RESEARCH STATEMENT

Feng Qian (fengqian@umich.edu)

Research Overview

My general research interests cover the broad area of computer networking, mobile computing, and network measurement, with special emphasis on cellular data networks, which experienced significant growth in the recent years particularly due to the emergence of smartphones. As reported by a major U.S. carrier [**?**], its cellular data traffic has experienced a growth of 5000% over 3 years [**?**]. Despite its popularity, there remain two major challenges associated with cellular carriers and their customers: *carriers operate under severe resource constraints, while mobile applications often utilize radio channels and consume handset energy inefficiently.*

From the carriers' perspective, compared to the Wi-Fi and wired networks, cellular systems operate under more resource constraints. To keep up with the explosive increase of their cellular traffic, all U.S. carriers are expected to spend 40.3 billion dollars on cellular infrastructures in 2011 [**?**]. Cellular networks employ a unique resource control mechanism to manage the limited resources [**?**]. However, my research identified significant inefficiency in the current resource management policy. For example, by analyzing the data collected from a large commercial 3G carrier, I found that up to 45% of the high-speed transmission channel occupation time is wasted on idling, because the critical parameters controlling the release of radio resources are configured in a static and ad-hoc manner. Cellular carriers therefore urgently need methods to systematically characterize and optimize resource usage for their networks.

From the customers' perspective, there is a plethora of mobile applications developed by both enthusiastic amateurs and professional developers. As of October 2011, the Apple app store had more than 500K mobile apps with 18 billion downloads. Smartphone applications are different from their desktop counterparts. Unfortunately, mobile application developers often overlook the severe resource constraints of cellular networks, and are usually unaware of the cellular specific characteristics that incur complex interaction with the application behavior. This potentially results in smartphone apps that are not cellular-friendly, *i.e.,* their bandwidth usage, radio channel utilization and energy consumption are inefficient. For example, I found that by improving the data transfer scheduling mechanism of professionally developed popular mobile apps such as Facebook and Pandora, their radio energy consumption can be reduced by up to 30% [**?**].

My research is dedicated to address both challenges, aiming at *providing practical, effective, and efficient methods to monitor and to reduce the resource utilization and bandwidth consumption in cellular networks.* I adhered to the following principles in the course of my research.

- **The measurement observations should be representative.** We collaborated with a commercial cellular ISP (AT&T) in the U.S., and collected cellular data of hundreds of thousands users from AT&T's cellular core network, to ensure the representativeness of our observations drawn from the data.
- **The solution should be general** in that it attacks the fundamental limitations of cellular networks. In fact, my proposed methodologies are directly applicable to any type of cellular networks including 2G GPRS/EDGE, 3G UMTS/HSPA, and 4G LTE networks that employ similar core resource management policies.
- **The solution should be practically deployable.** The ARO [**?**] and TOP [**?**] systems (described later) do not require any change to the cellular infrastructure. In particular, I am excited to see my ARO prototype [**?**] being productized by AT&T and now available to developers [**?**]. For resource inefficiencies found in popular smartphone applications such as Pandora and Facebook [**?**, **?**], we have contacted the corresponding developers, and the responses were encouragingly positive (watch the YouTube video [**?**] of comments from the Pandora CTO).
- **The underlying concept should have even longer-term impact.** My proposed frameworks of cross-layer analysis [**?**] and cooperative resource management [**?**] provide insightful guidelines for analyzing and optimizing general wireless network systems.

In the remainder of the statement, I detail the four aspects of my existing research, part of which has been reported on the AT&T Labs Research website [**?**], and sketch my future research plan.

Measuring the State of the Art: Characterizing Radio Resource Utilization for Cellular Networks

Understanding the current resource utilization for commercial cellular networks is the very first necessary step towards optimizing them. To achieve this goal, we collected cellular data of hundreds of thousands of 3G users from AT&T's cellular core network, then replayed the network traces against a novel RRC (Radio Resource Control) state machine simulator to obtain detailed statistics about radio resource utilization. To the best of my knowledge, my work is the first empirical study that investigates the optimality of cellular resource management policy using real cellular traces.

In a cellular system, a handset can be in one of several RRC states (*e.g.,* a high-power state, a low-power state, and an idle state), each with different amount of allocated radio resources. The state transitions also have significant impact

on the cellular network and the handset energy consumption: state promotions (resource allocation) incur signaling load and state demotions (resource release) are controlled by critical inactivity timers.

