the objective function is the total cost to minimize. The first constraint ensures that the data volume of each multimedia task i is satisfied where $\sum_{j:t(i)-T<t(j)\le t(i)} d_j + \sum_{j:t(i)-T<t(j)\le t(i)} o_j$ is the number of reserved plans which can cover $d(i)$. The second constraint ensures that the computing workload of each multimedia task i is

$$
\begin{aligned}
\min_{\{x_d(i), x_c(i), y_d(i), y_c(i), y_o(i)\}} \quad & C = \sum_{i=1}^{I} \Delta p (x_d(i) \lceil \frac{d(i)}{M} \rceil + x_c(i) \lceil \frac{c(i)}{N} \rceil + y_d(i) C_d + y_c(i) C_c + y_o(i) C_o) \\
\text{s.t.} \quad & x_d(i) + \sum_{j:t(i)-T<t(j)\le t(i)} d_j + \sum_{j:t(i)-T<t(j)\le t(i)} o_j > 0, \qquad \forall i \in [1, I], \\
& x_c(i) + \sum_{j:t(i)-T<t(j)\le t(i)} c_j + \sum_{j:t(i)-T<t(j)\le t(i)} o_j > 0, \qquad \forall i \in [1, I], \\
& x_d(i), x_c(i), y_d(i), y_c(i), y_o(i) \in \{0, 1\}, \forall i \in [1, I].
\end{aligned}
\tag{1}
$$

satisfied where $\sum_{j:t(i)-T<t(j)\le t(i)} c_j + \sum_{j:t(i)-T<t(j)\le t(i)} o_j$ is the number of reserved plans which can cover $c(i)$. In the offline plan reservation problem, task arrivals are known a priori, and this is a binary integer programming problem that is NP-hard [47]. In the online version of the plan reservation problem, data volume $d(i)$ and computing workload $c(i)$ of task i is only know when the task arrives at time $t(i)$. Therefore, the online plan reservation problem is even more challenging.

C. Measure of Competitiveness

To measure cost performance of an online plan reservation strategy, we adopt the conventional competitive analysis [15]. The main idea is to compare the cost of the online strategy with the optimal offline strategy which is obtained by solving (1) offline after the sequence of multimedia task arrivals is known.

Let $\mathbf{S} = \{(d(1), c(1), t(1)), \ldots, (d(I), c(I), t(I))\}$ denote the sequence of all multimedia tasks. A randomized online reservation algorithm A is *c-competitive* (c is a constant) if for all possible multimedia task sequences $\mathbf{S}$, we have

$$\mathbb{E}[C_A(\mathbf{S})] \le c \cdot C_{OPT}(\mathbf{S}), \tag{2}$$

where $C_A(\mathbf{S})$ is the overall cost of algorithm A given the input $\mathbf{S}$, and $C_{OPT}(\mathbf{S})$ is the overall cost of the optimal offline strategy given the input $\mathbf{S}$. Here, we also denote c as the competitive ratio of algorithm A. The smaller the competitive ratio is, the more closely the online algorithm A approaches the optimal solution. Our objective is to design the online reservation algorithm with the minimum competitive ratio[2].

The online reservation algorithm decides when to reserve which plans only based on the recent multimedia task sequence information. The deterministic online reservation algorithm makes the same plan reservation strategy given the same multimedia task sequence. The randomized online reservation algorithm is a probability distribution of the deterministic online reservation algorithm, which is more general and advanced. In the following, we will introduce and explain our randomized online reservation scheme, and prove that it achieves the minimum competitive ratio among all possible online algorithms.

[2]We note that the online plan reservation problem captures the ski rental problem as a special case [48] when a user only needs to reserve a single type of resource. However, our problem is substantially more complicated because there are multiple resources and an additional combo plan option which covers both the data volume and computing workload.

IV. Problem Solution: ROR Scheme

In this section, we present our randomized online reservation (ROR) scheme for the user to make the plan reservation strategy without future task information. Intrinsically, the ROR scheme reserves different types of plans probabilistically, and the probability is prudently tuned to minimize the cost (i.e., with a provable minimum competitive ratio). Later, we further present an insightful design guideline for the price of combo plan so that the combo plans are more attractive compared with individual data and computing plans.

A. The Randomized Online Reservation Scheme

First, we introduce the typical data cost and typical computing cost which are needed for a plan reservation decision in the ROR scheme. Recall that $y_d(i)$, $y_c(i)$, and $y_o(i)$ denote whether a data plan, computing plan, and combo plan are reserved at $t(i)$, respectively, $y_d(i), y_c(i), y_o(i) \in \{0, 1\}$. If a multimedia task i is not covered by plans, it must be satisfied by PAYG. *We define* that the typical data cost θ_d at time t is the sum of PAYG data cost during $(t-T, t]$, and can be expressed as (3). $\mathbb{1}(\cdot)$ in (3) is the indicator function, and $\mathbb{1}\Big(\sum_{j:t(i)-T<t(j)\le t(i)} y_d(j) + \sum_{j:t(i)-T<t(j)\le t(i)} y_o(j) < 1\Big)$ checks whether the data volume $d(i)$ is satisfied by a previous plan. *We define* that the typical computing cost θ_c at time t is the sum of PAYG computing cost during $(t-T, t]$, and can be expressed as (4).

Recall that $C_d\Delta p$, $C_c\Delta p$ and $C_o\Delta p$ are the prices of the data plan, computing plan, and combo plan, respectively. To achieve the minimum competitive ratio, we categorize the combo plan into different cases below according to two terms, $\sqrt{C_d(C_d+C_c)}\Delta p$ and $\sqrt{C_c(C_d+C_c)}\Delta p$, which is illustrated in Fig. 3.

Case 1: We *define* the combo plan is inexpensive if $C_o < \min\Big(\sqrt{C_d(C_d+C_c)}, \sqrt{C_c(C_d+C_c)}\Big)$.

Case 2.1: We *define* the combo plan is moderate if $\sqrt{C_d(C_d+C_c)} \le C_o < \sqrt{C_c(C_d+C_c)}$ and $C_d < C_c$.

Case 2.2: We *define* the combo plan is moderate if $\sqrt{C_c(C_d+C_c)} \le C_o < \sqrt{C_d(C_d+C_c)}$ and $C_d > C_c$ (Cases 2.1 and 2.2 can be ignored if $C_d = C_c$.).

Case 3: We *define* the combo plan is expensive if $C_o \ge \max\Big(\sqrt{C_d(C_d+C_c)}, \sqrt{C_c(C_d+C_c)}\Big)$.

These two terms are derived by comparing the minimum competitive ratios of reserving combo plans and reserving data and computing plans individually, which is given in the proof of Theorem 2. These terms indicate that we should use

$$\theta_d = \sum_{i:t-T<t(i)\le t} \mathbb{1}\Big(\sum_{j:t(i)-T<t(j)\le t(i)} y_d(j) + \sum_{j:t(i)-T<t(j)\le t(i)} y_o(j) < 1\Big)\lceil \frac{d(i)}{M}\rceil \Delta p, \tag{3}$$

$$\theta_c = \sum_{i:t-T<t(i)\le t} \mathbb{1}\Big(\sum_{j:t(i)-T<t(j)\le t(i)} y_c(j) + \sum_{j:t(i)-T<t(j)\le t(i)} y_o(j) < 1\Big)\lceil \frac{c(i)}{N}\rceil \Delta p. \tag{4}$$

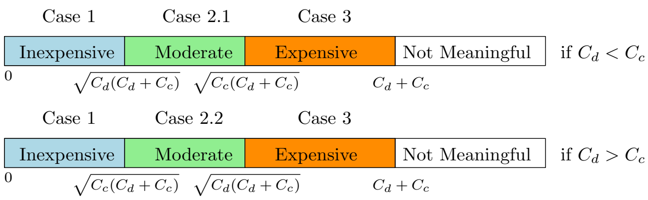


Fig. 3: Interpretation of Combo Plan Classification

different algorithms in different cases to ensure the minimum competitive ratio is reached. In the following, we are going to present ROR in different cases.

1) Case 1: Inexpensive Combo Plan: As the combo plan is inexpensive, we design Algorithm 1 to reserve combo plans whenever there are opportunities.

Algorithm 1 starts by initializing $y_d(i)$, $y_c(i)$, and $y_o(i)$, and threshold parameters $\gamma_o^{(1)}$ as well as $\gamma_o^{(2)}$ for combo plan reservations. The algorithm generates $\gamma_o^{(1)} \in \{1, 2, ..., C_d\}$ and $\gamma_o^{(2)} \in \{1, 2, ..., C_c\}$ according to probability mass functions (pmfs) $f_1(\gamma_o^{(1)})$ and $f_2(\gamma_o^{(2)})$, respectively, *which are defined as*

$$f_1(\gamma_o^{(1)}) = \Big(\frac{C_o - 1}{C_o}\Big)^{C_d - \gamma_o^{(1)}} \frac{1}{C_o\big(1 - (1 - 1/C_o)^{C_d}\big)}, \tag{5a}$$

$$f_2(\gamma_o^{(2)}) = \Big(\frac{C_o - 1}{C_o}\Big)^{C_c - \gamma_o^{(2)}} \frac{1}{C_o\big(1 - (1 - 1/C_o)^{C_c}\big)}. \tag{5b}$$

If there is no ongoing combo plan to cover the newly arrived multimedia task, the algorithm will update the typical data cost θ_d and typical computing cost θ_c with $t = t(i)$ according to (3) and (4). The algorithm will reserve a combo plan if we have "$\theta_d \ge \gamma_o^{(1)}\Delta p$" (Line 5) or "$\theta_c \ge \gamma_o^{(2)}\Delta p$" (Line 7). Otherwise, the task is satisfied by a PAYG payment. To introduce randomness, the threshold parameter in the satisfied condition is generated again according to its corresponding pmf. For example, if the combo plan is reserved due to "$\theta_d \ge \gamma_o^{(1)}\Delta p$" in Line 5, the threshold parameter $\gamma_o^{(1)}$ will be generate again according to $f_1(\gamma_o^{(1)})$.

