Ultra-Reliable Cloud Mobile Computing with Service Composition and Superposition Coding

Seyyed Mohammadreza Azimi and Osvaldo Simeone
CWCSPR, ECE Dept.
New Jersey Institute of Technology
Newark, NJ
Email: {sa677, osvaldo.simeone}@njit.edu

Onur Sahin
InterDigital
London, UK
Email: onur.sahin@interdigital.com

Petar Popovski
Aalborg University
Aalborg, Denmark
Email: petarp@es.aau.dk

***Abstract*—An emerging requirement for 5G systems is the ability to provide wireless ultra-reliable communication (URC) services with close-to-full availability for cloud-based applications. Among such applications, a prominent role is expected to be played by mobile cloud computing (MCC), that is, by the offloading of computationally intensive tasks from mobile devices to the cloud. MCC allows battery-limited devices to run sophisticated applications, such as for gaming or for the "tactile" internet. This paper proposes to apply the framework of reliable service composition to the problem of optimal task offloading in MCC over fading channels, with the aim of providing layered, or composable, services at differentiated reliability levels. Inter-layer optimization problems, encompassing offloading decisions and communication resources, are formulated and addressed by means of successive convex approximation methods. The numerical results demonstrate the energy savings that can be obtained by a joint allocation of computing and communication resources, as well as the advantages of layered coding at the physical layer and the impact of channel conditions on the offloading decisions.**

***Index Terms*—Ultra-reliable communications, 5G, mobile cloud computing, layered coding, call graph, application offloading.**

I. Introduction

An emerging requirement for 5G systems is the ability to provide wireless ultra-reliable communication (URC) services with close-to-full availability for cloud-based applications (see, e.g., [1]). Among such applications, a prominent role is expected to be played by mobile cloud computing (MCC), that is, by the offloading of computionally intensive tasks from mobile devices to the cloud [2]. MCC allows battery-limited devices to run sophisticated applications, such as for video processing, object recognition, gaming, automatic translation and medical monitoring, and can be an enabler of the "tactile" internet [3], [4]. Well-known applications that are based on MCC include Google Voice Search and Apple Siri.

Existing solutions for the optimization of the offloading decisions for MCC generally abstract the contribution of the underlying communication network by assuming reliable links with *fixed* achievable rates (see, e.g., [2], [5] and references therein). More recently, it was recognized that there is an important interplay between the offloading decisions at the application layer and the operation of the underlying communication network, which can provide different trade-offs between rate and energy expenditure at the mobile devices. As a result, the inter-layer optimization of offloading decisions and communication network parameters, such as transmission powers, were studied in [6], [7] and references therein, as well as in [8], [9]. In this line of work, the focus is on the resource allocation of communication and computing functionalities, and a key assumption is the reliability of the communication links at the rates specified by current channel conditions and by the power allocation. Furthermore, the applications to be offloaded can be assumed to be unsplittable as in [6] or splittable into constituent subtasks as in [8], [9].

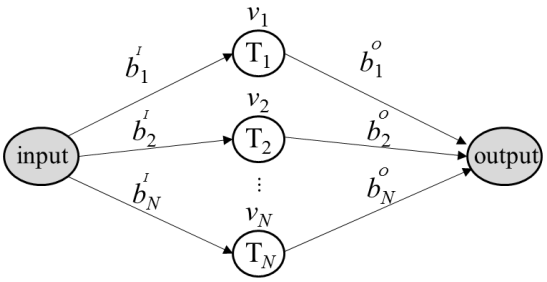

Figure 1: An example of a call graph in the class of map-reduce applications under study. Gray nodes correspond to tasks that must be run at the mobile.

The assumption of reliable communication is in practice too strong when communication takes place over wireless fading channels, especially when latency constraints prevent the use of retransmission protocols to reduce the probability of error. In light of this important motivation, this work aims at studying the problem of joint optimization of offloading decisions and communication system's parameters by accounting for the limited reliability of fading channels with given diversity degrees.

At the application layer, we postulate, as in [10] (see also [4]), that certain applications can be designed so as to ensure *service composition*: the application can be run at different levels of accuracy or quality of experience, with higher levels

The work of S.M. Azimi and O. Simeone was partially supported by the U.S. NSF through grant 1525629.

