Is there a Net Energy Saving by the use of Task Offloading in Mobile Devices?

Vishnu Cherusola Dev  Wichita State University, vxcherusoladev@wichita.edu
Vinod Namboodiri  Wichita State University, vinod.namboodiri@wichita.edu
Milind Tandel  Wichita State University, mhtandel@wichita.edu
Vijay Venkitachalam  Wichita State University, vxvenkitachalam@wichita.edu

Abstract. The cloud computing paradigm enables the work anywhere anytime paradigm by allowing application execution and data storage on remote servers. This is especially useful for mobile computing and communication devices that are constrained in terms of computation power and storage. Relying more on powerful remote servers further allows each individual device’s hardware to have reduced functionality, thus making them cheaper. Though there could be energy and performance benefits for mobile devices by offloading application tasks to remote servers, it is not clear whether there would be reduction in total end-to-end energy consumed when including energy consumed at the remote server and the network in between. The goal of this paper is to characterize the overall end-to-end energy consumed for task offloading in mobile devices. An analytical model of end-to-end energy consumed for task offloading is proposed and evaluated to understand the theoretical limits within which task offloading can save energy. Additionally, an empirical measurement-based evaluation is done with common off-the-shelf mobile devices to determine whether task offloading is more energy-efficient end-to-end compared to a scenario where the task is executed locally.

Introduction. Computing with mobile devices have always presented challenges in terms of storage, memory, processing, network connectivity, bandwidth, and battery lifetime in comparison to their static counterparts like desktop computers. With the technological advances in recent years improving ubiquitous connectivity and bandwidth, cloud computing has become feasible allowing these constrained devices to utilize the greater storage, memory, and processing capabilities of powerful remote servers.


Copyright © 2015 by Vishnu Cherusola Dev, Vinod Namboodiri, Milind Tandel, Vijay Venkitachalam, Licensed under CC-BY 3.0.

Cloud computing is typically a client server architecture, where the client can be any mobile device like a laptop, phone, browser, or any other operating system enabled device. Users of mobile devices that want to share documents, check email, and surf the Internet on the go, represent an increasing segment of the population that can utilize the cloud concept. Google’s introduction of the cloud based Chrome operating system (OS), for example, fuels this trend of increasingly relying on remote servers instead of local device resources.

The energy implication of offloading a task or set of tasks (termed task offloading) to a remote server for execution is not very clear. On one hand relying on a server can reduce energy spent on task computation; on the other hand, communicating the task to the server will incur network-wide energy consumption. Such task offloading scenarios have been studied extensively for performance benefits or for exploring the feasibility of thin-client computing [1], [2], [3], [4], [5],[6], [7]. More recently such paradigms have also been studied in terms of reducing energy consumed by mobile devices and increasing battery lifetimes [8], [9], [10], [11], [12], [13]. These results have provided good guidance on when it is beneficial in terms of energy consumption for a mobile device to offload tasks. However it is not clear whether there will be benefits end-to-end when the energy consumption of the network and the remote server are brought into consideration. A broader study of the energy implications of task offloading from an end-to-end perspective is still missing. Such an end-to-end study would be one of the first steps in the path towards environmentally sustainable mobile computing. When it comes to looking at energy efficiency and the concept of sustainability in information and communication technologies (ICT), the focus has invariably been on data centers and mobile infrastructures like cell towers, as these have been considered the most power intensive within the computing sector. With an increasing emphasis on utilizing cloud computing through mobile devices, the study and analysis of the complete end-to-end scenario from mobile devices to end-servers is needed to further sustainability efforts in the area of information and communication technologies (ICT).

The specific novel contributions made by this work are the following:

- An analytical model of energy consumption for task offloading to remote servers is developed that includes the end-user device energy consumption, the remote server energy consumption, and the energy consumed by the network in between.
- An evaluation of the theoretical limits of various parameters in the analytical model is done for which task offloading can be shown to reduce overall end-to-end energy consumption
- An empirical evaluation of common-off-the-shelf mobile devices is performed to determine whether task offloading can reduce end-to-end energy consumption. Our results with a remote server of modest capabilities and a highly computational task shows that a end-user device like a laptop with only slightly lesser computational capabilities (than the remote server) could reduce end-to-end energy consumption by 30% easily with task offloading. For a much more relatively computationally-constrained device like a smartphone, the energy consumption can be considerably reduced.