It can be seen that $f_1(\gamma_o^{(1)})$ and $f_2(\gamma_o^{(2)})$ control how the ROR scheme reserves combo plans probabilistically. The design of pmfs have a direct impact on the overall cost as well as the competitive ratio, and the best pmfs are derived by solving the competitive ratio minimization problem. The detailed derivation is given in the proof of Lemma 4.

2) Cases 2 and 3: Moderate and Expensive Combo Plans: In these cases, we design Algorithm 2 to reserve data and computing plans separately as the combo plan is not cheap.

Algorithm 1 ROR Scheme Case 1: Inexpensive Combo Plan

1: Initialize $\gamma_o^{(1)}$ and $\gamma_o^{(2)}$ according to (5a) and (5b), respectively, and set $y_d(i) \leftarrow 0$, $y_c(i) \leftarrow 0$, $y_o(i) \leftarrow 0$ for all $i = 1, 2, ...$
2: Upon the new arrival of task i. ▷ $d(i)$ and $c(i)$ are only known at $t(i)$.
3: **if** there is no ongoing combo plan **then**
4: Calculate θ_d and θ_c according to (3) and (4), respectively.
5: **if** $\theta_d \ge \gamma_o^{(1)}\Delta p$ **then**
6: Reserve a combo plan: $y_o(i) \leftarrow y_o(i) + 1$, and generate $\gamma_o^{(1)}$ by (5a).
7: **else if** $\theta_c \ge \gamma_o^{(2)}\Delta p$ **then**
8: Reserve a combo plan: $y_o(i) \leftarrow y_o(i) + 1$, and generate $\gamma_o^{(2)}$ by (5b).
9: **else**
10: Satisfy the task by PAYG.
11: **end if**
12: **end if**
13: Repeat from step 2.

The key difference between Algorithm 1 and Algorithm 2 is that data and computing plans are reserved.

The initialization stage of Algorithm 2 is similar to Algorithm 1. In addition to the threshold parameters of the combo plan, we also define the threshold $\gamma_d \in \{1, 2, ..., C_d\}$ for the data plan reservation, and threshold $\gamma_c \in \{1, 2, ..., C_c\}$ for the computing plan reservation. These thresholds are generated according to their respective pmfs, $g(\gamma_d)$ and $h(\gamma_c)$, which *are defined as*

$$g(\gamma_d) = \Big(\frac{C_d - 1}{C_d}\Big)^{[\frac{C_d C_o}{C_d + C_c}] - \gamma_d} \frac{1}{C_d\big(1 - (1 - 1/C_d)^{[\frac{C_d C_o}{C_d + C_c}]}\big)}, \tag{6}$$

$$h(\gamma_c) = \Big(\frac{C_c - 1}{C_c}\Big)^{[\frac{C_c C_o}{C_d + C_c}] - \gamma_c} \frac{1}{C_c\big(1 - (1 - 1/C_c)^{[\frac{C_c C_o}{C_d + C_c}]}\big)}. \tag{7}$$

In (6) and (7), $[\cdot]$ represents the nearest integer function. We can see that $g(\gamma_d)$ and $h(\gamma_c)$ control how the ROR scheme reserves data and computing plans probabilistically, and have the direct impact on the competitive ratio. The best pmfs are derived by solving the competitive ratio minimization problem and the derivation is given in the proof of Lemma 5.

If there is no ongoing data plan, Algorithm 2 will reserve a data plan when we have " $\theta_d \ge \gamma_d \Delta p$ "(Line 6).

Algorithm 2 will reserve a computing plan when we have "$\theta_c \geq \gamma_c \Delta p$" (Line 13), if there is no ongoing computing plan. Same as Algorithm 1, Algorithm 2 generates the threshold parameter in the satisfied condition to introduce randomness after a plan is reserved. If no plan is reserved and the task cannot be covered by previous plans, the task will be satisfied by PAYG.

Algorithm 2 ROR Scheme Cases 2 and 3: Moderate and Expensive Combo Plans

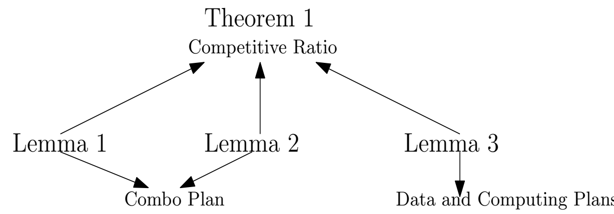
```
1:  Initialize γ_d and γ_c according to (6), and (7) respectively,
    and set y_d(i) ← 0, y_c(i) ← 0, y_o(i) ← 0 for all i = 1, 2, ...
2:  Upon the new arrival of task i.  ▷ d(i) and c(i) are only
    known at t(i).
3:  if there are no ongoing plans fully covering the task then
4:      Calculate θ_d and θ_c according to (3) and (4), respec-
        tively.
5:      if there is no ongoing data plan then
6:          if θ_d ≥ γ_d Δp then
7:              Reserve a data plan: y_d(i) ← y_d(i) + 1.
8:              Generate γ_d by (6).
9:          else
10:             Satisfy d(i) by PAYG.
11:         end if
12:     else if there no ongoing computing plan then
13:         if θ_c ≥ γ_c Δp then
14:             Reserve a computing plan: y_c(i) ← y_c(i) + 1.
15:             generate γ_c by (7).
16:         else
17:             Satisfy c(i) by PAYG.
18:         end if
19:     end if
20: end if
21: Repeat from step 2.
```

B. Performance Analysis of ROR Scheme

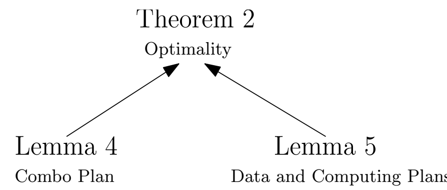
In the following, we analyze cost performance of our ROR scheme, and prove that it achieves the minimum competitive ratio among all online algorithms. To do this, we proceed to analyze the cost of the ROR scheme C_{ROR} relative to the cost of the optimal offline scheme C_{OPT}. The relationships of proofs are shown as Fig. 4. The competitive ratio of each threshold based plan reservation condition in each case of ROR scheme is derived in the following lemmas to support the competitive ratio analysis in Theorem 1.

Lemma 1. *When the ROR scheme reserves combo plans according to condition $\theta_d \geq \gamma_o^{(1)} \Delta p$, we have $\frac{\mathbb{E}[C_{ROR}(\mathbf{S})]}{C_{OPT}(\mathbf{S})} < \frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}$, $\forall$ $\mathbf{S}$.*

Proof. The competitive ratio in this situation can be derived as $\frac{1}{1-(1/e)^{\frac{C_d}{C_o}}}$ when the combo plans are reserved based on the typical data cost. We have $\frac{1}{1-(1/e)^{\frac{C_d}{C_o}}} < \frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}$, as $C_o < \sqrt{C_d(C_d+C_c)}$ in Case 1. The proof of Lemma 1 is given in the Appendix. $\square$



(a) Proof Relationship of Theorem 1



(b) Proof Relationship of Theorem 2

Fig. 4: Logic connection of proofs

Lemma 2. *When the ROR scheme reserves combo plans according to condition $\theta_c \geq \gamma_o^{(2)} \Delta p$ in Case 1, we have $\frac{\mathbb{E}[C_{ROR}(\mathbf{S})]}{C_{OPT}(\mathbf{S})} < \frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}$, $\forall$ $\mathbf{S}$.*

Proof. The competitive ratio in this situation can be derived as $\frac{1}{1-(1/e)^{\frac{C_c}{C_o}}}$ by following the similar steps in the proof of Lemma 1. We have $\frac{1}{1-(1/e)^{\frac{C_c}{C_o}}} < \frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}$, as $C_o < \sqrt{C_c(C_d+C_c)}$ in Case 1. $\square$

When the individual data and computing plans are reserved based on typical computing cost and typical data cost, respectively, the corresponding competitive ratios are derived in Lemma 3.

Lemma 3. *When the ROR scheme reserves data plans and computing plans individually, we have $\frac{\mathbb{E}[C_{ROR}(\mathbf{S})]}{C_{OPT}(\mathbf{S})} \leq \frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}$,$\forall$ $\mathbf{S}$.*

Proof. This corresponds to the ROR scheme satisfying the reservation conditions $\theta_d \geq \gamma_d \Delta p$ for data plans and $\theta_c \geq \gamma_c \Delta p$ for computing plans. The proof of Lemma 3 is given in the Appendix. $\square$

Based on the above lemmas, we derive the competitive ratio of ROR scheme in all cases in Theorem 1.

Theorem 1. *The ROR scheme is $\frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}$ - competitive. Formally, for any multimedia task sequence $\mathbf{S}$, we have*

$$\mathbb{E}[C_{ROR}(\mathbf{S})] \leq \frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}} \cdot C_{OPT}(\mathbf{S}). \tag{8}$$

Proof. The theorem is proved by showing that the competitive ratio of the ROR scheme under all cases (inexpensive combo plan, moderate combo plan, and expensive combo plan) is less than or equal to $\frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}$.