The work of P. Popovski has been in part supported by the European Research Council (ERC Consolidator Grant Nr. 648382 WILLOW) within the Horizon 2020 Program.

requiring a larger number of CPU cycles. For example, in an object recognition application based on video or image frames, the first service level may correspond to identification of dangerous obstructions, the second to the recognition of landmarks, the third to the search of businesses of possible interest, etc. We observe that the idea of service composition is already implemented in scalable video coding and, more generally, in successive refinement data compression. When coupled with transmission with differentiated reliability levels on the communication network, the approach will be referred to as *reliable service composition* [10].

This paper proposes to apply the framework of reliable service composition to the problem of optimal task offloading in MCC over fading channels, with the aim of providing layered, or composable, services at differentiated reliability levels. We focus on a simple application call graph, exemplified in Fig. 1, which is related to the popular "map-reduce" programming model, in which multiple parallel tasks operate between an input task that prepares the input ("map") and an output task that combines the outputs of the parallel tasks "reduce"). In MCC, each one of the parallel task may be offloaded or not. The application is designed, according to the service composition principle, so that running the first task T_1, along with input and output tasks, ensures the basic level of service, while the execution of successively more tasks $\mathrm{T}_2, \mathrm{T}_3, ...$ allows a higher-accuracy outcome to be obtained. As an example, the parallel tasks may correspond to the processing of different features to extract sufficient statistics in a detection application.

For communication, we consider and compare both time division (TD) transmission and superposition coding (SC), where the latter has been widely studied for the transmission successive refinement compression layers [11]. Inter-layer optimization problems, encompassing offloading decisions and communication resources, are formulated and addressed by means of successive convex approximation methods [12].

The paper is organized as follows. In Section II, first the system model is described and then optimization algorithms over both offloading decisions and communication parameters for TD and SC transmission are provided. Numerical results are provided in Section III and the paper is concluded in Section IV.

II. System Model and Problem Formulation

In this section, we define system model and problem formulation.

A. System Model

We focus on the optimization of offloading decisions and communication parameters for a given mobile user, which can communicate to a base station (BS) via a fading wireless channel. The BS is in turn connected to a cloud processor. As discussed, the application to be run at the mobile is characterized by a set of processing tasks that could be run locally or remotely at the cloud.

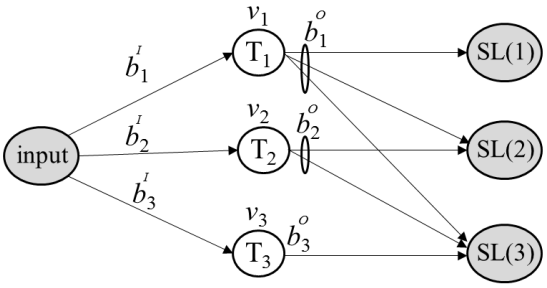

Figure 2: Example of a compound call hypergraph.

Call Graph: A call graph is used to describe the relationship between computing tasks (e.g., [2], [8]). In particular, in this work, we focus on the class of "map-reduce"-type call graphs, illustrated in Fig. 1, which is characterized by input ("map") task, to be run at the mobile; processing tasks T_i, for $i = 1, ...N$, which may be offloaded; and an output task ("reduce") to be run at the mobile. As seen in Fig. 1, the directed edge between input task and task T_i is labeled by the size b_i^I in bits of the data needed for task T_i to run; while the directed edge between each task T_i and the output task is labeled by the number b_i^O of bits produced by the task. Furthermore, each task T_i is labeled by the number of CPU cycles v_i required by its execution. Note that, if task T_i is offloaded, b_i^I bits need to be transmitted in the uplink and b_i^O bits should be received in the downlink direction.

Reliable Service Composition: According to the principle of reliable service composition [10] (see also [4]), the output task can provide services corresponding to different accuracy or quality of experience levels depending on the number of tasks T_i from which it receives data. In particular, it is assumed that the tasks are ordered so that receiving from T_1 allows to obtain a minimal acceptable performance, which is referred to as Service Level 1 (SL(1)); processing the inputs from T_1 and T_2 yields an enhanced performance, denoted as SL(2); and so on for every subset $\{\mathrm{T}_1, ..., \mathrm{T}_i\}$ for $i = 1, ..., N$, which yields a service level SL(i), with full quality obtained when the outputs of all tasks $\{\mathrm{T}_1, ..., \mathrm{T}_N\}$ are available at the output task. Note that extensions in which more general nested subsets correspond to different quality of experience metrics could be accommodated in the framework.

