Cloud Offloading for Multi-Radio Enabled Mobile Devices

S. Eman Mahmoodi
Dept. of Electrical
and Computer Engineering
Stevens Institute of Technology
Hoboken, New Jersey 07030–5991
Email: smahmood@stevens.edu

K. P. Subbalakshmi
Dept. of Electrical
and Computer Engineering
Stevens Institute of Technology
Hoboken, New Jersey 07030–5991
Email: ksubbala@stevens.edu

Vidya Sagar
Dept. of Electrical
and Computer Engineering
Stevens Institute of Technology
Hoboken, New Jersey 07030–5991
Email: vsagar@stevens.edu

***Abstract*—The advent of 5G networking technologies has increased the expectations from mobile devices, in that, more sophisticated, computationally intense applications are expected to be delivered on the mobile device which are themselves getting smaller and sleeker. This predicates a need for offloading computationally intense parts of the applications to a resource strong cloud. Parallely, in the wireless networking world, the trend has shifted to multi-*radio* (as opposed to multi-channel) enabled communications. In this paper, we provide a comprehensive *computation* offloading solution that uses the multiple radio links available for associated data transfer, optimally. Our contributions include: a comprehensive model for the energy consumption from the perspective of the mobile device; the formulation of the joint optimization problem to minimize the energy consumed as well as allocating the associated data transfer optimally through the available radio links and an iterative algorithm that converges to a locally optimal solution. Simulations on an HTC phone, running a 14-component application and using the Amazon EC2 as the cloud, show that the solution obtained through the iterative algorithm consumes only 3% more energy than the optimal solution (obtained via exhaustive search).**

I. Introduction

The "anywhere, anytime" promise of 5G networking has created a large demand for more sophisticated applications on energy constrained mobile devices [1], leading to a huge increase in computational demand on the end devices [2]. Meanwhile, the promise of 5G networking has also seen a surge in mobile device generated web traffic. In the year 2012 alone mobile web traffic increased by 70% and is expected to grow up to 13 times by 2017. One solution to this problem is to offload computations to the more resource strong cloud infrastructure [3]–[5].

The term cloud offloading can mean either data flow offloading in networking applications [6], [7] or offloading computation intense processes on to the cloud. In this paper, we refer to the latter. Cloud offloading can be classified into three categories: (a) those that always offload to the cloud [8]; (b) "all or nothing offloading" where either the entire application is offloaded to the cloud or executed locally, typically using an energy threshold to decide between offloading and not [9], [10]; and (c) piecewise decisions, where some parts are executed locally while the others are offloaded to the cloud [11]–[14]. The third category offers the most flexibility for trade-offs, and can be done either at the coarse component level [11], [15], [16] or at finer, method [13] or instruction levels [17].

While computation offloading to a resource strong cloud seems like the natural solution to the resource crunch at the mobile device level, it is essential to take into account the associated data transfer that must take place between the components that are executed in the cloud and their counterparts in the mobile device. Given the already increasing demands on the wireless backbone caused by the promise of 5G networking, this means that *computation offloading must be viewed in the context of the already increasing mobile traffic*. Hence it would be prudent to *optimally* use *all* of the radio interfaces (like WiFi, 3G, HSPA, and LTE), as appropriate, that are available in the multi-radio equipped mobile devices of today.

In this paper, we propose a solution that optimally decides which components of an application to offload and which to execute locally, *while simultaneously optimizing the percentage of data (associated with this offloading) to be sent via each radio interface*. Given recent advances in technologies that enable bandwidth aggregation in wireless devices [18], [19] our solution is implementable in practice. To the best of our knowledge, this is the first such solution that approaches cloud offloading for multi-radio enabled devices. Other works that fall under general umbrella of the radio-aware computation offloading include [11], where the best of the available wireless interfaces is chosen (only one of the wireless interfaces) for data transfer, rather than a solution that considers using all of the radio interfaces simultaneously. In [17] a cloud offloading scheduling mechanism is proposed for queue stability, but this work only deals with multi-channel systems, not multi-radio networks. Etime [8] is an "everything on the cloud" offloading strategy, which adapts to the condition of the wireless link, but this work does not consider multiple interfaces.

In this paper, we develop a comprehensive model for the energy consumed by the mobile device, including energy expended in communicating relevant data between the cloud and the device. We set up the computation offloading problem as a joint optimization to minimize the energy consumed on the device while at the same time maximizing the radio resources available to the device, under two constraints: (1) the total run time deadline of the application and (2) the maximum flow rate constraint on the radio resources. Since this optimization problem is non-linear and hence computationally intense, we also propose an iterative algorithm that converges to a local optimum. Simulations show that the proposed iterative algorithm performs very close to the optimal solution for a significant reduction in complexity.