The competitive ratio of the ROR scheme under Case 1 is less than $\frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}$ according to Lemmas 1, and 2. The same competitive ratio for Cases 2 and 3 is found according to Lemmas 3. As a result, the competitive ratio of the ROR scheme for all cases is $\frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}$. □

Compared with randomized online algorithms, deterministic online algorithms reserve plans according to thresholds of fixed values rather than thresholds generated from probability mass functions. Therefore, a deterministic online algorithm is a special case of randomized algorithms. In the following, we investigate the optimality of the proposed ROR scheme among all deterministic and randomized online algorithms.

In order for the competitive ratio of an online algorithm to be bounded in a period of T (length of reservation period), both data volume and computing workload of multimedia tasks need to be covered by plans, as there can be infinite numbers of multimedia tasks arriving during the period in the worst case. In terms of the plan reservation, reserving a combo plan with an individual plan during the same period leads to a higher cost than reserving the individual plan after the reservation period of the combo plan. Therefore, only two options are left for an online algorithm: Option 1) Satisfy the required data volume and computing workload by PAYG and the combo plan; Option 2) Satisfy the required data volume and computing workload by PAYG and individual plans. The following lemmas present the minimum competitive ratios for Options 1 and 2, respectively.

Lemma 4. *Let $\mathcal{C}$ denote the set of competitive ratios of all online algorithms which only satisfy the multimedia tasks by PAYG and combo plans. We have* $\min \mathcal{C} = \max(\frac{1}{1-(1/e)^{\frac{C_d}{C_o}}}, \frac{1}{1-(1/e)^{\frac{C_c}{C_o}}})$.

Proof. The proof of Lemma 4 is given in the Appendix. □

Lemma 5. *Let $\mathcal{I}$ denote the set of competitive ratios of all online algorithms which only satisfy the multimedia tasks by PAYG and individual data and computing plans. We have* $\min \mathcal{I} = \frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}$.

Proof. The proof of Lemma 5 is given in the Appendix. □

Based on Lemmas 4 and 5, the optimality of the ROR scheme is proved in Theorem 2.

Theorem 2. *The ROR scheme achieves the minimum competitive ratio among all deterministic and randomized online algorithms.*

Proof. The theorem is proved by showing that the competitive ratio of ROR scheme is the minimum among all randomized algorithms under all cases based on Lemmas 4 and 5, and the deterministic algorithms are special cases of randomized algorithms. A detailed proof is provided in the Appendix. □

C. Pricing Design Guideline of the Combo Plan

Given that the ROR scheme achieves the minimum competitive ratio, in the following discussion we assume that all rational users will adopt the ROR scheme for their plan reservations. The user behavior is summarized as follows. The user will select the combo plan by adopting Algorithm 1 in Case 1 when C_o is inexpensive, according to the condition $C_o < \min\left(\sqrt{C_d(C_d + C_c)}, \sqrt{C_c(C_d + C_c)}\right)$. This is the threshold value as the combo plan is more attractive compared with individual plans due to its price advantage. When the price of the combo plan is moderate or expensive, the user starts selecting the individual computing and data plans due to its flexibility (Algorithm 2). According to the above analysis, the price of the combo plan can be set to the threshold value to attract users whilst still maintaining profit.

V. Evaluation

In this section, we evaluate performance of the ROR scheme for a computing-intensive multimedia communications scenario based on actual mobile record traces for a real-world user in an urban university environment [49].

A. Trace Preprocessing and Plan Pricing

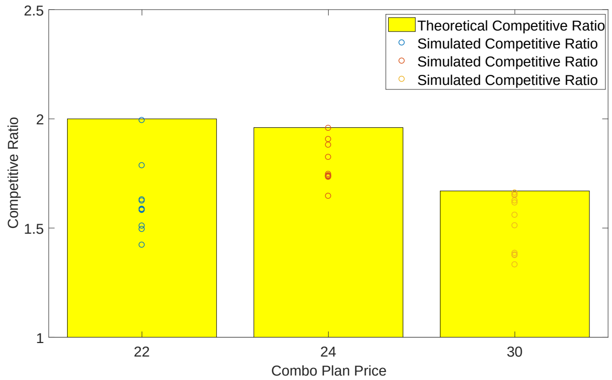
In our numerical evaluations, we use the record of time instances from [49] and associate it to different multimedia applications. We consider that the multimedia applications generate tasks including browsing, video streaming, machine learning, augmented reality, and video analysis according to Table II. A typical 480P video quality (default in YouTube [50]) is considered for both video streaming task as well as augmented reality task. Although both tasks have the same video length and quality, extra information is needed for the augmented reality task such as markers. For the video analysis task, 360P video quality is used to save the computing workload and data volume [51]. The data volume of per-minute video with different qualities can be seen at [52]. By changing the probability of arrival for each multimedia task, traces can be generated to represent different user behaviors.

TABLE II: Description of Multimedia Tasks

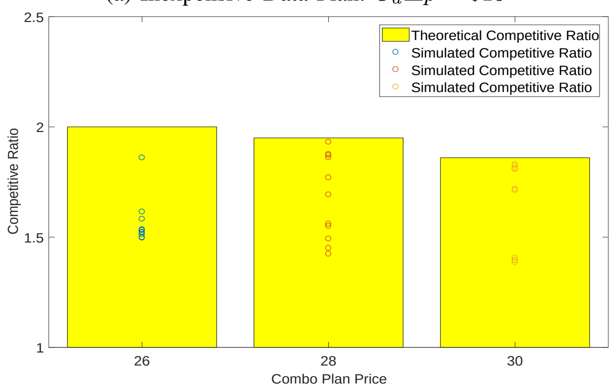
Tasks	Data Volume (MB)	Computing Workload (vCPU · hour)
Browsing	2 [53]	0
Video Streaming	15	0
Machine Learning	5 [54]	0.2185 [55]
Augmented Reality	20	0.87 [56]
Video Analysis	10	1.6 [57]

For the plan pricing, we set the pricing granularity to $\Delta p =$ \$0.05. Based on an Amazon EC2 t2.xlarge instance [8], the PAYG computing price is set to \$0.05 per vCPU·hour, and the computing plan upfront fee is set to $C_c\Delta p =$ \$18 for a week[3]. The communication prices are based on the AT&T [9]. PAYG data price is set to \$0.05 per 5 MB, and two data plan prices $C_d\Delta p =$ \$15 and $C_d\Delta p =$ \$21 for a week are considered to represent different QoS levels in data transmission.

[3]Since the trace spans for three months, we scale the plan reservation period to 1 week.



(a) Inexpensive Data Plan: $C_d\Delta p = \$15$



(b) Expensive Data Plan: $C_d\Delta p = \$21$

Fig. 5: Competitive Ratio Verification

B. Competitive Ratio Verification

We show that the competitive ratio is exactly $\frac{1}{1+e^{-\frac{C_d+C_c}{C_o}}}$ in this subsection. As it is NP-hard to get the optimal offline cost for the user trace used in the simulation, we produce some shorter traces (multimedia task sequences) by poisson process to verify the competitive ratio so that the optimal cost can be obtained by exhaustive search. The ratios between the expected costs of ROR scheme and optimal costs of multiple multimedia task sequences are plotted in Fig. 5 in cases of inexpensive, moderate, and expensive combo plans. It can be seen that all ratios are less than or equal to $\frac{1}{1+e^{-\frac{C_d+C_c}{C_o}}}$, and the theoretical competitive ratio is verified.

C. Cost Performance of Online and Offline Algorithms

In this subsection, we evaluate overall cost performance of our proposed ROR scheme by comparing it with the following benchmark strategies below.

PAYG Only (P): All tasks are satisfied by PAYG.

Combo Only (C): A combo plan is reserved when a new task arrives and the previous combo plan expires.

Individual Plan (I): A data or computing plan is reserved when new data volume or computing workload arrives and the previous plan expires.

Offline (O): The user knows all task arrivals in advance. Whenever a task arrives, the user decides to reserve a plan only if the upfront fee is less than the sum of PAYG costs for the plan reservation period.

Random Selection (S): When new data volume/computing workload arrives and the previous plan expires, PAYG data, data plans/computing plans, and combo plans are randomly selected with equal probabilities.

ROR without PAYG (W): As there is no PAYG in this scheme, we first *redefine* the typical data cost θ_d and typical computing cost θ_c in ROR without PAYG.

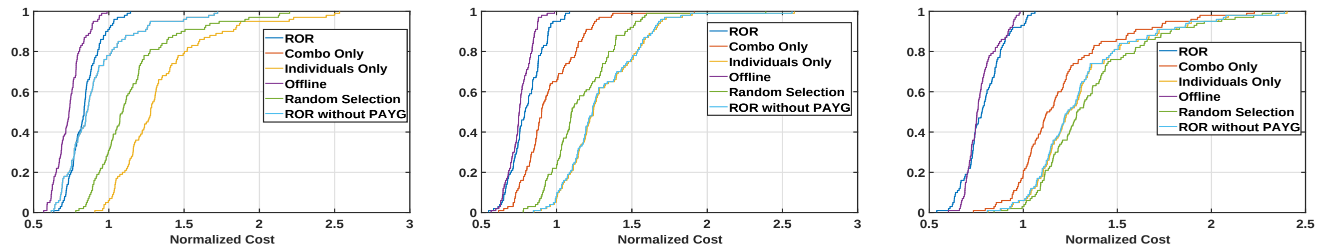
$$\theta_d = \sum_{i:t-T<t(i)\leq t} \lceil \frac{d(i)}{M} \rceil \Delta p, \tag{9}$$

$$\theta_c = \sum_{i:t-T<t(i)\leq t} \lceil \frac{c(i)}{N} \rceil \Delta p. \tag{10}$$

θ_d and θ_c are the accumulated costs of the data volume and computing workload in the last reservation period suppose they were all satisfied by PAYG. If no threshold value of any plan is exceeded by the θ_d and θ_c (e.g., Lines 5, 7 in Algorithm 1 of manuscript, Lines 6, 13 in Algorithm 2 of the manuscript), then the plan (data, computing, or combo) with the closest threshold value will be reserved.