Reliable service composition requires that the ith service level (SL(i)) be obtained with probability r_i, with $r_1 \geq r_2 \geq ... \geq r_N$. For example, in order to ensure ultra-reliability, SL(1) may be provided with reliability $r_1 = 99\%$, while a lower reliability may be sufficient for higher service levels. We define also the parameters $\tilde{r}_i = r_i/r_{i-1}$, with $\tilde{r}_1 = r_1$, where $\tilde{r}_i$ measures the probability that SL(i) is realized given that SL($i-1$) is also attained. This follows from the definition of SLs, which imply that SL(i) can only be realized if SL($i-1$) is.

Offloading: The offloading decisions are described by binary variables I_i. Specifically, the ith task can be either offloaded, which is indicated by setting $I_i = 1$, or computed

locally, indicated as $I_i = 0$. The set of offloaded tasks is represented by $\mathcal{T}$, i.e., $\mathcal{T} = \{i \in \{1, ..., N\} : I_i = 1\}$. If task T_i is offloaded, a transmission power P_i^I is allocated to send b_i^I bits in the uplink, while P_i^O is the allocated transmission power to send b_i^O bits in the downlink direction. Rayleigh fading is used to model the communication channel between user and BS, with a diversity order of d in both uplink and downlink. For the sake of simplicity and concreteness, selection diversity is utilized to exploit the diversity. It is assumed that mobile has no knowledge about the channel while the BS has full knowledge. Channels in the uplink and downlink direction are independent of each other. Furthermore, we let f^M and f^C be mobile and cloud computing frequencies, respectively, in CPU cycles per second. We also denote as P_M the power needed to compute locally at the mobile device. Finally, the application latency constraint states that maximum allowed delay, including the time need for communication and computing, is L_{max} second.

Compound Hypergraph: To simplify the interpretation of the reliable service composition requirements, we now introduce an alternative graphical representation that we refer to as compound hypergraph. While this is not necessary for what follows, we believe it to be a useful way to visualize the reliability requirements. To elaborate, given a call graph as in Fig. 1, a compound call hypergraph can be constructed in order to represent the reliable service composition requirements as follows:

- Set input node and edges between input task and tasks T_i as in original call graph;
- Replicate the output task N times, the first corresponding to SL(1), the second to SL(2) and so on. We refer to each output node by the corresponding service level;
- Connect task T_i to the output nodes SL(j) with $j = i, ..., N$ via a directed hyperedge with head in T_i and tail given by the set {SL(i),SL($i+1$),...,SL(N)}. The hyperedge is labeled by the size of the output of task T_i, namely b_i^O.

Fig. 2 is an example of a compound call hypergraph. The hypergraph formalism is useful to capture the fact that, if task T_i is offloaded, b_i^O bits need to be received for all the connected SL output tasks.

Time-division vs. superposition coding Two transmission modes are considered for offloading, namely:

- Time-division (TD) transmission: Input bits b_i^I for $i \in \mathcal{T}$ on the uplink and output bits b_i^O for $i \in \mathcal{T}$ on the downlink are transmitted in separate time slots, each of duration length L_i^I and L_i^O, respectively. More in detail, in a time slot of duration L^I, the bits for all the offloaded tasks are encoded into different codewords of the same length that are summed, i.e., superimposed, for transmission in the uplink. The same is done for the downlink in a time-slot of duration L^O.
- Superposition coding (SC): Input bits b_i^I for $i \in \mathcal{T}$ on the uplink and output bits b_i^O for $i \in \mathcal{T}$ on the downlink are transmitted using superposition coding in two separate time-slots of length L^I and L^O, respectively. The BS in the uplink and the mobile in the downlink decode in lexicographical order starting from the bits corresponding to the lower index i by means successive interference cancellation.