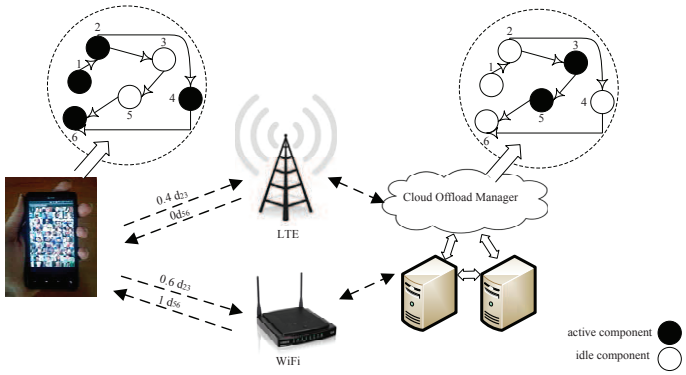

Figure 1. An example of application offloading to the cloud. In this figure, the dots represent components of the application. There are 6 components in this application. Components 1,2, 4, and 6 run on the device, whereas components 3 and 5 are executed in the cloud. Two radio links are available to the mobile device for offloading components to the cloud and the diagram shows the ratio of data that is sent via each radio interface. The terms active (idle) components refers to the components that are executed (or not) in that particular entity, mobile device or the cloud.

II. System Model

Consider a mobile device with K radio interfaces, running computationally intense applications with M components (See Fig. 1, with an example where $K = 2$ and $M = 6$). Any given component may require data from the other components to complete execution. This data dependency is determined based on the corresponding application call graph (dependency matrix). In this example, the optimal offloading strategy stipulates that Components 1, 2, 4 and 6 be executed in the mobile device, and Components 3 and 5 be offloaded to the cloud. In Fig. 1, Component 3 requires d_{23} units of data from Component 2 to complete execution. In this example, 60% ($\nu_{2,1}$ = 0.6) of this data is sent through the Radio Interface 1 (WiFi, say), and 40% ($\nu_{2,2} = 0.4$) through Interface 2 (LTE) to give us the most performance efficient offloading strategy. Once Component 3 and 5 have finished execution, the data needed by Component 6 from Component 5 (d_{56}) must be sent to the mobile device via one of the radio interfaces (for example in Fig. 1 is WiFi). We assume that only one radio interface is used for data reception ($\sum_{k=1}^{K} \gamma_{i,k} = 1$), leaving the optimization of radio resource allocation for the downlink as future work. Also, we assume that the energy consumption and the time required to transfer data within components that are executing in the same entity (whether cloud or mobile) is negligible in comparison to when the data must be transferred between entities. We also assume that the components of the application are executed in a predetermined manner [11], [13]. This is not an unreasonable assumption as the compiler usually predetermines this order.

The parameters needed to set up the optimization are described in Table I. We model the energy consumed by the mobile device in running application component i, as $E_i = E_i^{(\mathrm{m})} + E_i^{(\mathrm{c})} + E_i^{(\mathrm{com})}$, where $E_i^{(\mathrm{m})}$, $E_i^{(\mathrm{c})}$ and $E_i^{(\mathrm{com})}$ are all defined in Table I. The energy consumed to execute component i locally, in the mobile device, is expressed as $E_i^{(\mathrm{m})} = (1 - I_i)P_{\mathrm{ac}}^{(\mathrm{m})}(i)\tau_i^{(\mathrm{m})}$. If the component is executed remotely, then the mobile will only spend the idle power

Table I
Parameter Definitions.

Parameters	Definitions
M	Number of components in the application.
K	Number of radio interfaces in the system model.
$P_{\mathrm{ac}}^{(\mathrm{m})}(i)$	Power consumed by the mobile device when it is actively processing component i.
$P_{\mathrm{id}}^{(\mathrm{m})}$	Power consumed by the mobile in the idle mode.
$P_k^{(\mathrm{Tx})}$ ($P_k^{(\mathrm{Rx})}$)	Transmit (Received) power consumed by the mobile device at radio interface k .
$\tau_i^{(\mathrm{m})}$ ($\tau_i^{(\mathrm{c})}$)	Time to process component i in mobile (cloud).
$\tau_{ij,k}^{(\mathrm{mc})}$ ($\tau_{ij,k}^{(\mathrm{cm})}$)	Time to transfer data required by component j to mobile (cloud) from component i in the cloud (mobile), using radio interface k.
$T_i^{(\mathrm{com})}$	Time to transfer necessary data between the cloud and mobile, to execute component i.
α_{ij}	Component dependency indicator: 1 if component i must be processed before j, 0 otherwise.
I_i	Processing place indicator: 1 if component i is processed on cloud, 0 if processed on mobile.
$\nu_{i,k}$	Percentage of data upload using radio interface k, for execution of component i in the cloud.
$\gamma_{i,k}$	Radio receiving indicator: 1 if transferred data of component i is received at radio k, 0 otherwise.
d_{ij}	Data size required by component j from i.
$R_k^{(\mathrm{d})}$ ($R_k^{(\mathrm{u})}$)	Downlink (Uplink) service rate for radio k.
r_k	Demand rate for radio interface k.
E_i	Total energy consumed by the mobile device to run component i.
$E_i^{(\mathrm{m})}$ ($E_i^{(\mathrm{c})}$)	Energy consumed by the mobile device to run component i in the mobile (cloud).
$E_i^{(\mathrm{com})}$	Energy consumed by the mobile for data transfer of component i between cloud and mobile.