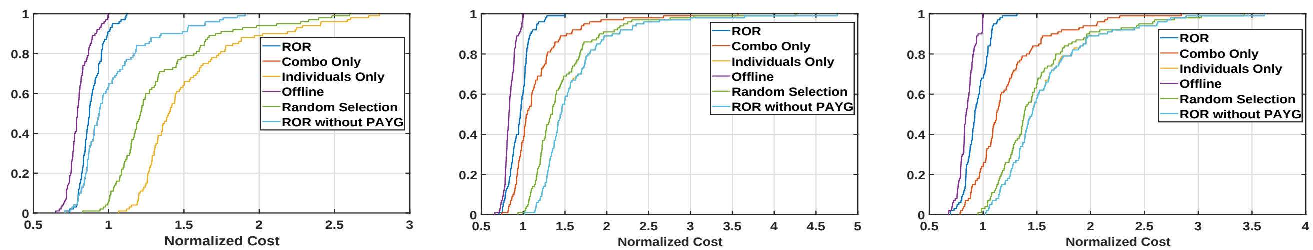
Fig. 6 shows cost performance for the inexpensive, moderate, and expensive combo plan cases with inexpensive data plans. We set $(C_d\Delta p, C_c\Delta p) = (\$15, \$18)$. The plan reservation period in days is set to $T = 7$. The costs of all schemes (except PAYG Only) are normalized with the cost of PAYG Only scheme. In the figure, we see that overall cost performance of our proposed ROR scheme is better than that of all online algorithms in all three cases. As expected, the Combo Only scheme saves money compared with the Individual Only scheme, as we consider $C_o \leq C_d + C_c$. In Fig. 6a, the performance of the ROR scheme is close to the Combo Only scheme, as both schemes take advantage of the inexpensive combo plan to save the cost. As the ROR without PAYG scheme only purchases combo plans when the combo plan is inexpensive, its performance overlaps with the Combo Only scheme in Fig. 6a. When the price of combo plan increases, we see that the ROR scheme saves more money compared with the Combo Only scheme in Fig. 6b. When the price of combo plan is expensive, about 70% of users save money by using the ROR scheme compared with the Combo Only scheme. Besides, the ROR without PAYG scheme approaches the Individuals Only scheme. Besides, the ROR scheme approaches the performance of the Offline scheme even though the Offline scheme has information of all task arrivals in advance.

Fig. 7 shows cost performance for various combo plan cases with expensive data plans. We set $(C_d\Delta p, C_c\Delta p) = (\$21, \$18)$. As the prices of data plan and combo plan increase, cost performance of the Individual Plan, Combo Only, Random Selection, and ROR without PAYG schemes become worse. However, the increase of the data and combo plan prices has no significant impact on both Offline and ROR schemes, which illustrates the robustness of our ROR scheme. Furthermore, from performance of Random Selection and ROR without PAYG, it can be seen that selecting PAYG and plans randomly can result in a very high cost, and purely reserve plans can waste money if there are only a few tasks.



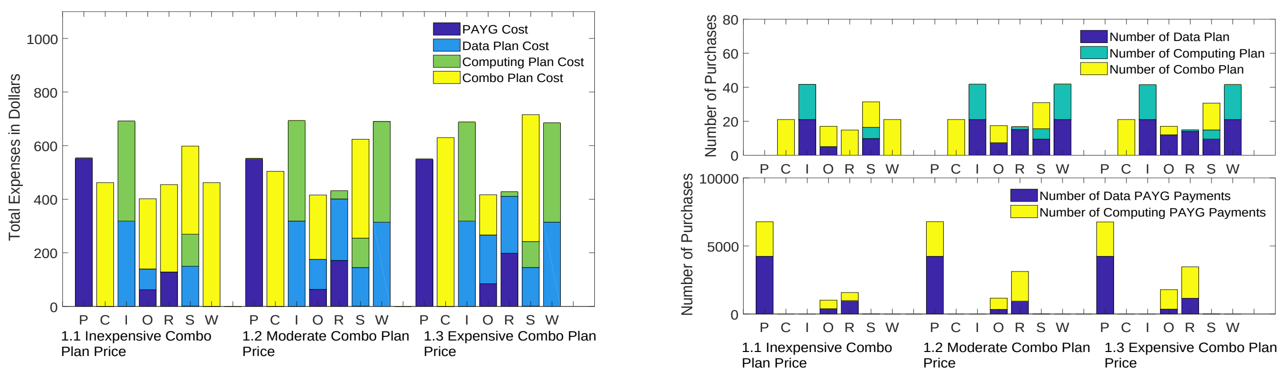
(a) Inexpensive Combo Plan: $C_o\Delta p = \$22$. (b) Moderate Combo Plan: $C_o\Delta p = \$24$. (c) Expensive Combo Plan: $C_o\Delta p = \$30$.

Fig. 6: Cost Performance of Algorithms with Inexpensive Data Plans: $C_d\Delta p = \$15$



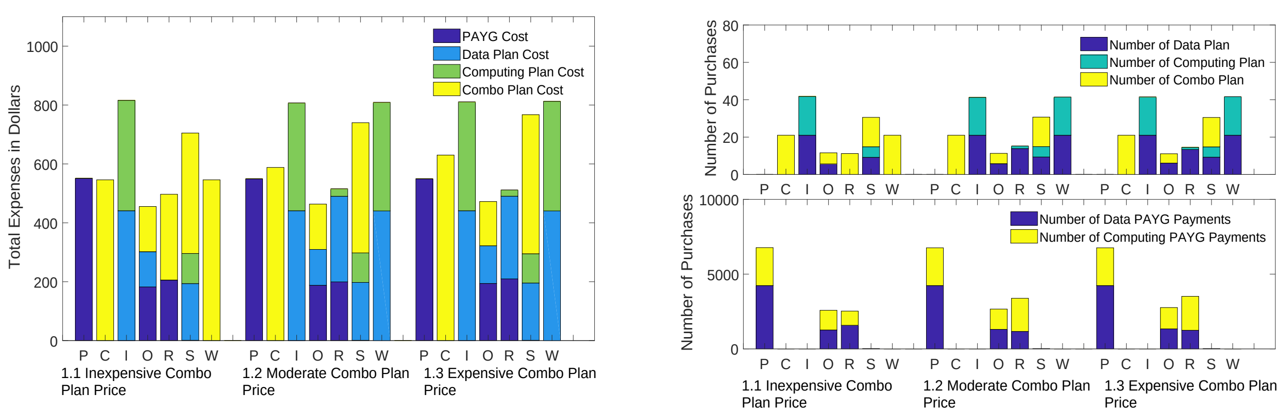
(a) Inexpensive Combo Plan: $C_o\Delta p = \$26$. (b) Moderate Combo Plan: $C_o\Delta p = \$28$. (c) Expensive Combo Plan: $C_o\Delta p = \$30$.

Fig. 7: Cost Performance of Algorithms with Expensive Data Plans: $C_d\Delta p = \$21$



(a) Total Cost Analysis. (b) Plan Reservation Analysis.

Fig. 8: Cost and Plan Reservation Analysis of Online and Offline Algorithms with Inexpensive Data Plan: $C_d\Delta p = \$15$



(a) Total Expenses and Cost Analysis. (b) Plan Reservation Analysis.

Fig. 9: Cost and Plan Reservation Analysis of Online and Offline Algorithms with Expensive Data Plan: $C_d\Delta p = \$21$

D. Cost and Plan Reservation Analysis of Online and Offline Algorithms

In this subsection, we analyze the user's behavior in the plan reservation strategies for all online and offline algorithms with different combo plan prices. Each multimedia task in Table II is considered to have equal arrival probability and the plan reservation period in days is set to $T = 7$. The plan pricing setting in Fig. 8 is the same as in Fig. 6, while the plan pricing setting in Fig. 9 is the same as in Fig. 7. In both Figs. 8b and 9b, the labels P, C, I, O, R, S, W indicate the PAYG Only, Combo Only, Individual Plan, Offline, ROR schemes, Random Selection, and ROR without PAYG schemes respectively.

Fig. 8 shows the cost and plan reservation analysis with inexpensive data plans for various combo plans. When the combo plan is inexpensive, both the Offline and ROR schemes reserve a significant number of combo plans to save money. Compared with the ROR scheme, the Offline scheme reserves data plans to save more money. When the price of combo plan increases, more individual plans and PAYG payments are reserved to replace combo plans in the ROR scheme. This illustrates that the ROR scheme is adaptable to different plan prices. Our proposed ROR scheme reserves less combo plans than the Offline scheme when the combo plan is expensive due to the high random threshold for the combo plan. This is why the total expense of the ROR scheme approaches the expense of the Offline scheme in expensive combo plan case. As the Random Selection scheme still reserves a large portion of combo plans when the combo plan is expensive, its cost is the highest among all schemes.

Fig. 9 shows the cost and plan reservation analysis with the expensive data plans and various combo plan prices. We see that the Individual Plan and ROR without PAYG schemes spends more on data plans than computing plans due to the higher price of data plan. As prices of both combo and data plans increase, the thresholds to reserve data and combo plans become higher in the proposed ROR scheme. Therefore, more PAYG payments are reserved to replace the combo and data plan reservations for cost savings in ROR scheme. A similar plan reservation strategy is also adopted in the Offline scheme.