B. Design Problem Formulation

We focus on the problem of minimizing the energy consumption at the mobile subject to the mentioned maximum latency constraint and reliability constraints, as well as power constraints at the base station. The resulting optimization problem is stated as:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{N} \left(I_i P_i^I L_i^I + \frac{(1-I_i)v_i}{f^M} P_M \right) \\ \text{subject to} & \sum_{i=1}^{N} \left(I_i (L_i^I + L_i^O) + \frac{I_i v_i}{f^C} + \frac{(1-I_i)v_i}{f^M} \right) \leq L_{\max} \\ & \rho_i^I(\mathbf{P}^I, L_i^I, \mathbf{I}) \geq \sqrt{\tilde{r}_i} \qquad \text{for } i \in \mathcal{T} \\ & \rho_i^O(\mathbf{P}^O, L_i^O, \mathbf{I}) \geq \sqrt{\tilde{r}_i} \qquad \text{for } i \in \mathcal{T} \\ & P_i^O \leq P_{max}^{DL} \qquad \text{for } i \in \mathcal{T} \\ & P_i^I \geq 0, P_i^O \geq 0, L_i^I \geq 0, L_i^O \geq 0 \\ & I_i \in \{0, 1\} \\ \text{variables} & \{I_i, P_i^I, P_i^O, L_i^I, L_i^O\} \end{array} \tag{1}$$

where $\mathbf{P}^I = (P_1^I, ..., P_N^I)$ and $\mathbf{P}^O = (P_1^O, ..., P_N^O)$ are the vectors of transmission powers in uplink and downlink directions, respectively; $\mathbf{I} = (I_1, ..., I_N)$ is the vector collecting all the offloading decisions; L_i^I and L_i^O are the uplink and downlink transmission times, respectively, as introduced above. Note that problem (1) applies to both TD and SC transmissions, with the only caveat that, with SC, we have the additional constraint that $L_i^I = L^I$ and $L_i^O = L^O$ for all $i = 1, ..., N$. The functions $\rho_i^O(\mathbf{P}^O, L_i^O, \mathbf{I})$ and $\rho_i^I(\mathbf{P}^I, L_i^I, \mathbf{I})$ represent the probabilities of success for the transmissions in the uplink and in the downlink, respectively, for the offloading of task T_i. These functions depend on whether the transmission takes place via TD or SC, as further discussed below.

The objective function in (1) is the sum of transmission energy at the user, which accounts for the offloaded tasks, and of the local computing energy, for tasks that are run locally. In a similar manner, the first constraint accounts for the latency of both transmission and computing. The following reliability constraints in (1) are justified by the fact that the reliability of SL(i), conditioned on SL($i-1$), is given by the product of the probabilities of success for uplink and downlink transmissions. This is because task $\mathrm{T}_i \in \mathcal{T}$ is successfully offloaded as long as both uplink and downlink transmissions are successful. The problem formulation in (1) is obtained by imposing equal reliability requirements on uplink and downlink. A problem formulation with a more general balancing could be easily defined, but is not further considered here. Finally, the fourth constraint imposes a power limit on the transmission of the BS, due to the power-limited, rather than energy-limited, nature of BS transmission.

Problem (1) is a mixed integer program. To solve this problem, we perform an exhaustive search over the binary variable I_i and adopt the successive convex approximation method of [12] to optimize over the remaining variables

namely $\{P_i^I, P_i^O, L_i^I, L_i^O\}$ for fixed offloading variables. This method is invoked since, as further detailed below, for fixed offloading variables, problem (1) is not convex. For instance, the objective function of problem (1) is a non-convex bilinear function in the optimization variables (P_i^I, L_i^I).

We now specialize problem (1) to TD and SC transmission.

1) Time Division Transmission: For TD transmission, using outage capacity arguments, the probability of a successful transmission for the uplink can be written as (see Appendix):

$$\rho_i^I(P_i^I, L_i^I, I_i)=\left(1-\left(1-\exp\left(-\frac{2^{\frac{b_i^I}{L_i^I W^I}}-1}{\gamma^I P_i^I}\right)\right)^d\right) \tag{2}$$

and analogously for the downlink by substituting the superscript "O" for "I". In (2), γ^I stands for average signal-to-noise ratio (SNR) of the uplink channel for a unitary transmit power, i.e., for $P_i^I = 1$. We define the downlink average SNR γ^O in a similar way. The original problem (1) can be seen to be non-convex due to the bilinearity of the objective. Using the successive convex approximation method in [12], the original problem (1) is solved as outlined in Table I by means of an iterative procedure in which the current iterate is denoted as $s^t = \{p_i^I, l_i^O, l_i^I, l_i^O\}$, where t is the iteration index. At each iteration, the following strictly convex problem is solved:

$$\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^N \left(I_i\Big(p_i^I(L_i^I - l_i^I) + \tfrac{\tau_i^I}{2}\|L_i^I - l_i^I\|^2 \right. \\
& \left. + l_i^I(P_i^I - p_i^I) + \tfrac{\tau_i^P}{2}\|P_i^I - p_i^I\|^2\Big) + \tfrac{(1-I_i)v_i}{f^M} P_M \right) \\
\text{subject to} \quad & \sum_{i=1}^N \left(I_i(L_i^I + L_i^O) + \tfrac{I_i v_i}{f^C} + \tfrac{(1-I_i)v_i}{f^M} \right) \le L_{\max} \\
& \tfrac{2^{\frac{b_i^I}{L_i^I W^I}}-1}{\gamma^I P_i^I} + \ln(1-(1-\sqrt{\tilde{r}_i})^{\frac{1}{d}}) \le 0 \quad \text{for } i \in \mathcal{T} \\
& \tfrac{2^{\frac{b_i^O}{L_i^O W^O}}-1}{\gamma^O P_i^O} + \ln(1-(1-\sqrt{\tilde{r}_i})^{\frac{1}{d}}) \le 0 \quad \text{for } i \in \mathcal{T} \\
& P_i^O \le P_{max}^{DL} \qquad \text{for } i \in \mathcal{T} \\
& P_i^I \ge 0, P_i^O \ge 0, L_i^I \ge 0, L_i^O \ge 0 \\
\text{variables} \quad & \{P_i^I, P_i^O, L_i^I, L_i^O\}.
\end{aligned} \tag{3}$$

Note that all the constraints in the problem above are convex. Also, the second and third constraints are obtained from simple algebraic manipulations from the corresponding constraints in (1). In Table I, the step sizes are updated as $\lambda^{t+1} = \lambda^t(1-\epsilon\lambda^t)$ for $t \ge 0$ with $\lambda^0 \in (0, 1]$ and $\epsilon^0 \in (0, 1)$. The algorithm in Table I is repeated until convergence for every value of I_i. The minimum value of the objective over all possible choices of I_i is taken as the final solution.

2) Superposition Coding Transmission: With SC, the probability of success for uplink transmission can be written as (see Appendix):

$$\rho_i^I(P_i^I, L^I, I_i)=1-\left(1-\exp\left(-\frac{2^{\frac{b_i^I}{L^I W^I}}-1}{\gamma^I P_i^I - \left(2^{\frac{b_i^I}{L^I W^I}}-1\right)\sum_{j=i+1}^N \gamma^I P_j^I}\right)\right)^d \tag{4}$$

and analogously for the downlink. Note that here the transmission periods L^I and L^O do not depend on the task i as explained above. As for TD, the reliability constraint can be expressed as

$$\frac{2^{\frac{b_i^I}{L^I W^I}}-1}{\gamma^I P_i^I} - \frac{1}{\gamma^I \sum_{j=i+1}^N P_j^I - \left(\ln(1-(1-\sqrt{\tilde{r}_i})^{\frac{1}{d}})\right)^{-1}} \le 0 \tag{5}$$

for the uplink and analogously for the downlink. Unlike TD, these constraints are non-convex. However, they can be written as the difference of two convex functions, which may be dealt with as explained in [12] in the successive convex approximation method by linearizing the negative term. This yields the approximate reliability function:

$$\begin{aligned}
&\tfrac{2^{\frac{b_i^I}{L^I W^I}}-1}{\gamma^I P_i^I} - \tfrac{1}{\gamma^I \sum_{j=i+1}^N p_j^I - \left(\ln(1-(1-\sqrt{\tilde{r}_i})^{\frac{1}{d}})\right)^{-1}} + \\
&\tfrac{\gamma^I}{\left(\sum_{j=i+1}^N \gamma^I p_j^I - \left(\ln(1-(1-\sqrt{\tilde{r}_i})^{\frac{1}{d}})\right)^{-1}\right)^2}\Big(\sum_{j=i+1}^N (P_j^I - p_j^I)\Big) \le 0
\end{aligned} \tag{6}$$

where $\{p_i^I\}$ represents the previous iterate. Following Table I, the problem to be solved at each iteration is then (3), with $L_i^I = L^I$ and $L_i^O = L^O$ as well as with the constraint above, and the corresponding downlink constraint, in lieu of the third and fourth constraints in (3).