for the duration of this execution. Hence, the energy consumed by the mobile when component i is being remotely executed, is given by $E_i^{(\mathrm{c})} = I_i P_{\mathrm{id}} \tau_i^{(\mathrm{c})}$. $E_i^{(\mathrm{com})}$ comes into play when either the component immediately preceding the component i, or immediately succeeding component i is executed in the other entity. $E_i^{(\mathrm{com})}$ can be written as $E_i^{(\mathrm{com})} = \sum_{j=1}^{M} \sum_{k=1}^{K} (\alpha_{ij}\varepsilon_{ij,k} + \alpha_{ji}\varepsilon_{ji,k})$, where $\varepsilon_{ij,k}$ (or $\varepsilon_{ji,k}$) is the energy consumed in transferring data from component i (j) to component j (i) using radio interface k, when component i (j) is executed immediately before component j (i). They can be written as follows:

$$\varepsilon_{ij,k} = I_i(1 - I_j)\gamma_{j,k}P_{\mathrm{id}}\tau_{ij,k}^{(\mathrm{cm})} + (1 - I_i)I_j\nu_{i,k}P_k^{(\mathrm{Tx})}\tau_{ij,k}^{(\mathrm{mc})}, \quad (1)$$

$$\varepsilon_{ji,k} = I_i(1 - I_j)\nu_{j,k}P_{\mathrm{id}}\tau_{ji,k}^{(\mathrm{mc})} + (1 - I_i)I_j\gamma_{i,k}P_k^{(\mathrm{Rx})}\tau_{ji,k}^{(\mathrm{cm})}. \quad (2)$$

The first terms on the RHS of equations (1) and (2) represent the idle powers consumed when the relevant component is being executed in the cloud, and second terms represent the energy consumed in transmitting or receiving the relevant data. The time needed to transfer data in the downlink communication (cloud to mobile) and uplink communication (mobile to cloud) are given by $\tau_{ij,k}^{(\mathrm{cm})} = \frac{d_{ij}}{R_k^{(\mathrm{d})}}$ and $\tau_{ij,k}^{(\mathrm{mc})} = \frac{d_{ji}}{R_k^{(\mathrm{u})}}$ respectively, where $R_k^{(\mathrm{d})}$ and $R_k^{(\mathrm{u})}$ are the downlink and uplink rates respectively, on radio interface k. d_{ij} is the size of the data that must be transferred from component i to j.

III. Energy Efficient and Radio Resource Optimized Offloading

A. Problem Formulation

In this section, we formulate an optimization problem to minimize the total energy consumed by the mobile user in executing a given application under total execution time constraints. Specifically, we will formulate an optimization problem that will determine which components should be executed where (in the device or cloud) and what percentage of data should be allocated to each radio link for necessary uplink data transfer. This minimization is subject to the following constraints: deadline on the execution time of the application; flow rate control on each radio link used for computation offloading; and the total value of data percentage allocated to the radio interfaces for each offloaded component. The optimization problem is mathematically formulated as

$$\min_{\boldsymbol{\nu},\boldsymbol{I}} E \triangleq \sum_{i=1}^{M} E_i, \tag{3}$$

where $\boldsymbol{I} = [I_1 I_2 ... I_M]$ and $\boldsymbol{\nu}$ is a matrix with entries $\nu_{i,k}$, $\forall i,k$ and I_i's and $\nu_{i,k}$'s are defined in Table I. The constraint on the total application execution time is given by

$$\sum_{i=1}^{M} T_i \leq T_{\text{req}}, \tag{4}$$

where T_{req} is the execution time deadline of the application, and $T_i = T_i^{(\text{m})} + T_i^{(\text{c})} + T_i^{(\text{com})}, \forall i$. $T_i^{(\text{m})}$ represents the time taken for component i to execute in the mobile device, and is given by $T_i^{(\text{m})} = (1 - I_i)\tau_i^{(\text{m})}$. Similarly, $T_i^{(\text{c})} = I_i \tau_i^{(\text{c})}$ is the time taken to execute component i in the cloud. $T_i^{(\text{com})}$ is the time taken to complete the necessary data transfer for execution of component i, and is given by

$$\begin{aligned} T_i^{(\text{com})} = \sum_{j=1}^{M} \sum_{k=1}^{K} \big(& I_i(1 - I_j)(\alpha_{ji}\nu_{j,k}\tau_{ji,k}^{(\text{mc})} + \alpha_{ij}\gamma_{j,k}\tau_{ji,k}^{(\text{cm})}) + \\ & (1 - I_i) I_j (\alpha_{ij}\nu_{i,k}\tau_{ij,k}^{(\text{mc})} + \alpha_{ji}\gamma_{i,k}\tau_{ij,k}^{(\text{cm})}) \big). \end{aligned} \tag{5}$$

This constraint allows us to take into consideration the potential time delays in sending and receiving the data related to each component via radio links ($T_i^{(\text{com})}, \forall i$) and trading it off optimally for energy consumption on the device.