E. Reservation Period Analysis

Next, we analyze the impact of the reservation period on cost performance of all online and offline algorithms. The reservation period T is set to 7, 14, 30, and 60 days. The price of all plans are set proportional to the length of the reservation period. Each multimedia task in Table II is considered to have equal arrival probability.

Fig. 10 shows cost performance versus various reservation periods with inexpensive data plan. We set $(C_d\Delta p, C_c\Delta p) = (\$[\frac{15T}{7}], \$[\frac{18T}{7}])$. For inexpensive, moderate, and expensive combo plans, $C_o\Delta p$ is set to $\$[\frac{22T}{7}]$, $\$[\frac{24T}{7}]$, and $\$[\frac{30T}{7}]$, respectively. The figure shows that our proposed ROR scheme outperforms other online algorithms in most cases. As the length of the reservation period increases, the cost of all online and offline algorithms increases. This is because the plan reservation is less flexible, and the user is more likely to be in a long-term plan even if there is no multimedia task.

Fig. 11 shows cost performance versus various reservation periods with expensive data plan. We set $(C_d\Delta p, C_c\Delta p) = (\$[\frac{21T}{7}], \$[\frac{18T}{7}])$. For inexpensive, moderate, and expensive combo plans, $C_o\Delta p$ is set to $\$[\frac{26T}{7}]$, $\$[\frac{28T}{7}]$, and $\$[\frac{30T}{7}]$, respectively. As the prices of combo and data plan increase, the costs of the Combo Only and Individual Plan schemes are higher than the PAYG. When $T = 60$, even the proposed ROR scheme costs more than the PAYG. These results suggest that multimedia service providers should avoid lengthy plan reservation periods, since this will result in higher costs for users, and eventually result in more PAYG selections.

VI. Conclusion and Future Work

To fill needs of communication and computing resources, combo plans will emerge for computing-intensive multimedia applications. Due to the bursty and unpredictable nature of multimedia traffic, users do not have the future task information to reserve plans. Therefore, we propose a randomized online reservation (ROR) scheme for users to minimize their cost without considering any future information. We prove that the competitive ratio of our proposed ROR scheme is constant for any multimedia task sequences, and is the minimum among all possible deterministic and randomized online algorithms. By considering all users adopt the ROR scheme due to its cost advantage, we derive a threshold value which will be beneficial for the design of combo plan pricing. Trace-driven simulation results show the cost advantage of ROR scheme and characterize how different combo plan prices influence plan reservation strategies of users.

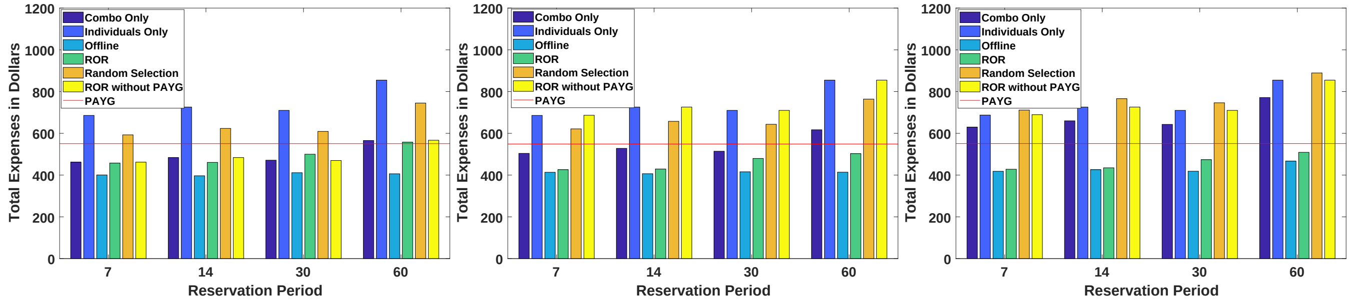
For the future work, there can be extensions of resource reservation problems on scenarios with multiple service providers, multiple type of data plans (high data rate and low data rate) and computing plans (different cores and memories), and multiple types of multimedia tasks (real-time and non real-time).

VII. Appendix

For the support of following proofs, *we define* period PAYG data cost V and period PAYG computing cost W here. The period PAYG data cost V is the sum of PAYG data cost during a period with length T given that no plan is reserved in this period. The period PAYG computing cost W is the sum of PAYG computing cost during a period with length T given that no plan is reserved in this period.

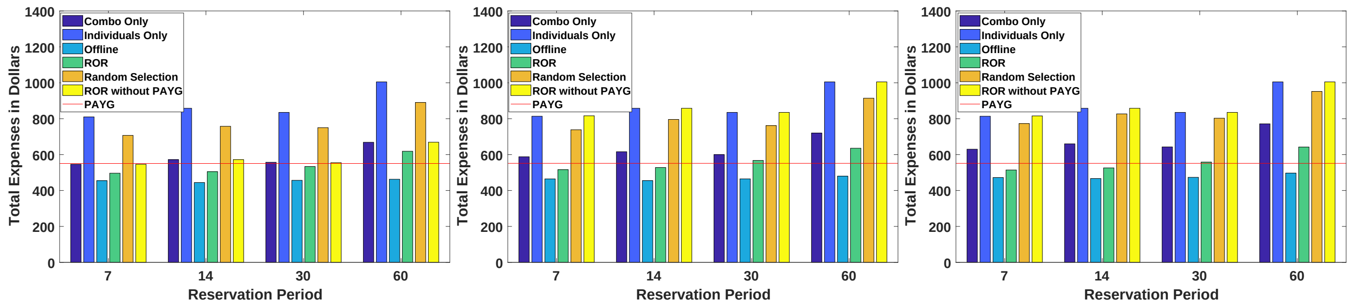
A. Proof of Lemma 1

Let $p_{\gamma_o^{(1)}}$ denote the probability when the data plan is reserved under the condition $\theta_d \geq \gamma_o^{(1)}\Delta p$. We have $p_{\gamma_o^{(1)}} = f_1(\gamma_o^{(1)})$ in (5a). Given that $V = m\Delta p$ and $W = n\Delta p$,



(a) Inexpensive Combo Plan. (b) Moderate Combo Plan. (c) Expensive Combo Plan.

Fig. 10: Reservation Period Analysis of Algorithms with Inexpensive Data Plan.



(a) Inexpensive Combo Plan. (b) Moderate Combo Plan. (c) Expensive Combo Plan.

Fig. 11: Reservation Period Analysis of Online and Offline Algorithms with Expensive Data Plan.

the expected cost of ROR scheme in the period of T can be expressed as

$$\begin{aligned}\mathbb{E}[C_{ROR}(\mathbf{S})] &= \sum_{\gamma_o^{(1)}=1}^{m}(C_o+m+1)\Delta p p_{\gamma_o^{(1)}} + \sum_{\gamma_o^{(1)}\geq m+1} m\Delta p p_{\gamma_o^{(1)}} + n\Delta p \\ &= \frac{m\Delta p}{1-(1-1/C_o)^{C_d}} + n\Delta p \\ &\leq \frac{m\Delta p}{1-(1/e)^{\frac{C_d}{C_o}}} + n\Delta p. \qquad (11)\end{aligned}$$

In the third step, $(1-1/C_o)^{C_o}$ is less than $1/e$ and tends to $1/e$ when C_o is at infinity.

According to (5), (6) and (7), the ROR scheme definitely reserves a plan when $V \geq C_d\Delta p$ or $W \geq C_c\Delta p$ or $V+W \geq C_o\Delta p$. The cost of ROR scheme becomes flat after a plan is reserved, and the worst case competitive ratio of ROR scheme happens before the cost becomes flat. As a result, we consider $m \leq C_d$, $n \leq C_c$, and $m+n \leq C_o$. The optimal offline strategy in the period of T is $C_{OPT}(\mathbf{S}) = (m+n)\Delta p$. As a result, the competitive ratio is

$$\frac{\mathbb{E}[C_{ROR}(\mathbf{S})]}{C_{OPT}(\mathbf{S})} \leq \frac{\frac{m\Delta p}{1-(1/e)^{\frac{C_d}{C_o}}} + n\Delta p}{(m+n)\Delta p} \leq \frac{1}{1-(1/e)^{\frac{C_d}{C_o}}}. \qquad (12)$$

B. Proof of Lemma 3

We prove this lemma by considering three cases: 1) Only the data plan is reserved according to $\theta_d \geq \gamma_d\Delta p$; 2) Only the computing plan is reserved according to the $\theta_c \geq \gamma_c\Delta p$; and 3) Both data and computing plans are reserved.

1) Data Plan Reserved: Let p_{γ_d} denote the probability when the data plan is reserved under the condition $\theta_d \geq \gamma_d\Delta p$. We have $p_{\gamma_d} = g(\gamma_d)$ in (6). Given that $V = m\Delta p$ and $W = n\Delta p$, the expected cost of ROR scheme in the period of T can be expressed as

$$\begin{aligned}\mathbb{E}[C_{ROR}(\mathbf{S})] &= \sum_{\gamma_d=1}^{m}(C_d+m-1)\Delta p p_{\gamma_d} + \sum_{\gamma_d\geq m+1} m\Delta p p_{\gamma_d} + n\Delta p \\ &= \frac{m\Delta p}{1-(1-1/C_d)^{[\frac{C_dC_o}{C_d+C_c}]}} + n\Delta p \\ &\leq \frac{m\Delta p}{1-(1/e)^{\frac{C_o}{C_d+C_c}}} + n\Delta p. \qquad (13)\end{aligned}$$

In the third step, we consider that $[\frac{C_dC_o}{C_d+C_c}] \simeq \frac{C_dC_o}{C_d+C_c}$[4]. Given that $(1-1/C_d)^{C_d}$ is less than $1/e$ and tends to $1/e$ when C_d is at infinity, we have $(1-1/C_d)^{\frac{C_dC_o}{C_d+C_c}} \leq (1/e)^{\frac{C_o}{C_d+C_c}}$.