**Problem Statement.** This section will focus on an analytical characterization of the task offloading problem with a formal problem definition, and an evaluation of the theoretical limits of model parameters under which task offloading can reduce overall end-to-end energy consumed.

**Problem Definition.** We will consider one mobile device (also termed as the client) that has to execute a task. It has the following two options: execute this task on the device locally, or offload the task to a remote
server over the network and have it executed there and have the results communicated back. Let $E_{\text{off}}$ be the total energy consumed to execute the task on the client device itself when no offloading is done. Thus, this energy consumption under the no offloading scenario to execute the task equals the product of the average power consumed to execute the task locally ($P_c$) and the time to take to complete the task $t_c$.

$$E_{\text{off}} = P_c \times t_c \quad \text{(1)}$$

Under the offloading scenario, the energy consumed end-to-end is the sum of (i) energy consumed by the client $E_{\text{co}}$, (ii) energy consumed in the network $E_{\text{nw}}$, and (iii) energy consumed at the remote server executing the task $E_{\text{serv}}$. That is,

$$E_{\text{off}} = E_{\text{co}} + E_{\text{nw}} + E_{\text{serv}} \quad \text{(2)}$$

$$= P_{\text{co}} \times t_{\text{co}} + P_{\text{nw}} \times t_{\text{nw}} + P_s \times t_s \quad \text{(3)}$$

For a given task, we can say that offloading saves energy end-to-end if

$$E_{\text{off}} < E_{\text{off}} \quad \text{(4)}$$

**Theoretical Limits.** Here we numerically investigate the theoretical limits within which offloading could reduce the energy consumed for executing a task. We introduce some assumptions and some parameters to evaluate the condition in Equation 4.

Let $p_{\text{nw/c}}$ be the ratio of average power consumed by the network to that of the client to execute the task. That is,$ p_{\text{nw/c}} = P_{\text{nw}}/P_c$. Similarly, let $p_{\text{sc}} = P_s/P_c$ be the ratio of average power consumed by the server to execute the task compared to the client locally. If we introduce another ratio $p_{\text{co/c}}$ to denote the ratio of average power consumed by the client when offloading versus executing the task locally, Equation 2 can be re-written as

$$E_{\text{off}} = P_{\text{co}} \times t_{\text{co}} + p_{\text{nw/c}}/p_{\text{co/c}} \times P_{\text{co}} \times t_{\text{nw}} + p_{\text{sc}}/p_{\text{co/c}} \times P_{\text{co}} \times t_s \quad \text{(5)}$$

Equation 1 can be re-written as

$$E_{\text{off}} = P_{\text{co}}/ p_{\text{co/c}} \times t_s/ r_{\text{s/c}} \quad \text{(6)}$$

where $r_{\text{s/c}} = t_s/t_c$ is the ratio of relative capabilities of the server and client respectively in terms of processing times for the same task. From Equations 5 and 6, the energy saving condition for task offloading reduces to

$$t_{\text{co}} + (p_{\text{nw/c}}/p_{\text{co/c}}) \times t_{\text{nw}} + p_{\text{sc}}/p_{\text{co/c}} < t_s / (p_{\text{co/c}} \times r_{\text{s/c}}) \quad \text{(7)}$$

If we neglect the delays for packet processing, transmission, and queuing for simplicity, and the total time for task offloading at the client $t_{\text{co}}$ would be equal to the sum of the total time spent in the network (both directions), $t_{\text{nw}}$, and the execution time at the server $t_s$, or $t_{\text{co}} = t_{\text{nw}} + t_s$. If we also introduce a ratio $r_{\text{nw/c}} = t_{\text{nw}}/t_c$ as the ratio of time spent to offload the task over the network to the client local task execution time, then Equation 7 can be re-written as the condition.
Is there a Net Energy Saving by the use of Task Offloading in Mobile Devices?

\[
1 - \frac{p_{\text{co}} \cdot r_{z/c} - r \cdot p_{z/c}}{r_{\text{nw/c}} (p_{\text{co/c}} + p_{\text{nw/r}})} > 1
\]

\[\text{...........................................} (8)\]