In order for the system to be stable, the transmit data rate on the radio interfaces must be less than the service rate of each radio interface. This is represented by the second constraint:

$$\sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \alpha_{ij}(1 - I_i) I_j \nu_{i,k} r_k < R_k^{(\text{u})}, \forall k. \tag{6}$$

The final constraint ensures that for each component, the total data allocations to the radio interfaces sums up to the total data that needs to be transferred, and is expressed as

$$\sum_{\substack{j=1 \\ j \neq i}}^{M} \alpha_{ij} \sum_{k=1}^{K} \nu_{i,k} \leq 1, \qquad \forall i. \tag{7}$$

B. Proposed Solution

The objective function of the optimization problem is represented in Eq. (3) with the constraints in Eqns (4), (6), and (7). The objective function and the constraint in Eq. (6) involve product terms of two non-negative variables, thereby forming a nonlinear convex function. Thus, the problem can be solved using MIP (Mixed Integer Programming) using Lagrangian multipliers: $\kappa, \zeta_k, \phi_i, \forall i,k$. The Lagrangian, $L = L(\boldsymbol{\nu}, \boldsymbol{I}, \kappa, \boldsymbol{\zeta}, \boldsymbol{\phi})$, is expressed as

$$\begin{aligned} L = & \sum_{i=1}^{M} E_i(\nu_{i,k}, I_i) + \kappa \sum_{i=1}^{M} (T_i(I_i, \nu_{i,k}) - T_{\text{req}}) + \\ & \sum_{k=1}^{K} \zeta_k \Big(\sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \alpha_{ij}((1 - I_i) I_j \nu_{i,k} r_k - R_k^{(\text{u})}) \Big) + \\ & \sum_{i=1}^{M} \phi_i \Big(\sum_{\substack{j=1 \\ j \neq i}}^{M} \alpha_{ij} \sum_{k=1}^{K} \nu_{i,k} - 1 \Big). \end{aligned} \tag{8}$$

Minimizing L will involve finding the best set of values for the parameters $\nu_{i,k}$, and I_i, $\forall i,k$. To obtain the best offloading policy (values of I_i), we write L_i as a function of I_i and a constant term (c_1) that does not depend on I_i. That is, $L_i = \Delta_i I_i + c_1$, where

$$\Delta_i = \Lambda_i + \sum_{\substack{j=1 \\ j \neq i}}^{M} (1 - I_j)\Gamma_{i,j}^{(\text{c})} - \sum_{\substack{j=1 \\ j \neq i}}^{M} I_j \Gamma_{i,j}^{(\text{m})}, \tag{9}$$

and Λ_i is independent of $\nu_{i,k}$, and can be written as

$$\Lambda_i = P_{\text{id}}\tau_i^{(\text{c})} - P_{\text{ac}}^{(\text{m})}(i)\tau_i^{(\text{m})} + \kappa(\tau_i^{(\text{c})} - \tau_i^{(\text{m})}), \tag{10}$$

and

$$\Gamma_{i,j}^{(\text{c})} = (P_{\text{id}} + \kappa) \sum_{k=1}^{K} (\alpha_{ji}\nu_{j,k}\tau_{ji,k}^{(\text{mc})} + \alpha_{ij}\gamma_{j,k}\tau_{ji,k}^{(\text{cm})}), \tag{11}$$

and

$$\begin{aligned} \Gamma_{i,j}^{(\text{m})} = \sum_{k=1}^{K} \Big(& \alpha_{ij}\nu_{i,k}P_k^{(\text{Tx})}\tau_{ij,k}^{(\text{mc})} + \alpha_{ji}\gamma_{i,k}P_k^{(\text{Rx})}\tau_{ij,k}^{(\text{cm})} + \\ & \kappa(\alpha_{ij}\nu_{i,k}\tau_{ij,k}^{(\text{mc})} + \alpha_{ji}\gamma_{i,k}\tau_{ij,k}^{(\text{cm})}) + \zeta_k \alpha_{ij}\nu_{i,k}r_k \Big). \end{aligned} \tag{12}$$

In Algorithm 1, we present an iterative algorithm to find the optimal values of $\nu_{i,k}$ and I_i for each component. The algorithm is initialized with values for the Lagrange multipliers ($\kappa, \zeta_k, \phi_i, \forall i,k$) as well as an initial allocation of where the given component i will be executed (values of I_i). The iteration index r is set to 0, and the initial value of $I_i^{(r)}$ is given by:

$$I_i^{(r)} = \begin{cases} 1 & \Lambda_i < 0, \\ 0 & \Lambda_i \geq 0. \end{cases} \tag{13}$$

This initial schedule of components implies that the component i will be scheduled to run in the cloud if the trade-off between energy consumption and execution time for running it on the cloud is favorable to running it on the mobile. To obtain optimum $\nu_{i,k}$s, we rewrite L for Component i and Radio Interface k as: $L_{i,k} = \nu_{i,k}\Omega_{i,k} + c_2$, where $\Omega_{i,k} = \sum_{j=1}^{M} \{\alpha_{ij}(1 - I_i) I_j [\tau_{ij,k}^{(\text{mc})}(P_k^{(\text{Tx})} + \kappa) + \zeta_k r_k] + \phi_i\}$,

and c_2 is a constant w.r.t $\nu_{i,k}$. The optimal value of $\nu_{i,k}$, $\nu^*_{i,k}$ for a given value of I_i is calculated as

$$\nu^*_{i,k} = \begin{cases} (1-I_i)(1-\frac{\Omega_{i,k}}{\sum_{i=1}^{M}\sum_{k=1}^{K}\Omega_{i,k}}) & \sum_{\substack{j=1\\ j\neq i}}^{M}\alpha_{ij} \neq 0 \\ 0 & \sum_{\substack{j=1\\ j\neq i}}^{M}\alpha_{ij} = 0 \end{cases} \quad (14)$$