III. Numerical Results

In this section, we provide some numerical examples based on the analysis developed in the previous sections. We set $P_M = 0.4$ Watts; $f_M = 10^9$ CPU cycles/s (e.g., Apple iPhone 6 processor has maximum clock rate of 1.4 GHz); $f_C = 10^{10}$ CPU cycles/s (e.g., AMD FX-9590 has a clock rate of 5 Ghz); bandwidth $W^I = 1MHz$ and $W^O = 1MHz$; and SNR levels $\gamma^I = \gamma^O = 0$ dB.

We start by considering just the basic service level, namely SL(1), in order to simplify the interpretation of the results and to gain insight into the role of the diversity level d on the offloading decisions. The reliability of SL(1) is set to $r_1 = 0.99$. Note that, given the presence of only one offloadable task, TD and SC yield the same performance.

Table I: Successive convex approximation algorithm for TD

Initialization: Set $t = 0$, s^0=$\{p_i^I, p_i^O, l_i^I, l_o^I\}$ feasible $\lambda^0 \in (0, 1]$, $\epsilon^0 \in (0, 1)$
Step 1) If s^t satisfies a termination criterion: STOP
Step 2) Compute $\hat{s}(s^t)$ as the solution of (3).
Step 3) Set $s^{t+1} = s^t + \lambda^t(\hat{s}(s^t) - s^t)$.
(S.4) $t \leftarrow t + 1$ and go to step 1.

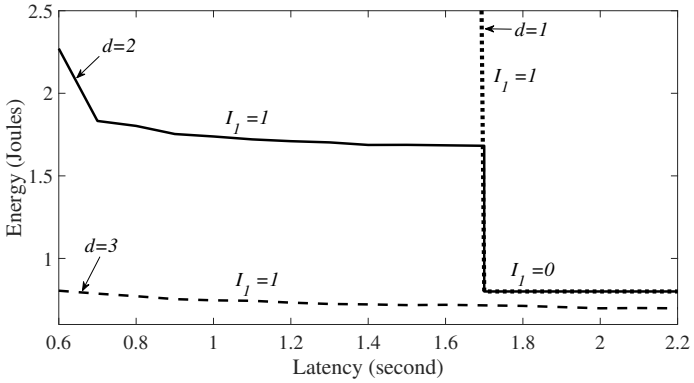

Figure 3: Mobile energy expenditure versus latency constraint for a single service level ($v = 2 \times 10^9$ CPU cycles and $b_1^I = b_1^O = 1.4 \times 10^5$ bits).

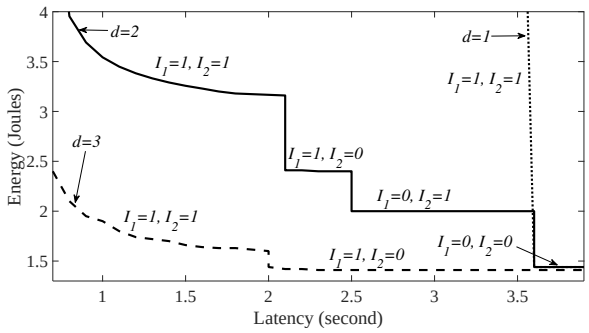

Figure 4: Mobile energy expenditure versus latency constraint for a single service level ($v_1 = 2 \times 10^9$ CPU cycles, $v_2 = 1.6 \times 10^9$ CPU cycles, $b_1^I = b_1^O = 1.4 \times 10^5$ bits and $b_2^I = b_2^O = 2.8 \times 10^5$ bits).

Fig. 3 presents the mobile energy versus latency constraint L_{max} for different diversity orders. If the tolerable latency is low, then it is necessary to offload the task to the cloud since local computing here takes around 1.7 seconds. For diversity $d = 1$, the energy required to offload is outside the range shown in the Fig. 1. An increase of diversity order provides more reliable communication between mobile and cloud, and therefore the offload of the task can be performed with a lower energy expenditure. In particular, if $d = 3$, then it is optimal to offload the task even for latencies larger than 1.7 seconds. We emphasize that the discontinuity in the curves is due to changes in the optimal offloading decisions.