**Numerical Evaluation.** Here we evaluate the condition in Equation 8 for a range of values for the power and processing capability ratios introduced above. The ratios \(p_{\text{co/c}}, p_{\text{nw/c}}, r_{s/c}, \text{ and } r_{\text{nw/c}}\) are varied to study the impact on the range of \(p_{s/c}\) over which task offloading consumes less energy. For a given value of \(r_{\text{nw/c}}\), it can be observed from the results in Figure 1 that a smaller client offload to no offload power ratio \(p_{\text{co/c}}\) increases the range under which energy can be saved. Thus, efficiency of the low-power state when a client is waiting for a server to return a completed task is of great significance. It can also be seen that a lower network to client offload power ratio \(p_{\text{nw/c}}\) helps increase the range over which offloading is superior. For small values of \(r_{s/c}\), offloading is always preferable for all values of \(p_{s/c}\). That is, when the server can execute the task much faster than the client, offloading always saves energy overall. For \(r_{s/c}\) closer to 1, offloading is better only at lower values of \(p_{\text{nw/c}}, p_{\text{co/c}}, \text{ and } p_{s/c}\).

Comparing across different values of \(r_{\text{nw/c}}\) Figure 1 shows that as an offloaded task takes more time to be communicated back and forth from the server, the impact of increasing values of \(p_{\text{co/c}}\) and \(p_{\text{nw/c}}\) becomes more significant. That is, as a task incurs more communication delays, the range of power values of model parameters under which task offloading saves energy becomes narrower. The analytical model and its evaluations provide a means to study the impact of different parameters relative to each other and quantify their impact. The presented model is also useful to generalize the potential of end-to-end energy savings for any combination of end-user device, server, and network power consumption and processing capabilities, and for any type of tasks. In the next two sections, we look at specific current COTS devices and deployed networks and study whether offloading would reduce overall energy consumed for a specific task under consideration.

**Empirical Methodology.** In this section we describe our empirical approach to understand the conditions under which task offloading could save energy end-to-end. This section will focus more on the experimental methodology with evaluation results presented in the following section.

**Overview.** The empirical approach was divided into two cases. The first case involved measuring the energy consumed to execute a defined task only on a mobile device. The second case involved the mobile client offloading the task to a remote server which the latter completes and sends back. The energy consumed at the client, the server, and the network in between is measured in the second case with the sum of the three reported as the total energy consumed to execute the task. The results of the energy measurements from these two cases are then used to evaluate the condition first presented in Equation 4 if \(E_{\text{off}} < E_{\text{noff}}\).
Experimental Setup.

Computing Devices. Mobile clients of two classes (based on computational capability) were tested, a laptop, and two smartphones. The laptop was a Lenovo SL3000 with an Intel core 2 duo CPU with each core running at 2.4GHz and 2 GB RAM. One of the smartphones was a Motorola Milestone XT 270 and ran Android OS v2.1 with 600 MHz CORTEX-A 8 CPU. The other smartphone was a HTC Desire A8181 with a 1 GHz Scorpion CPU from Qualcomm and ran Android OS v2.2. All devices used the Wi-Fi interface (802.11g) for communicating over a network. The server machine emulating a cloud server was a more powerful laptop machine with an Intel core i3 CPU running at 2.53GHz and 4 GB RAM. Thus, the laptop end-user client was only slightly less computationally-capable than the server, while the smartphones were much more computationally-constrained than the server. Between the smartphones, the Motorola device considered was more computationally capable.

Task. The task considered was to sort a file containing \( N \) random integers repeatedly 200,000 times using the common insertion sort algorithm. Sorting is a task whose computation intensity and execution time can be easily varied by varying \( N \) making it very useful for the problem under consideration. Insertion sort is a commonly used sorting algorithm that is well-understood. Note that the particular sorting algorithm used could be easily changed and will not have any bearing on the conclusions drawn. The time to complete the task at the client (no offloading scenario)

Figure 1. Impact of various parameter values on range over which task offloading can save energy over local execution with \( r_{nw/c} = 0.05, 0.25, \) and \( 0.5, \) and \( p_{cc/c} = 0.1 \) and 1.
involved using a simple timer. For the time to complete a task at the server (offloading scenario) the timer was set at the client before the task was sent over the network and the value was saved when the results were received back. Due to the wide variance in processing capabilities, we kept $N = 10000$ for the laptop and $N = 100$ for the smartphone. Experiments were also conducted to understand the impact of task complexity on the results by varying the size of the sorting input as $N; 2N; 4N$, and $8N$.

**Power Measurements.** Energy consumption was computed as the product of average power consumed over the time to complete the task. The time duration to complete the task was explained above. The power consumption was measured through the Watts up? PRO meter for all devices except the smartphone for which the Monsoon Power Monitor tool was used. The Watts up? PRO meter allows recording power values over a period of time over a USB port and can provide average power and energy consumed as output. This meter had adequate resolution for capturing tasks that took well over 5-10 seconds to complete and has a stated accuracy range within ±1.5% of the actual value. For the smartphone, the commonly used Monsoon Power Monitor was used. It has a stated accuracy range within ±1% of the actual value.