Now by using the value of $\nu^*_{i,k}$ by using (14), I_i can be updated by

$$I_i^{(r)} = \begin{cases} 1 & \Delta_i < 0, \\ 0 & \Delta_i \geq 0. \end{cases} \quad (15)$$

The iterations continue until Eq. (8) is minimized. The algorithm converges, when the Langrange parameters have converged. The details are given in Algorithm 1.

C. Convergance and Complexity of the Algorithm

In line 1, I_i and $\nu_{i,k}$ are initialized. In a nested loop, these two variable parameters are modified such that the Lagrangian formulation in Eq. (8) is minimized. The strategy of Lines 3-17 of the algorithm has been discussed in subsection B. The variables I_i and $\nu_{i,k}$ are opportunistically updated using Eqns (15) and (14), respectively so that the objective function is minimized (lines 12,13 of the algorithm). The outer loop updates the Lagrangian multipliers using the subgradient method. Using the logic in [20], we see that the updated multipliers (κ, ζ_k, and ϕ_i, $\forall i,k$) will converge to the optimum values of I_i and $\nu_{i,k}$, $\forall i,k$.

Complexity of the modification loop (Lines 9-23) of the algorithm is O($r_{\max}M$), where $r_{\max}$ is the maximum number of iterations required to find the optimum vector $\boldsymbol{I}$. Note that we assume $M > K$. Overall, the complexity of the algorithm is O(($s_{\max}r_{\max}M$), where $s_{\max}$ is the maximum required number of iterations to satisfy all the constraints in the optimization problem. The value of $s_{\max}$ depends on the initial values in line 6 and ϵ values in lines 28, 29 of the algorithm. In the simulations (Section VI), we observe that the mean values of $s_{\max}$ and $r_{\max}$ are 3 and 2, respectively. The complexity of the exhaustive search method is O($2^M \times k$), which is prohibitively high.

Algorithm 1 Proposed Radio Aware Offloading Schedule.

1: **initialization**:
2: Set $r \leftarrow 0$, modification index, $s \leftarrow 1$
3: Set $I_i^{(0)}$ using Eq. (13)
4: Set $\Delta^{(0)}$ using Eq. (10)
5: Set $\nu_{i,k}^{(0)}$ using Eq. (14)
6: Set initial values for parameters $\kappa^{(s)}, \zeta_k^{(s)}, \phi_i^{(s)}$
7: Set $X_r = X_s \leftarrow$False
8: **repeat**:
9: **if** $\Delta_i^{(r)} < 0, \forall i$ **then**
10: **while** X_r=False **do**
11: calculate $\Delta_i^{(r+1)} = \Delta_i|_{I_i=I_i^{(r)},\nu_{i,k}=\nu_{i,k}^{(r)}}$ by (9)
12: calculate $I_i^{(r+1)}$ by Eq. (15)
13: calculate $\nu_{i,k}^{(r+1)}$ by Eq. (14)
14: **if** $\exists i : \Delta_i^{(r+1)}\Delta_i^{(r)} < 0$ **then**
15: **Find** $\min_{\tilde{i}}(\Delta_i^{(r+1)}; \forall i)$
16: $I_{\tilde{i}} \rightarrow 1 - I_{\tilde{i}}$,
17: **end if**
18: **if** $\sum_{i=1}^{M} L_i^{(r+1)} \geq \sum_{i=1}^{M} L_i^{(r)}$ **then**
19: X_r=True,
20: **end if**
21: $r \rightarrow r+1$,
22: **end while**
23: **end if**
24: $\kappa^{(s+1)} = \kappa^{(s)} - \varepsilon_\kappa(T_{\text{req}} - \sum_{i=1}^{M} T_i)$
25: $\zeta_k^{(s+1)} = \zeta_k^{(s)} - \varepsilon_\zeta \times$
26: $(R_k^{(\text{u})} - \sum_{i=1}^{M}\sum_{\substack{j=1\\ j\neq i}}^{M} |I_i - I_j|(1-I_i)\alpha_{ij}\nu_{i,k}r_k), \forall k$
27: $\phi_i^{(s+1)} = \phi_i^{(s)} - \varepsilon_\phi(1 - \sum_{\substack{j=1\\ j\neq i}}^{M}\alpha_{ij}\sum_{k=1}^{K}\nu_{i,k}), \forall i$
28: **if** $\frac{|\kappa^{(s+1)}-\kappa^{(s)}|}{\kappa^{(s+1)}} < \varepsilon_\kappa$ & $\frac{|\zeta_k^{(s+1)}-\zeta_k^{(s)}|}{\zeta_k^{(s+1)}} < \varepsilon_\zeta$ &
29: $\frac{|\phi_i^{(s+1)}-\phi_i^{(s)}|}{\phi_i^{(s+1)}} < \varepsilon_\phi, \forall i,k$ **then**
30: X_s = True,
31: **end if**
32: $s \rightarrow s+1$
33: **until** any constraint in Eqs (4),(6),(7) is not satisfied: (X_s=False).