We now consider reliable service composition with two service levels, namely $N = 2$, accounting for both TD and SC transmission modes. The reliability for the second level is set to $\tilde{r}_2 = 0.9$ and the first is still $r_1 = 0.99$. The corresponding mobile energy versus latency trade-offs are shown in Fig. 4 and Fig. 5, respectively. Note that here computing both tasks locally requires a latency of approximately 3.6 seconds. Considering first TD transmission, we observe from Fig. 4 that achieving low latency requires tolerating a high energy cost by offloading both tasks. When increasing the latency, the mobile has incentive to first offload the task with higher computation cost, here the first task, while the second task is run locally due to the lower energy consumption. For $d = 2$ and higher latencies, when the first task can be run locally, it becomes optimal to offload only the second task; while, when the latency is large enough, both tasks should be run locally. With a larger latency, instead, the solution $(I_1 = 1, I_2 = 0)$ turns out to be optimal over a larger range of latencies.

Comparing TD with SC, by observing Fig. 5, we note that SC enables a drastic energy reduction for offloading and hence makes the decision to offload both tasks optimal for all latencies up to 3.6 seconds when $d = 2$, and for the entire range of considered latencies when $d = 3$.

IV. Concluding Remarks

In this paper, the mobile energy versus latency tradeoff was explored for mobile cloud computing applications over fading channels by accounting for the principle of reliable service level composition at the application layer. The aim of this approach is providing layered, or composable, services at differentiated reliability levels. Inter-layer optimization problems, encompassing offloading decisions and communication resources, were formulated and addressed by means of successive convex approximation methods. The numerical results demonstrated the energy savings that may be obtained by a joint allocation of computing and communication resources, as well as the advantages of superposition coding at the physical layer and the impact of channel conditions on the offloading decisions.

Appendix: Calculation of the Probabilities of Success

For TD transmission, assuming diversity $d = 1$ and Rayleigh fading channel gain G^I, the probability of success

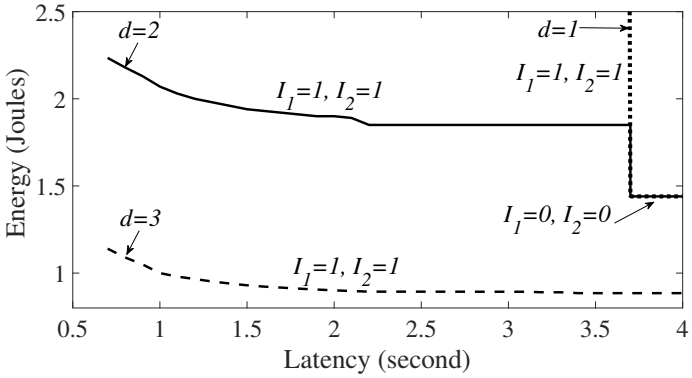

Figure 5: Mobile energy expenditure versus latency constraint for a single service level ($v_1 = 2 \times 10^9$ CPU cycles, $v_2 = 1.6 \times 10^9$ CPU cycles, $b_1^I = b_1^O = 1.4 \times 10^5$ bits and $b_2^I = b_2^O = 2.8 \times 10^5$ bits).

is the complement of the outage probability, namely

$$\Pr\left[L_i^I W^I \log(1+G^I P_i^I \gamma^I) \geq b_i^I\right] = \Pr\left[G^I \geq \frac{2^{\frac{b_i^I}{L_i^I W^I}}-1}{\gamma^I P_i^I}\right] = \exp\left(-\frac{2^{\frac{b_i^I}{L_i^I W^I}}-1}{\gamma^I P_i^I}\right). \quad (7)$$

Generalizing, with a diversity order $d \geq 1$ and selection diversity, we obtain

$$\rho_i^I(P_i^I, L_i^I, I_i) = 1 - \left(1 - \exp\left(-\frac{2^{\frac{b_i^I}{L_i^I W^I}}-1}{\gamma^I P_i^I}\right)\right)^d, \quad (8)$$

and similar calculations apply for $\rho_i^O(P_i^O, L_i^O, I_i)$.