According to (6) and (7), the ROR scheme definitely reserves a data plan when $V \geq [\frac{C_dC_o}{C_d+C_c}]\Delta p$, and reserves a computing plan when $W \geq [\frac{C_cC_o}{C_d+C_c}]\Delta p$. As the worst case competitive ratio of ROR scheme happens before the cost becomes flat, we consider $m \leq [\frac{C_dC_o}{C_d+C_c}]$ and $n \leq [\frac{C_cC_o}{C_d+C_c}]$.

The optimal offline strategy in the period of T is $C_{OPT}(\mathbf{S}) = (m+n)\Delta p$, given $m \leq [\frac{C_dC_o}{C_d+C_c}]$ and $n \leq [\frac{C_cC_o}{C_d+C_c}]$. As a result, the competitive ratio when only the data

[4] As the data plan can cover the unlimited data volume in the following reservation period, its price C_d should be far greater than 1. For example, the unlimited data plan of AT&T is \$70 [9]. As we have $0.5 \leq \frac{C_o}{C_d+C_c} \leq 1$, $\frac{C_dC_o}{C_d+C_c}$ is far greater than 1, and we have $[\frac{C_dC_o}{C_d+C_c}] \simeq \frac{C_dC_o}{C_d+C_c}$.

plan is reserved is

$$\frac{\mathbb{E}[C_{ROR}(\mathbf{S})]}{C_{OPT}(\mathbf{S})} \leq \frac{\frac{m\Delta p}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}+n\Delta p}{(m+n)\Delta p} \leq \frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}. \tag{14}$$

The worst competitive ratio happens when no computing workload is required during the data plan reservation. This competitive ratio is for a single period of T and is also valid for multiple periods of T.

2) Computing Plan Reserved: The competitive ratio can also be derived as $\frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}$ by following the similar procedures in the case above.

3) Data and Computing Plans Reserved: As the competitive ratio is $\frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}$ for each plan reservation, the competitive ratio of reserving both plans is still $\frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}$.

C. Proof of Lemma 4

We prove this lemma by considering three cases of the multimedia task sequence: 1) Only the data volume is required in multimedia task sequence; 2) Only the computing workload is required in the multimedia task sequence; and 3) Both data volume and computing workload are required the multimedia task sequence.

1) Only Data Volume Required: Let $\mathbf{S}$ denote any task arrival sequence with no computing workload in the period of T. Given $V = n\Delta p$, the optimal offline cost under this case is

$$C_{OPT}(\mathbf{S}) = \begin{cases} n\Delta p, n \leq C_d, \\ C_d\Delta p, n > C_d. \end{cases} \tag{15}$$

Let p_i denote the probability to reserve the combo plan when $\theta_d \geq \gamma_o^{(1)}\Delta$. Let A denote the the randomized algorithm. The expected cost of algorithm A when $V = n\Delta p$, $n \leq C_d$, is

$$\mathbb{E}[C_A(\mathbf{S})] = \sum_{\gamma_o^{(1)}=1}^{n} (C_o + n - 1)p_{\gamma_o^{(1)}} + \sum_{\gamma_o^{(1)} \geq n+1} np_{\gamma_o^{(1)}} \tag{16}$$

To minimize the competitive ratio c, we have the following optimization problem:

$$\begin{aligned} \underset{p_{\gamma_o^{(1)}}}{\text{minimize}} \quad & c \\ \text{subject to} \quad & \mathbb{E}(C_A(\mathbf{S})) \leq c \cdot C_{OPT}(C_A(\mathbf{S})), \forall \theta_d \leq C_d\Delta p, \\ & 0 \leq p_{\gamma_o^{(1)}} \leq 1, \qquad 1 \leq \gamma_o^{(1)} \leq C_d, \\ & p_1 + ... + p_{C_d} = 1, \end{aligned} \tag{17}$$

As the optimal offline cost becomes flat after $C_d\Delta p$, the randomized algorithm should also reserve the combo plan before $C_d\Delta p$ to bound the gap. This is why we have the last two constraints in the above optimization problem. To achieve the minimum c, we need to have $\mathbb{E}(C_A(\mathbf{S})) = c \cdot C_{OPT}$, $\forall \theta_d \leq C_d\Delta p$. Otherwise, we can always make c smaller. As such, we can obtain that $f_1(\gamma_o^{(1)}) = p_{\gamma_o^{(1)}} = \left(\frac{C_o-1}{C_o}\right)^{C_d-\gamma_o^{(1)}} \frac{1}{C_o\left(1-(1-1/C_o)^{C_d}\right)}$, and the minimum c is solved as $\frac{1}{1-(1/e)^{\frac{C_d}{C_o}}}$. For each period of T, in the worst case, the optimal competitive ratio is $\frac{1}{1-(1/e)^{\frac{C_d}{C_o}}}$. Therefore, $\frac{1}{1-(1/e)^{\frac{C_d}{C_o}}}$ is also the optimal competitive ratio for multiple periods of T.

2) Only Computing Workload Required: This case is symmetric to the case above. By following the same procedure, the competitive ratio of reserving combo plans when only computing workload is required can be derived as $\frac{1}{1-(1/e)^{\frac{C_c}{C_o}}}$ and we can also obtain the expression of $f_2(\gamma_o^{(2)})$.

3) Both Data Volume and Computing Workload Required: Let $\mathbf{S}$ denote any task arrival sequence in period of T. Given $V = n\Delta p$ and $W = m\Delta p$, the optimal offline cost under this case is

$$C_{OPT}(\mathbf{S}) = \begin{cases} (m+n)\Delta p, n \leq C_d, m \leq C_c, m+n \leq C_o \\ (C_d+m)\Delta p, n > C_d, m \leq C_c, C_d+m \leq C_o \\ (C_c+n)\Delta p, n \leq C_d, m > C_c, C_c+n \leq C_o \\ C_o\Delta p, m+n > C_o. \end{cases} \tag{18}$$

As the randomized algorithm should reserve the combo plan before the optimal offline cost becomes flat, the competitive ratio for the first three cases in (18) can be derived as $\frac{1}{1-(1/e)^{\frac{C_o}{C_o}}}$, $\frac{\frac{m}{1-(1/e)^{\frac{C_d}{C_o}}}+n}{m+n}$, $\frac{\frac{n}{1-(1/e)^{\frac{C_c}{C_o}}}+m}{m+n}$, respectively.

As $\max(\frac{\frac{m}{1-(1/e)^{\frac{C_d}{C_o}}}+n}{m+n}, \frac{\frac{n}{1-(1/e)^{\frac{C_c}{C_o}}}+m}{m+n}) \leq \max(\frac{1}{1-(1/e)^{\frac{C_d}{C_o}}}, \frac{1}{1-(1/e)^{\frac{C_c}{C_o}}})$, the worst case happens when only one type of the resource is required. Therefore, the competitive ratio is $\max(\frac{1}{1-(1/e)^{\frac{C_d}{C_o}}}, \frac{1}{1-(1/e)^{\frac{C_c}{C_o}}})$ and the lemma is proved.

D. Proof of Lemma 5

Let p_{γ_d} denote the probability to reserve the data plan when the typical data cost $\theta_d = \gamma_d\Delta p$, $p_{\gamma_d} = g(\gamma_d)$. Let p_{γ_c} denote the probability to reserve the computing plan when the typical computing cost $\theta_c = \gamma_c\Delta p$, $q_{\gamma_c} = h(\gamma_c)$. Let A denote the the randomized algorithm. The expected cost of algorithm A in this case is (19). To minimize the competitive ratio c, we have the following optimization problem:

$$\begin{aligned} \underset{p_{\gamma_d}, q_{\gamma_c}, a, b}{\text{minimize}} \quad & c \\ \text{subject to} \quad & \mathbb{E}(C_A(\mathbf{S})) \leq c \cdot C_{OPT}(C_A(\mathbf{S})), \forall \mathbf{S}, \\ & 0 \leq p_{\gamma_d} \leq 1, \\ & 1 \leq \gamma_d \leq C_d - a, \\ & 0 \leq q_{\gamma_c} \leq 1, \\ & 1 \leq \gamma_c \leq C_c - b, \\ & p_1 + ... + p_{C_d-a} = 1, \\ & q_1 + ... + q_{C_c-b} = 1, \\ & a + b = C_d + C_c - C_o \end{aligned} \tag{20}$$

$$\mathbb{E}[C_A(\mathbf{S})] = \sum_{\gamma_d=1}^{n} (C_d + \gamma_d - 1)p_{\gamma_d} + \sum_{\gamma_d \geq n+1} np_{\gamma_d} + \sum_{\gamma_c=1}^{m} (C_c + \gamma_c - 1)q_{\gamma_c} + \sum_{\gamma_c \geq m+1} mq_{\gamma_c} \tag{19}$$