**Populating the Energy Model.** Evaluating the energy consumed for the offloading and no offloading scenarios involves calculating each term in the energy models proposed in Equations 1 and 2. Energy consumed with no offloading: Calculating $E_{\text{noff}}$ in Equation 1 was as simple as measuring the average power consumed by the client to complete the task locally $P_c$ and multiplying it by the time taken to complete the task $t_c$. $P_c$ was computed by subtracting out a “base” power value for the device when idling and thus including only the additional power consumed to execute the task. This base power value denotes the power consumed by the device when it is not executing any tasks and includes power to run the device’s display and the Wi-Fi interface in a low-power sleep state when not actively communicating.

Energy consumed for task offloading. Calculating $E_{\text{off}}$ in Equation 2 was more involved due to multiple components spread over the network including a remote server. Our server as mentioned before was a more powerful laptop device. The network between the two was a LAN environment to get a best case result on top of which an analytical model was superimposed to vary the actual server location and WAN network characteristics.

For energy consumed by the client device when offloading $t_{co}$ was computed by computing the time required to send the task to the remote server and getting it back, including the communication network latency. $P_{co}$ was computed over this time and was found to be close to the base power of the device when it was awaiting results back from the remote server with spikes only for packet processing.

For energy consumed by the server device when executing the task, the energy consumed was found in a similar fashion to the client device. The time taken to complete the task at the server, $t_s$, was computed along with the average power consumed in the process $P_s$ after subtracting off the base idling power of the server device. Note that $t_s < t_{co}$, with the difference primarily being the round trip network delay between the client and server. We also account for the overhead due to heating and cooling costs by the commonly used Power Usage Effectiveness (PUE) factor for only this server giving
According to the Uptime Institute in a survey of data centers in 2012, the typical range of PUE was between 1.8 to 1.89. Thus we used the value of 1.8 for PUE in our calculations. Finally, we computed the energy consumed in the network. A router's energy consumption was modeled as in [15] by

\[ E_{\text{router}} = V_r \times P_r / C \times U \]  

(10)

where \( V_r \) is the total routed data volume in bits over the period of consideration, \( P_r \) is the average power consumption of the router, \( U \) is the utilization of the router, and \( C \) is the data handling capacity of router in bits per second. If we ignore protocol overhead, the data volume in our case will be equal to the data sent as part of task offloading to the remote server and the result back from the server. This is easily known based on the number of integers \( N \) that are being sorted for the example considered in this paper. We assume a utilization \( U \) of 25% as done in [15]. The parameters \( P_r \) and \( C \) can be easily found from specifications of a specific router.

Our network model consisted of a client router, \( k_e \) edge routers, and \( k_c \) core routers. The overall energy consumed in the network was computed as

\[ E_{\text{nw}} = E_{\text{user}} + k_e \times E_{\text{edge}} + k_c \times E_{\text{core}} \]  

(11)

where we put in values corresponding to specifications for each type of router as listed in Table I. Our results in the next section are presented with \( k_e = 2 \) and \( k_c = 18 \) to model a long-distance WAN with 21 network hops overall. Note that this assumption likely overestimates the energy consumed by the network as, in most cases, \( k_c \) is likely to be much smaller than 18.

Empirical Results. In this section we present results empirically measured on different COTS mobile devices by applying the methodology described in the previous section. The first set of results are for a laptop, an end-user device which was only slightly less computationally-capable than the server, while the second set of results are for smartphones, a much more computationally constrained device than the server. The presented results try to answer two main questions for each class of end-user device: (i) how does the energy consumption compare between executing the task locally and remotely, and how does the energy consumption in the offloading scenario break down across the individual components \( E_{\text{co}}, E_{\text{nw}}, \) and \( E_{\text{serv}} \), and (ii) how does the energy consumed in the offloading and no offloading case vary as the computation required (and hence processing duration) by the task increases. Each experiment was run 12 times with plots showing the mean value and standard deviation denoted S.D.