The RRC state machine is the key for cellular resource management but it is hidden from the mobile applications. This motivated me to design algorithms to accurately infer it through a light-weight probing scheme, then systematically characterize the impact of operational RRC state machine settings. The key observation is that the radio resource utilization is surprisingly inefficient: up to 45% of the occupation time of the high-speed transmission channel is wasted on the idle time period matching the inactivity timer value, which is called *tail time*, before releasing radio resources [**?**]. I further explored the optimal state machine settings in terms of several critical timer values evaluated using real network traces. My findings revealed that the fundamental limitation of the current cellular resource management mechanism is its *static nature of treating all traffic according to the same RRC state machine*, making it difficult to balance tradeoffs among the radio resource usage efficiency, the signaling load, the handset radio energy consumption, and the performance. Such an important observation drove me to delve into the optimization of cellular resource utilization described below.

Exposing the Visibility: Profiling Smartphone Apps for Identifying Inefficient Resource Usage

From cellular customers' perspective, as mentioned before, there remain far more challenges associated with mobile applications compared to their desktop counterparts, leading to smartphone applications that are not cellular-friendly, *i.e.,* their radio channel utilization and handset energy consumption are inefficient because of a lack of transparency in the lower-layer protocol behavior. To fill such a gap, I developed a novel data analysis and visualization framework called ARO (mobile **A**pplication **R**esource **O**ptimizer) [**?**]. ARO is the first tool that *exposes the cross-layer interaction* for layers ranging from higher layers such as user input and application semantics down to the lower protocol layers such as HTTP, transport, and very importantly radio resources. Correlating behaviors of all these layers helps reveal inefficient resource usage due to a lack of transparency in the lower-layer protocol behavior, leading to suggestions for improvement.

One key observation is that, from applications' perspective, given the aforementioned static nature of the current resource management policy, the low resource efficiency in cellular networks is fundamentally attributed to *short traffic bursts* carrying small amount of user data while interleaved with long idle periods during which a handset keeps the radio channel occupied. ARO employs a novel algorithm to identify them and to distinguish which factor triggers each such burst, *e.g.,* user input, TCP loss, or application delay, by synthesizing the cross-layer analysis results. Discovering such triggering factors is crucial for understanding the root cause of inefficient resource utilization.

ARO revealed that many popular applications (Pandora, Facebook, Fox News *etc.*) have significant resource utilization inefficiencies that are previously unknown. For example, for Pandora, a popular music streaming application on smartphones, due to the poor interaction between the RRC state machine and the application's data transfer scheduling mechanism, 46% of its radio energy is spent on periodic audience measurements that account for only 0.2% of received user data. Improving the data transfer scheduling mechanism can reduce its radio energy consumption by 30%.

Enabling the Cooperation: Optimizing Radio Resource Usage Through Adaptive Resource Release

I have investigated the resource optimization problem from perspectives of the network and customer applications, respectively. As mentioned before, analyses from both sides indicate the resource inefficiency origins from the *release* of radio resources controlled by static inactivity timers. The timeout value itself, known as the *tail time*, can last for more than 10 seconds, leading to significant waste in radio resources of the cellular network and battery energy of user handsets. Naively decreasing the timer is usually not an option because it may significantly increase the signaling load.

Therefore, to eliminate tail times, we need to change the way resources are released *from statically to adaptively*. This requires the cooperation between the network and handsets, since the latter have the best knowledge of application traffic patterns determining resource allocation and release. Towards this goal, I proposed **T**ail **O**ptimization **P**rotocol (TOP), a cooperative resource management protocol that eliminates tail times [**?**]. Intuitively, applications can often accurately predict a long idle time. Therefore a handset can notify the network on such an imminent tail, allowing the network to *immediately* release resources. However, doing so aggressively may incur unacceptably high signaling load. TOP employs a set of novel algorithms to address this key challenge by *(i)* letting individual applications predict tails and *(ii)* designing an efficient and effective scheduling algorithm that coordinates tail prediction of concurrent applications. The handset requests for immediate resource release only when the combined idle time prediction across all applications is long.

Interestingly, I found that the basic building block for realizing TOP is already supported by most cellular networks. It is a recent proposal of 3GPP specification called fast dormancy [**?**], a mechanism for a handset to request for an immediate RRC state demotion. TOP thus requires no change to the cellular infrastructure or the handset hardware given that fast dormancy is widely deployed. The experimental results based on real AT&T traces showed that with reasonable prediction accuracy, TOP saves the overall radio energy (17%) and radio resources (14%) by reducing tail times by up to 60%. For applications such as multimedia streaming, TOP can achieve more significant savings of radio energy (60%) and radio resources (50%).

Reducing the Footprint: Eliminate Redundant Data Transfers in Cellular Data Networks

Another important topic in cellular networks is to reduce the amount of data transferred without compromising the application semantics. Compared to wired and Wi-Fi networks, this issue is particularly critical in cellular networks. From carriers' perspective, cellular networks operate under severe resource constraints. Even a small reduction of the total traffic volume by 1% leads to savings of tens of millions of dollars for carriers [**?**]. The benefits are also significant from customers' perspective, as fewer network data transfers cut cellular bills, improve user experience, and reduce handset energy consumption.