As the optimal offline strategy becomes flat when θ_d exceeds $C_d\Delta p$, θ_c exceeds $C_c\Delta p$, and $\theta_d + \theta_c$ exceeds $C_o\Delta p$, we have $\gamma_d < C_d$, $\gamma_c < C_c$, and $\gamma_d + \gamma_c \leq C_o$. We first fix a and b, and optimize p_{γ_d} and q_{γ_c}. We can obtain that $g(\gamma_d) = p_{\gamma_d} = \left(\frac{C_d-1}{C_d}\right)^{C_d-a-\gamma_d} \frac{1}{C_d\left(1-(1-1/C_d)^{C_d-a}\right)}$ and $h(\gamma_c) = q_{\gamma_c} = \left(\frac{C_c-1}{C_c}\right)^{C_c-b-\gamma_c} \frac{1}{C_c\left(1-(1-1/C_c)^{C_c-b}\right)}$. Therefore, We can thus obtain the competitive ratio $c = \frac{\frac{m}{1-(1/e)^{\frac{C_d-a}{C_d}}} + \frac{n}{1-(1/e)^{\frac{C_c-b}{C_c}}}}{m+n}$, when $m \leq C_d, n \leq C_c, m + n \leq C_o$. To further minimize c for all m and n, we need to have $\frac{1}{1-(1/e)^{\frac{C_d-a}{C_d}}} = \frac{1}{1-(1/e)^{\frac{C_c-b}{C_c}}}$. With the constraint $a + b = C_o$, we can obtain that $a = \frac{C_d(C_d+C_c-C_o)}{C_d+C_c}$ and $b = \frac{C_c(C_d+C_c-C_o)}{C_d+C_c}$. Therefore, we have the expression for $g(\gamma_d) = \left(\frac{C_d-1}{C_d}\right)^{[\frac{C_dC_o}{C_d+C_c}]-\gamma_d} \frac{1}{C_d\left(1-(1-1/C_d)^{[\frac{C_dC_o}{C_d+C_c}]}\right)}$ and $h(\gamma_c) = \left(\frac{C_c-1}{C_c}\right)^{[\frac{C_cC_o}{C_d+C_c}]-\gamma_c} \frac{1}{C_c\left(1-(1-1/C_c)^{[\frac{C_cC_o}{C_d+C_c}]}\right)}$. The minimum competitive ratio is thus solved as $\frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}$.

E. Proof of Theorem 2

We consider each case of the ROR scheme. In Case 1 (inexpensive combo plan), as we have $C_o < \min\left(\sqrt{C_d(C_d+C_c)}, \sqrt{C_c(C_d+C_c)}\right)$, the minimum competitive ratio of reserving the combo plan is less than the minimum competitive ratio of reserving individual plans according to Lemma 4 and Lemma 5 in Case 1. Therefore, the minimum competitive ratio in Case 1 is $\max(\frac{1}{1-(1/e)^{\frac{C_d}{C_o}}}, \frac{1}{1-(1/e)^{\frac{C_c}{C_o}}})$. As the competitive ratio of ROR scheme in Case 1 is $\max(\frac{1}{1-(1/e)^{\frac{C_d}{C_o}}}, \frac{1}{1-(1/e)^{\frac{C_c}{C_o}}})$ according to Lemma 1 and Lemma 2, ROR scheme achieves the minimum competitive ratio in Case 1.

As we have $C_o \geq \min\left(\sqrt{C_d(C_d+C_c)}, \sqrt{C_c(C_d+C_c)}\right)$ in Cases 2 and 3, the minimum competitive ratio of reserving the combo plan is greater than or equal to the minimum competitive ratio of reserving individual plans according to Lemmas 4 and 5 in Cases 2 and 3. Therefore, the minimum competitive ratio in Cases 2 and 3 is $\frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}$. As the competitive ratio of the ROR scheme is also $\frac{1}{1-(1/e)^{\frac{C_o}{C_d+C_c}}}$ according to Lemma 3, the ROR scheme achieves the minimum competitive ratio in Cases 2 and 3.

We note that the deterministic algorithms are special cases of the randomized algorithms. A deterministic algorithm is a randomized algorithm whose certain threshold has probability equal to 1. As the ROR scheme achieves the minimum competitive ratio among all the randomized algorithms, it also achieves the minimum competitive ratio among all the deterministic algorithms.

References

[1] Amazon, "Amazon sumerian," available at https://aws.amazon.com/sumerian/?hp=tile&so-exp=below/

[2] AT&T, "Talking cars, bikes and streetlights may someday make your commute easier," Oct. 2017, available at http://about.att.com/newsroom/talking_cars_and_streetlights.html

[3] Verizon, "4G LTE speeds vs. your home network," May 2013, available at https://www.verizonwireless.com/articles/4g-lte-speeds-vs-your-home-network/

[4] Ericsson, "This is 5G," 2017, available at https://www.ericsson.com/assets/content/7324512f2c704e6cbcfd6c4897eab110/this_is_5g_digital_rev_a3.pdf

[5] Amazon, "Amazon EC2 instance types," 2018, available at https://aws.amazon.com/ec2/instance-types/#burst

[6] Pokemon Support, "Pokemon go plus system requirements and compatibility," available at https://support.pokemongo.nianticlabs.com/hc/en-us/articles/221958248-Supported-devices

[7] A. Colon, "IPhone 4S voice and data plans: AT&T vs. Sprint vs. Verizon," Oct. 2011, available at https://www.pcmag.com/article2/0,2817,2394224,00.asp

[8] Amazon, "Amazon EC2 on-demand pricing," 2018, available at https://aws.amazon.com/ec2/pricing/on-demand/

[9] AT&T, "AT&T wireless plans," 2018, available at https://www.att.com/plans/wireless.html

[10] H.-L. Chung, Y.-S. Lin, and J.-L. Hu, "Bundling strategy and product differentiation," *Journal of Economics*, vol. 108, no. 3, pp. 207–229, 2013. [Online]. Available: http://www.jstor.org/stable/42001105

[11] J. Carbajo, D. de Meza, and D. J. Seidmann, "A strategic motivation for commodity bundling," *The Journal of Industrial Economics*, vol. 38, no. 3, pp. 283–298, 1990. [Online]. Available: http://www.jstor.org/stable/2098499

[12] K. Liu, X. Ling, X. Shen, and J. W. Mark, "Performance analysis of prioritized MAC in UWB WPAN with bursty multimedia traffic," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 4, pp. 2462–2473, July 2008.

[13] W. Miao, G. Min, Y. Wu, H. Wang, and J. Hu, "Performance modelling and analysis of software-defined networking under bursty multimedia traffic," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 12, no. 5s, pp. 77:1–77:19, Sep. 2016.

[14] P. Unterbrunner, G. Giannikis, G. Alonso, D. Fauser, and D. Kossmann, "Predictable performance for unpredictable workloads," *Proc. VLDB Endow.*, vol. 2, no. 1, pp. 706–717, Aug. 2009.

[15] A. Borodin and R. El-Yaniv, *Online Computation and Competitive Analysis*. New York, NY, USA: Cambridge University Press, 1998.

[16] C. A. Gizelis and D. D. Vergados, "A survey of pricing schemes in wireless networks," *IEEE Communications Surveys Tutorials*, vol. 13, no. 1, pp. 126–145, First 2011.

[17] N. C. Luong, D. T. Hoang, P. Wang, D. Niyato, D. I. Kim, and Z. Han, "Data collection and wireless communication in Internet of Things (IoT) using economic analysis and pricing models: A survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 4, pp. 2546–2590, Fourthquarter 2016.

[18] B. Heidarpour, Z. Dziong, W. C. Lau, and S. Vakilinia, "Pricing the volume-based data services in cellular wireless markets," *IEEE Transactions on Network and Service Management*, vol. 15, no. 1, pp. 260–274, March 2018.

[19] L. Duan, J. Huang, and B. Shou, "Optimal pricing for local and global WiFi markets," in *2013 Proceedings IEEE INFOCOM*, April 2013, pp. 1088–1096.

[20] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "Smart data pricing: Using economics to manage network congestion," *Commun. ACM*, vol. 58, no. 12, pp. 86–93, Nov. 2015. [Online]. Available: http://doi.acm.org/10.1145/2756543

[21] Q. Ma, Y. F. Liu, and J. Huang, "Time and location aware mobile data pricing," *IEEE Transactions on Mobile Computing*, vol. 15, no. 10, pp. 2599–2613, Oct 2016.

[22] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, vol. 49, no. 3, pp. 237–252, Mar 1998.

[23] X. Wang and H. Schulzrinne, "Incentive-compatible adaptation of internet real-time multimedia," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 2, pp. 417–436, Feb 2005.

[24] H. Wu, L. Liu, X. Zhang, and H. Ma, "Quality of video oriented pricing incentive for mobile video offloading," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, April 2016, pp. 1–9.

[25] W. Ji, P. Frossard, B. Chen, and Y. Chen, "Profit optimization for wireless video broadcasting systems based on polymatroidal analysis," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2310–2327, Dec 2015.

[26] M. Yang, T. Groves, N. Zheng, and P. Cosman, "Iterative pricing-based rate allocation for video streams with fluctuating bandwidth availability," *IEEE Transactions on Multimedia*, vol. 16, no. 7, pp. 1849–1862, Nov 2014.

[27] N. H. Mastronarde and M. van der Schaar, "Automated bidding for media services at the edge of a content delivery network," *IEEE Transactions on Multimedia*, vol. 11, no. 3, pp. 543–555, April 2009.

[28] J. Liao, P. A. Chou, C. Yuan, Y. Hu, and W. Zhu, "Online allocation of communication and computation resources for real-time multimedia services," *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 670–683, April 2013.

[29] N. Rahimi and R. Koosawangsri, "Selection of smart phone plans," in *2013 Proceedings of PICMET '13: Technology Management in the IT-Driven Services (PICMET)*, July 2013, pp. 426–448.

[30] D. Niyato and E. Hossain, "Competitive spectrum sharing in cognitive radio networks: a dynamic game approach," *IEEE Transactions on Wireless Communications*, vol. 7, no. 7, pp. 2651–2660, July 2008.