Results for laptops. The initial result of interest for laptops was whether task offloading would save energy in laptops, and by how much. Figure 2(a) plots \( E_{\text{nooff}}, E_{\text{off}}, \) and the individual terms that comprise \( E_{\text{off}}, \) namely, \( E_{\text{co}}, E_{\text{nw}}, \) and \( E_{\text{serv}} \). All terms except \( E_{\text{nw}} \) were measured empirically by repeatedly taking 12 different measurements with the mean value shown in the plot. The standard deviation of these measurements is also indicated above each bar. It can be observed that task offloading reduced the energy consumed by almost 50% over the no-offloading scenario.
Is there a Net Energy Saving by the use of Task Offloading in Mobile Devices?

The energy consumed by the server to execute the task was the largest component in $E_{off}$ with the network and client energy consumption relatively much smaller. The next result of interest for laptops was the impact of the complexity and duration to execute the task. This was studied by varying the number of integers to be sorted as per the sequence $N$, $2N$, $4N$, $8N$, and $16N$ for

**Figure 2.** Energy consumed to execute task with and without offloading on laptops.
$N = 10000$. This study was more or less an analysis of the sensitivity of energy consumption to task complexity and duration. Figure 2(b) plots $E_{\text{off}}$ and $E_{\text{off}}$ for varying input sizes. It can be seen that both increase with the increase in input sizes in line with the $O(n^2)$ complexity of insertion-sort. In absolute terms, greater the input size, greater the energy saved by offloading for the scenario considered. In relative percentage terms, however, the converse seems true with the ratio of $E_{\text{off}}/ E_{\text{off}}$ decreasing with greater input size. For the inputs tested, and relative performance difference between the laptop and server used, 30% savings in end-to-end energy savings was easily possible by task offloading.

**Results for Smartphones.** The initial result of interest for mobile phones was again whether task offloading would save energy in a computationally-constrained form-factor device like smartphones and by how much. Plots for two different smartphones that we considered are shown in Figures 2(c) and 2(d). An important distinction from the experiments for laptops is that $N$ is set now equal to only 100 due to the much lesser computational capabilities of these devices. Both plots show that even for smartphone, task offloading seems to consume much less energy, an even more pronounced difference than in laptops. This is because $E_{\text{off}}$ values are much larger for local execution in phones whose processing capabilities are limited and the duration to complete the task being relatively much larger than on the powerful server. This trend is seen on both phones, though $E_{\text{off}}$ is much smaller on the Motorola phone as it is newer hardware model with higher processing capability and a newer OS. The end-to-end energy consumption was reduced considerably in both cases. The next result of interest for smartphones was based on the sensitivity to input size. Just as the case of laptops, the input size was varied in multiples of two. Figures 2(e) and 2(f) show these results. The sensitivity analysis for both phones to varying input size demonstrated that a powerful server can handle much larger input sizes easily compared to a phone and hence offers a continuing advantage that is much higher than what was seen for laptops.

**Discussion of Results.** The results for both classes of mobile devices studied in this section highlight that the benefits of task offloading increases as the computational-capability of the end-user device decreases. For the laptop, reduction in end-to-end energy consumption was only of the order of 30% because the server itself was a laptop, albeit a more powerful one. If a more commercial grade server were used, much greater benefits to task offloading could perhaps be seen. Surprisingly, the energy consumed in the network was much lesser than we had anticipated before our experiments in spite of the worst case assumption of 21 hops as network diameter. After studying these empirical results, it becomes apparent that the analytical model presented in Section II is a useful complementary contribution as well in generalizing when task offloading is likely to reduce end-to-end energy consumption. For example, one can compute (omitted due to space limitations) the range at which offloading will save energy for some given power consumption values and processing capabilities at the client, network, and server and predict the benefits of offloading for other unknown values.

**Conclusions.** This paper characterized the overall end-to-end energy consumed for task offloading in mobile devices. An analytical model of end-to-end energy consumed for task offloading was proposed and evaluated to understand the theoretical limits within which task offloading can save energy. Additionally, an empirical measurement-based evaluation was done with common-off-the-shelf mobile devices to determine whether task offloading is more energy-efficient end-to-end compared to a scenario where the task is executed locally. Our results show that task offloading can significantly reduce end-to-end energy consumed, with greater benefits seen when end-user mobile devices are less powerful; for a smartphone considerable reduction in energy
Is there a Net Energy Saving by the use of Task Offloading in Mobile Devices?

...consumption is seen as opposed to a laptop for which energy reduction is 30%. Future work to be done could include studying how results may vary when the profile of tasks to be offloaded is changed. Multiple types of servers could also be experimented with studying the impact of server processing capability empirically. The analytical model of the network energy consumption could also be improved by adding more details.

References