IV. Performance Analysis

In this section, we investigate the efficiency of the proposed approach using an HTC Vivid smartphone with a 1.2 GHz dual core processor. This phone is equipped with two radio interfaces (k = 2): WiFi, and LTE. Moreover, we assume that whereas LTE is always available, the WiFi interface can sometimes be unavailable (as is common in real life scenarios). A multi-component video navigation application was used for the experiments. This application uses video processing, face detection, graphics, and clustering the video points. Graphics library tools are used from the OpenGL mobile Android applications [21], face detection is used from [22], and all of the video processing features are available in [23]. We used fourteen component applications to form the codeset in our work. Note that the first and last components are executed in the mobile device, because most mobile initiated applications must start in the mobile device and usually have an output/display that happens on the mobile device. We measured execution time of the components in the HTC phone and the cloud, uplink and downlink rates and delays for WiFi and LTE. We obtained the dependency matrix of this application, and the size of the data that needs to be transferred between components. The Amazon Elastic Compute Cloud (Amazon EC2) was used for cloud computing capacity [24]. The average transmit power levels of the mobile device for WiFi, and LTE services are 300 and 600 mWs, respectively. The average received power levels were 100 and 250 mWs, respectively. The active and idle power levels of the phone are 644.9 and 22 mWs, respectively. The power

consumption for the last component in the mobile device was 55 mWs. These power measurements are obtained by using CurrentWidget: Battery monitor application [25]. The average wireless service rates for WiFi, LTE are 0.80 and 2.96 Mbps for the uplink transmission and 1.76 and 4 Mbps for the downlink transmission, respectively. Also, local execution time of the fourteen components are measured as [30 340 345 125 30 80 70 30 185 125 650 571 904 56] ms. The number of arriving requests is modeled as a Poisson distributed variable with average rate of 1.5 Mbps. The initial multiplier values for κ, ϕ and ζ were set to 0.1, 0.1, and 10^{-6}, respectively. The results shown are averages of 1000 independent test runs.

Four scenarios are compared in this section. First, we consider the scenario that all components are executed locally in the mobile. The energy consumed in this scenario is used to normalize all energy values. The second scenario consists of executing the entire application on the cloud (other than the first and the last components). In this scenario, all data must be uploaded to the cloud. The third scenario is a brute force exhaustive search for the best values of I_i for each component. That is, we manually schedule components $i = 2$ through 13 to run on either the cloud or the mobile and calculate the associated energy and time. Note, that since the first and last component must run on the mobile, we are left with $2^{(14-2)}$ combinations of possible values for the I_i's. For each combination of $\mathbf{I}$, the problem turns out to be a linear optimization over the variable set $\boldsymbol{\nu}$. Thus, the radio allocation percentages are calculated using linear programming. The sets of I_i and $\nu_{i,k}$, $\forall i, k$, values which minimize the energy consumption give the over all optimal solution. The approach in this scenario is called "Exhaustive search". Finally, the fourth set of results is obtained by our iterative algorithm.

Fig. 2 shows the average energy consumption for four different approaches while the application execution time equals to 3.54 seconds. We observe that the proposed approaches (exhaustive search and the proposed iterative algorithm) result in lower energy consumptions in comparison to the others. Note that 3.54 seconds is the minimum execution time to execute the application locally, so that the execution time deadline is satisfied in all of the approaches. On an average, the proposed iterative algorithm consumes 3% more energy in comparison to the proposed optimal solution (Exhaustive search approach) for $T_{\text{req}} = 580$ ms. This is a fairly good trade-off for the reduced complexity of the proposed iterative algorithm. Fig. 3 presents the execution time of different approaches in different scenarios. While local and remote execution approaches require longer application execution time, the proposed scheme gives us 29% and 27% faster execution time in comparison to these approaches respectively with the same amount of energy consumption to the remote execution approach. If we desire to save 9% of energy, then we have still 9% and 6% faster execution time in comparison to local and remote execution respectively. On the other hand, if only fast execution of the application is important for us, then by costing 11% more energy than remote execution, we can achieve 50% and 48% faster run rather than local and remote execution. Fig. 4 plots the energy-execution time trade-off in the proposed scheme in comparison to the local and remote execution, while the proposed scheme takes advantage of three scenarios for radio resources: 1. WiFi and LTE are used jointly; 2. only WiFi is used for offloading; and 3. only LTE is used for offloading. The four points in the plot show local and remote executions by using only LTE, only WiFi, or both. We see that although remote execution by using LTE consumes much more energy in comparison to the others, the execution time for this scenario would be less than the others. Thus, there is a trade off between energy consumption and execution time of the application which is relied on the delay of offloading. On the other hand by using the proposed offloading scheme lesser energy is consumed with reasonable value for execution time. When the execution time deadlines are longer, there is more flexibility in offloading jobs to the cloud and hence energy consumptions for the mobile device reduces. Also, it is clear that joint use of radio resources gives less energy consumption and requires less execution time.