[31] P. Maille and B. Tuffin, "Pricing the internet with multibid auctions," *IEEE/ACM Transactions on Networking*, vol. 14, no. 5, pp. 992–1004, Oct 2006.

[32] Y. Zhang, S. Bi, and Y. J. A. Zhang, "Joint spectrum reservation and on-demand request for mobile virtual network operators," *IEEE Transactions on Communications*, vol. 66, no. 7, pp. 2966–2977, July 2018.

[33] H. Jin, X. Wang, S. Wu, S. Di, and X. Shi, "Towards optimized fine-grained pricing of IaaS cloud platform," *IEEE Transactions on Cloud Computing*, vol. 3, no. 4, pp. 436–448, Oct 2015.

[34] H. Xu and B. Li, "Dynamic cloud pricing for revenue maximization," *IEEE Transactions on Cloud Computing*, vol. 1, no. 2, pp. 158–171, July 2013.

[35] M. Aazam, E. Huh, M. St-Hilaire, C. Lung, and I. Lambadaris, "Cloud customer's historical record based resource pricing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 7, pp. 1929–1940, July 2016.

[36] B. Sharma, R. K. Thulasiram, P. Thulasiraman, S. K. Garg, and R. Buyya, "Pricing cloud compute commodities: A novel financial economic model," in *2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID 2012)*, May 2012, pp. 451–457.

[37] V. D. Valerio, V. Cardellini, and F. L. Presti, "Optimal pricing and service provisioning strategies in cloud systems: A stackelberg game approach," in *2013 IEEE Sixth International Conference on Cloud Computing*, June 2013, pp. 115–122.

[38] L. Lu, J. Yu, Y. Zhu, and M. Li, "A double auction mechanism to bridge users' task requirements and providers' resources in two-sided cloud markets," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 4, pp. 720–733, 2018.

[39] K. Liu, J. Peng, B. Yu, W. Liu, Z. Huang, and J. Pan, "An instance reservation framework for cost effective services in geo-distributed data centers," *IEEE Transactions on Services Computing*, pp. 1–1, 2018.

[40] S. Chaisiri, B. Lee, and D. Niyato, "Optimization of resource provisioning cost in cloud computing," *IEEE Transactions on Services Computing*, vol. 5, no. 2, pp. 164–177, April 2012.

[41] W. Wang, B. Liang, and B. Li, "Optimal online multi-instance acquisition in IaaS clouds," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 12, pp. 3407–3419, Dec 2015.

[42] S. Zang, W. Bao, P. L. Yeoh, B. Vucetic, and Y. Li, "Paying less for more? combo plans for edge-computing services," in *USENIX Workshop on Hot Topics in Edge Computing (HotEdge 18)*. Boston, MA: USENIX Association, 2018.

[43] W. Hanson and R. K. Martin, "Optimal bundle pricing," *Management Science*, vol. 36, no. 2, pp. 155–174, 1990.

[44] R. Venkatesh and V. Mahajan, "A probabilistic approach to pricing a bundle of products or services," *Journal of Marketing Research*, vol. 30, no. 4, pp. 494–508, 1993.

[45] B. L. Simonin and J. A. Ruth, "Bundling as a strategy for new product introduction: Effects on consumers' reservation prices for the bundle, the new product, and its tie-in," *Journal of Business Research*, vol. 33, no. 3, pp. 219 – 230, 1995.

[46] B. A. Harlam, A. Krishna, D. R. Lehmann, and C. Mela, "Impact of bundle type, price framing and familiarity on purchase intention for the bundle," *Journal of Business Research*, vol. 33, no. 1, pp. 57 – 66, 1995.

[47] Wikipedia, "Integer programming," available at https://en.wikipedia.org/wiki/Integer_programming

[48] A. R. Karlin, C. Kenyon, and D. Randall, "Dynamic TCP acknowledgement and other stories about e/(e-1)," in *Proceedings of the Thirty-third Annual ACM Symposium on Theory of Computing*. New York, NY, USA: ACM, 2001, pp. 502–509.

[49] M. Ficek, "Crawdad dataset ctu/personal (v. 2012-03-15)," Czech Technical University, Tech. Rep., 2012, available at https://crawdad.org/ctu/personal/20120315

[50] MiniMatters, "The best video length for different videos on YouTube," Aug. 2017, available at https://www.minimatters.com/youtube-best-video-length/

[51] Aimetis, "Improve video analytics performance," available at http://cdn.aimetis.com/public/Library/Video\%20Analytic\%20Performance_Aimetis\%20White\%20Paper.pdf

[52] C. Gallup, "How much data does YouTube use?" Nov. 2017, available at https://www.whistleout.com.au/MobilePhones/Guides/How-Much-Data-Does-YouTube-Use

[53] B. Carson, "The overweight web: Average web page size is up 15% in 2014," Dec. 2014, available at https://gigaom.com/2014/12/29/the-overweight-web-average-web-page-size-is-up-15-in-2014/

[54] Kaggle, "Kaggle public datasets," available at https://www.kaggle.com/datasets

[55] X. Zhang, Y. Wang, and W. Shi, "pCAMP: Performance comparison of machine learning packages on the edges," in *USENIX Workshop on Hot Topics in Edge Computing (HotEdge 18)*. Boston, MA: USENIX Association, 2018.

[56] M. Schneider, J. Rambach, and D. Stricker, "Augmented reality based on edge computing using the example of remote live support," in *2017 IEEE International Conference on Industrial Technology (ICIT)*, March 2017, pp. 1277–1282.

[57] N. Chen, Y. Chen, Y. You, H. Ling, P. Liang, and R. Zimmermann, "Dynamic urban surveillance video stream processing using fog computing," in *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)*, April 2016, pp. 105–112.

Shizhe Zang received the B.E. degree (university medal) in electrical engineering from the University of Sydney, Australia, in 2015. He is currently working toward the Ph.D. degree in Centre of Excellence for IoT and Telecommunications, University of Sydney, Australia. His current research interest includes wireless communication and computer networking, with emphasis on Internet of Things, mobile edge computing, and machine learning.

Wei Bao (S'10–M'16) received the B.E. degree in communications engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2009, the M.A.Sc. degree in electrical and computer engineering from The University of British Columbia, Vancouver, Canada, in 2011, and the Ph.D. degree in electrical and computer engineering from the University of Toronto, Toronto, Canada, in 2016. He is currently a Lecturer with the School of Computer Science, The University of Sydney, Sydney, Australia. His research covers the area of network science, with particular emphasis on edge computing, Internet of Things, and mobile networks. He received the Best Paper Awards from the ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems in 2013 and the IEEE International Symposium on Network Computing and Applications in 2016.

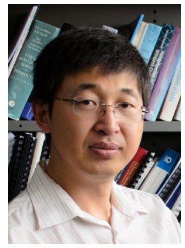
Phee Lep Yeoh (S'08-M'12) received the B.E. degree with University Medal from the University of Sydney, Australia, in 2004, and the Ph.D. degree from the University of Sydney, Australia, in 2012. From 2008 to 2012, he was with the Telecommunications Laboratory at the University of Sydney and the Wireless and Networking Technologies Laboratory at the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia. From 2012 to 2016, he was with the Department of Electrical and Electronic Engineering at the University of Melbourne, Australia. In 2016, he joined the School of Electrical and Information Engineering at the University of Sydney, Australia.

Dr Yeoh is a recipient of the 2017 Alexander von Humboldt Research Fellowship for Experienced Researchers and the 2014 Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA). He has served as TPC chair for the 2016 Australian Communications Theory Workshop (AusCTW) and TPC member for IEEE GLOBECOM, ICC, and VTC conferences. He has received best paper awards at IEEE ICC 2014 and IEEE VTC-Spring 2013, and the best student paper award at AusCTW 2013. His current research interests include wireless security, ultra-reliable and low-latency communications (URLLC), ultra-dense networks, and multiscale molecular communications.

Branka Vucetic is an ARC Laureate Fellow and Director of the Centre of Excellence for IoT and Telecommunications at the University of Sydney.

Her current research work is in wireless networks and the Internet of Things. In the area of wireless networks, she works on communication system design for millimetre wave frequency bands. In the area of the Internet of Things, Vucetic works on providing wireless connectivity for mission critical applications.

Branka Vucetic is a Fellow of IEEE, the Australian Academy of Technological Sciences and Engineering and the Australian Academy of Science.



Yonghui Li (M'04-SM'09-F'19) received his PhD degree in November 2002 from Beijing University of Aeronautics and Astronautics. From 1999 – 2003, he was affiliated with Linkair Communication Inc, where he held a position of project manager with responsibility for the design of physical layer solutions for the LAS-CDMA system. Since 2003, he has been with the Centre of Excellence in Telecommunications, the University of Sydney, Australia. He is now a Professor in School of Electrical and Information Engineering, University of Sydney. He is the recipient of the Australian Queen Elizabeth II Fellowship in 2008 and the Australian Future Fellowship in 2012.

His current research interests are in the area of wireless communications, with a particular focus on MIMO, millimeter wave communications, machine to machine communications, coding techniques and cooperative communications. He holds a number of patents granted and pending in these fields. He is now an editor for IEEE transactions on communications and IEEE transactions on vehicular technology. He also served as a guest editor for several special issues of IEEE journals, such as IEEE JSAC special issue on Millimeter Wave Communications. He received the best paper awards from IEEE International Conference on Communications (ICC) 2014, IEEE PIMRC 2017 and IEEE Wireless Days Conferences (WD) 2014. He is a Fellow of IEEE.