There are multiple ways to achieve this goal, such as caching, compression, and offloading transfers to Wi-Fi. I have led the first network-wide study of HTTP caching on smartphones in collaboration with my colleagues, because HTTP traffic generated by mobile browsers and smartphone applications far exceeds any other type of traffic. Also caching on handsets (compared to caching in the network) is particularly important as it eliminates all network-related overheads.

Our study focuses on redundant transfers caused by *inefficient handset web caching implementation* [**?**]. We used a dataset collected from 3 million smartphone users of AT&T, as well as another five-month-long trace contributed by 20 smartphone users at the University of Michigan. Surprisingly, our findings suggest that redundant transfers contribute 18% and 20% of the total HTTP traffic volume in the two datasets. Even at the scope of *all* cellular data traffic, they are responsible for 17% of the bytes and 9% of the radio resource utilization. As confirmed by our local experiments, most of such redundant transfers are caused by the smartphone web caching implementation that does not fully support or strictly follow the protocol specification, or by developers not fully utilizing the caching support provided by the libraries. Our finding suggested that improving the cache implementation on handsets will bring considerable reduction of network traffic volume, cellular resource consumption, handset energy consumption, and user-perceived latency, benefiting both cellular carriers and customers.

Other Research: Network Measurement and Computer Security

While my dissertation mainly focuses on making cellular systems more resource efficient, I have a wide range of interests in the areas of network measurement and computer security. For example, I have worked on examining key properties of TCP behavior observed on today's Internet [**?**], characterizing spam campaigns launched by botnets [**?**], building an unsupervised anti-spam system leveraging campaign signatures [**?**], and building an unsupervised anomaly detection system for popular applications used in a community [**?**]. Also I have been interested in computer graphics and visualization since I joined the IGST (image-guided surgery and therapy) lab at SJTU when I was an undergraduate student. I would like to actively pursue opportunities for inter-disciplinary research in the future.

Future Research Agenda

In the course of my research, I have noticed that understanding the underlying radio resource control mechanism and its implications helps balance the key tradeoffs in cellular data networks and improve the resource efficiency for mobile applications. In the near future, I am interested in further leveraging this guideline to make wireless systems more resource efficient, as well as identifying the new challenges of cellular network and smartphone applications.

The network: from 3G to 4G. Currently 3G (UMTS, EvDO, and HSPA) is the main-stream cellular access technology. In 2009, 4G LTE (Long Term Evolution) started entering the commercial markets and is available now in more than 10 countries with a fast-growing user base. Besides the higher bit rate and lower latency that significantly outperform those of 3G, LTE employs a more complex RRC state machine with DRX (Discontinuous Reception where the handset periodically wakes up to check paging messages and sleeps for the remaining time) enabled even when a handset is occupying the high-speed transmission channel, in order to save the energy. We have done preliminary analysis on understanding the RRC policy, the power model, and the impact of DRX on application performance in 4G LTE networks [**?**]. A more in-depth exploration is an important part of my future work.

The apps: from individuals to the full spectrum. Given the extreme popularity of smartphone applications, a critical missing piece of information needed by carriers is their efficiencies of resource usage (radio resource utilization, signaling load, and handset energy consumption), which may have little correlation with their bandwidth consumption. Based on my experience of developing ARO, I plan to build a real-time monitoring system that can be leveraged by cellular carriers, which already have infrastructures to capture packet data in their core networks, to monitor resource efficiencies for a wide range of applications used by millions of customers. For those resource-inefficient applications detected by the system, the carrier can contact their developers for improvement.

We face three challenges towards building such a monitoring system. First, unlike ARO that can obtain various types of data directly from handsets, the only type of data available to the monitoring system is the network packet traces, whose payload needs to be carefully mined to extract useful information. Second, since the data collection is completely passive, multiple applications can be simultaneously running on a handset. However, the RRC state transitions are

determined by the aggregated traffic of all applications. Therefore we need to separate the resource impact of each of the concurrent applications running on a handset. Third, as a real-time monitoring system, it should be well scalable with small computation and storage overhead.

The optimization techniques: from uniform to diverse. Besides handset-based HTTP caching, I plan to pursue other directions for reducing the amount of data transferred in cellular networks, such as efficient compression, delta encoding [**?**], and the offline application cache provided by HTML5, which is expected to be supported by almost all smartphones by 2013 [**?**]. There remain three main challenges: *(i)* how to select the most effective technique for a particular content type or content provider, *(ii)* how to handle the complex interplay among multiple techniques when they are used together (if this brings additional benefits), and *(iii)* how to make the entire mechanism as transparent as possible from the application developers' perspective.