Fig. 5 plots the percentage of data stream to the cloud through WiFi (radio interface 1) versus RTT of the WiFi and LTE in the proposed scheme. We observe that by increase of RTT in WiFi for the range of 40-160 ms, less data stream is allocated to WiFi and more data stream is allocated through LTE for computation offloading. On the other hand, when RTT of LTE increases in the range of 50-200 ms, more data stream is allocated to WiFi and less data is allocated to LTE. Finally,

V. Conclusion

We studied the problem of offloading computationally intense applications from mobile devices to a cloud infrastructure for multi-radio equipped mobile devices. We presented a comprehensive model for the energy consumed in offloading components to the cloud. We modeled the decision to offload any given component to the cloud as an optimization problem that seeks to resolve the conflicting goals of reducing computation costs while keeping the execution time of the application below its deadline. We showed that this is a non-linear optimization problem. We proposed an iterative algorithm to find the local optima for the offload schedule of the components as well as the percentage of the data to be carried on each radio interface. We showed that the proposed algorithm consumes within 4% of the optimal solution (obtained via brute force search) and also offers 31% less energy consumption in comparison to offloading the entire application to the cloud.

References

[1] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. Johnson, "M2M: From mobile to embedded internet," *IEEE Communication Magazine*, vol. 49, no. 4, pp. 36–43, 2011.

[2] K. Kumar and Y. H. Lu, "Cloud computing for mobile users: Can offloading computation save energy?" *Computer*, vol. 43, pp. 51–56, 2010.

[3] N. Vallina-Rodriguez and J. Crowcroft, "Energy management techniques in modern mobile handsets," *IEEE Communications Surveys Tutorials*, vol. 15, no. 1, pp. 179–198, 2013.

[4] X. Gu, K. Nahrstedt, A. Messer, I. Greenberg, and D. Milojicic, "Adaptive offloading for pervasive computing," *IEEE Pervasive Computing*, vol. 3, no. 3, pp. 66–73, 2004.

[5] B.-G. Chun, S. Ihm, P. Maniatis, M. Naik, and A. Patti, "Clonecloud: elastic execution between mobile device and cloud," in *Proceedings of the sixth conference on Computer systems*, 2011, pp. 301–314.

[6] S. Merlin, N. Vaidya, and M. Zorzi, "Resource allocation in multi-radio multi-channel multi-hop wireless networks," in *Proc. of INFOCOM*, April 2008.

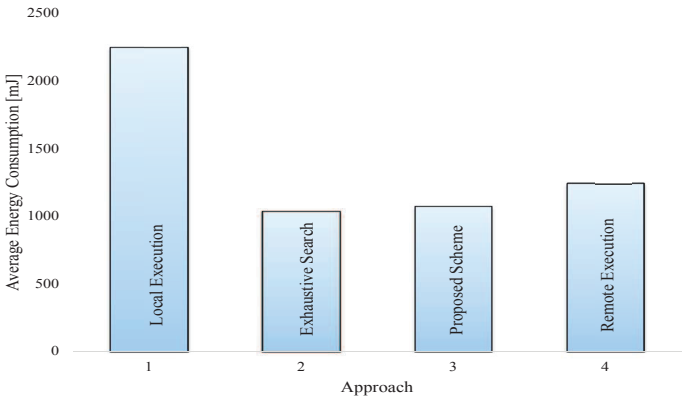

Figure 2. Average Energy Consumption of the Four Approaches, while execution time equals 3.54 seconds.

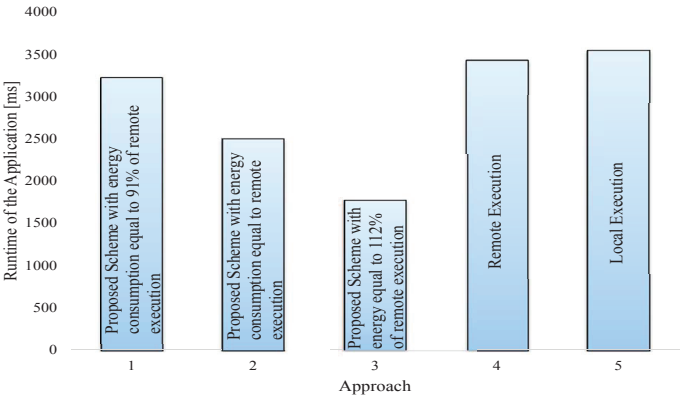

Figure 3. Execution time of Different Approaches.

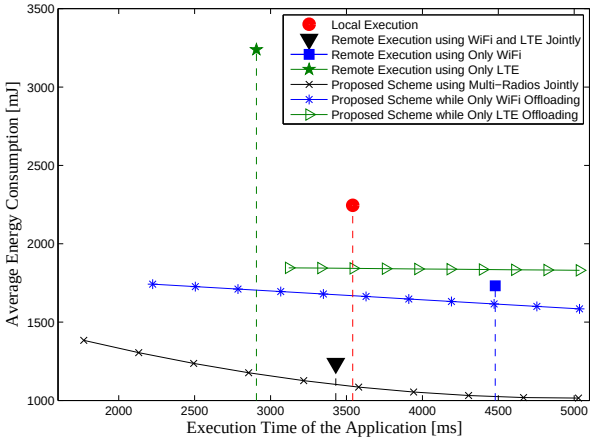

Figure 4. Average Energy Consumption versus execution time of the application. Jointly, it shows the trade-off between the cost of energy consumption and execution time in the proposed scheme. We observe that the application can be executed in half of the time (0.52) it takes to be executed in the cloud with the cost of 12% more energy consumption, and also the application can be executed with 20% more energy saving in comparison to the remote execution with the cost of 42% execution time extension in comparison to remote execution.