For SC transmission, assuming for $d = 1$, following similar arguments, we have

$$\begin{aligned}
&\rho_i^I(\mathbf{P}^I, L^I, \mathbf{I}) \\
&= \Pr\left[L^I W^I \log_2\left(1+\frac{G^I P_i^I \gamma^I}{1+\gamma^I G^I \sum_{i=i+1}^N P_i^I}\right) \geq b_i^I\right] \\
&= \Pr\left[G^I P_i \gamma^I \geq \left(2^{\frac{b_i^I}{L^I W^I}}-1\right)\left(1+\gamma^I G^I \sum_{i=i+1}^N P_i^I\right)\right] \\
&= \Pr\left[G^I \geq \frac{2^{\frac{b_i^I}{L^I W^I}}-1}{\gamma^I P_i - \left(2^{\frac{b_i^I}{L^I W^I}}-1\right)\gamma^I \sum_{i=i+1}^N P_i^I}\right] \\
&= \exp\left(-\frac{2^{\frac{b_i^I}{L_i^I W^I}}-1}{\gamma^I P_i - \left(2^{\frac{b_i^I}{L^I W^I}}-1\right)\gamma^I \sum_{i=i+1}^N P_i^I}\right),
\end{aligned}$$

where all layers beyond i are treated as noise in the decoding. With selection diversity, we obtain the reliability function stated in the text.

REFERENCES

[1] A. Osseiran and et al, "Scenarios for 5G mobile and wireless communications: the vision of the metis project," *IEEE Commun. Mag.*, vol. 52, no.5, pp. 26–35, May 2014.

[2] S. Abolfazli, Z. Sanaei, E. Ahmed, A. Gani, and R. Buyya, "Cloud-based augmentation for mobile devices: Motivation, taxonomies, and open challenges," *IEEE Commun. Mag.*, vol. 16, no.1, pp. 337–368, Jan. 2014.

[3] G. Fettweis, "The Tactile Internet: Applications and Challenges," *IEEE Veh. Technol. Mag.*, vol. 9, no.1, pp. 64–70, 2014.

[4] A. Aijaz, M. Dohler, A. H. Aghvami, V. Friderikos, and M. Frodigh, "Realizing The Tactile Internet: Haptic Communications over Next Generation 5G Cellular Networks," *Submitted to IEEE Wireless Commun. Mag., arXiv:1510.02826*.

[5] Y.-H. Kao, B. Krishnamachari, M.-R. Ra, and F. Bai, "Hermes: Latency Optimal Task Assignment for Resource-constrained Mobile Computing," in *Computer Communications (INFOCOM), 2015 IEEE Conference on*, 2015, pp. 1894–1902.

[6] S. Barbarossa, S. Sardellitti, and P. D. Lorenzo, "Communicating While Computing: Distributed Mobile Cloud Computing over 5G Heterogeneous Networks," *IEEE Signal Process. Mag.*, vol. 31, no.6, pp. 45–55, 2014.

[7] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile-edge computing," *arXiv preprint arXiv:1412.8416*, 2014.

[8] P. D. Lorenzo, S. Barbarossa, and S. Sardellitti, "Joint Optimization of Radio Resources and Code Partitioning in Mobile Cloud Computing," *Submitted to IEEE T. Mob. Comput., arXiv:1307.3835*.

[9] S. Khalili and O. Simeone, "Inter-Layer Per-Mobile Optimization of Cloud Mobile Computing: A Message-Passing Approach," *Submitted to IEEE T. Com., arXiv:1509.01596*.

[10] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *1st International Conference on 5G for Ubiquitous Connectivity (5GU)*, 2014, pp. 146–151.

[11] C. N. D. Gunduz, A. Goldsmith, and E. Erkip, "Distortion Minimization in Gaussian Layered Broadcast Coding with Successive Refinement," *IEEE Trans. Inf. Theory*, vol. 55, no.11, pp. 5074–5086, Nov. 2009.

[12] G. Scutari, F. Facchinei, L. Lampariello, and P. Song, "Distributed Methods for Constrained Nonconvex Multi-Agent Optimization-Part I: Theory," *Submitted, arXiv:1410.4754*.