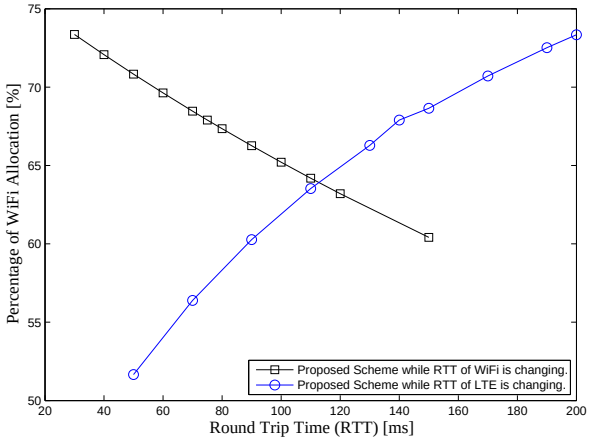

Figure 5. Percentage of WiFi Allocation in the Proposed Scheme versus Round Trip Time (RTT) of WiFi and LTE.

[7] V. Bhandari and N. H. Vaidya, "Scheduling in multi-channel wireless networks," *bookchapter of Distributed Computing and Networking, Springer*, 2010.

[8] P. Shu, F. Liu, H. Jin, M. Chen, F. Wen, and Y. Qu, "etime: Energy-efficient transmission between cloud and mobile devices," in *Proc. of INFOCOM*, April 2013, pp. 195–199.

[9] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. Wu, "Energy-optimal mobile cloud computing under stochastic wireless channel," *IEEE Transactions on Wireless Communications*, vol. 12, no. 9, pp. 4569–4581, 2013.

[10] Y. Wen, W. Zhang, and H. Luo, "Energy-optimal mobile application execution: Taming resource-poor mobile devices with cloud clones," in *Proc. of the IEEE International Conference on Computer Communications (INFOCOM)*, 2012, pp. 2716–2720.

[11] D. Huang, P. Wang, and D. Niyato, "A dynamic offloading algorithm for mobile computing," *IEEE Transactions on Wireless Communications*, vol. 11, no. 6, pp. 1991–1995, 2012.

[12] D. Kovachev, T. Yu, and R. Klamma, "Adaptive computation offloading from mobile devices into the cloud," in *IEEE Symposium on Parallel and Distributed Processing with Applications (ISPA)*, 2012, pp. 784–791.

[13] E. Cuervo, A. Balasubramanian, D.-k. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl, "MAUI: Making smartphones last longer with code offload," in *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '10. ACM, 2010, pp. 49–62.

[14] S. Kosta, A. Aucinas, P. Hui, R. Mortier, and X. Zhang, "Thinkair: Dynamic resource allocation and parallel execution in the cloud for mobile code offloading," in *Proc. of IEEE International Conference on Computer Communications (INFOCOM)*, 2012, pp. 945–953.

[15] X. Wang, A. V. Vasilakos, M. Chen, Y. Liu, and T. T. Kwon, "A survey of green mobile networks: Opportunities and challenges," *Mobile Network Applications*, vol. 17, no. 1, pp. 4–20, Feb. 2012.

[16] A. K. S. G. Xinwen Zhang, Sangoh Jeong, "Towards an elastic application model for augmenting computing capabilities of mobile platforms," in *Mobile Wireless Middleware, Operating Systems, and Applications*, vol. 48. Springer, 2011, pp. 161–174.

[17] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Computation offloading for mobile cloud computing based on wide cross-layer optimization," in *Journal of Future Network and Mobile Summit*, July 2013, pp. 1–10.

[18] K. Hong, S. Sengupta, and R. Chandramouli, "Spiderradio: A cognitive radio implementation using ieee 802.11 components," *IEEE Transactions on Mobile Computing*, vol. 12, no. 11, pp. 2105–2118, November 2013.

[19] D. Kaspar, "Multipath aggregation of heterogeneous access networks," *SIGMultimedia Rec.*, vol. 4, no. 1, March 2012.

[20] S. Boyd and A. Mutapcic, "Subgradient methods," *Lecture Notes of EE364b, Stanford Univ., Stanford, CA*, Spring quarter 2008.

[21] Mar. 2014. [Online]. Available: http://www.opengl.org/.

[22] Jul. 2014. [Online]. Available: http://www.developer.com/ws/android/programming/face-detection-with-android-apis.html.

[23] Apr. 2014. [Online]. Available: http://opencv.org/.

[24] Jul. 2014. [Online]. Available: http://aws.amazon.com/ec2/.

[25] Jul. 2014. [Online]. Available: http://code.google.com/p/currentwidget/.