In Situ Network and Application Performance Measurement on Android Devices and the Imperfections

Mohammad A. Hoque  
University of Helsinki, Finland  
mohammad.a.hoque@helsinki.fi

Ashwin Rao  
University of Helsinki, Finland  
aswin.rao@helsinki.fi

Sasu Tarkoma  
University of Helsinki, Finland  
sasu.tarkoma@helsinki.fi

ABSTRACT
Understanding network and application performance are essential for debugging, improving user experience, and performance comparison. Meanwhile, modern mobile systems are optimized for energy-efficient computation and communications that may limit the performance of network and applications. In recent years, several tools have emerged that analyze network performance of mobile applications in situ with the help of the VPN service. There is a limited understanding of how these measurement tools and system optimizations affect the network and application performance. In this study, we first demonstrate that mobile systems employ energy-aware system hardware tuning, which affects application performance and network throughput. We next show that the VPN-based application performance measurement tools, such as Lumen, PrivacyGuard, and Video Optimizer, aid in ambiguous network performance measurements and degrade the application performance. Our findings suggest that sound application and network performance measurement on Android devices requires a good understanding of the device, networks, measurement tools, and applications.

1 INTRODUCTION
In situ Internet traffic measurement tools, such as Video Optimizer (VoP) [10], Lumen [11], PrivacyGuard (PvG) [12], and MopEye [14], are essential for debugging, improving user experience, and performance comparison of mobile applications. The alternative is rooting the device and using tcpdump for offline analysis.

The above traffic measurement tools shed light on the network and application performance. However, they may also contribute to imperfect and ambiguous results, as we might measure something which we do not intend to measure. Studying the sources of these imperfections is vital to calibrate the measurement procedures and to improve the tools. At present, there is a limited understanding of the impact of in situ mobile Internet traffic measurement tools and how device hardware optimization affects the network and application performance.

In this work, we quantify the performance impact of system hardware optimization and also evaluate the impact of VoP, Lumen, and PvG on network performance metrics, and application traffic. We focus on these three applications, as they exemplify state-of-the-art traffic measurement and analysis tools. These tools have similar designs and use the Android VPN interface. However, they do not route the traffic to a remote VPN server. VoP [4], formerly known as ARO [10], is a popular open-source tool for collecting traffic from mobile devices without rooting the device, and it also enables various diagnosis and optimization of applications, network, CPU and GPU [4] through offline analysis. In contrast, Lumen and PvG are two online traffic analysis tool helping users to find privacy leaking incidents. Lumen also provides insights on the TLS usage of mobile applications [11], the CDN usage by mobile applications [9], and the DNS [6]. MopEye is another similar application. It is currently unavailable in the Google Play Store and also in popular source code hosting websites, such as GitHub.

This article investigates the imperfections in traffic measurements on Android devices due to system optimization and in situ traffic measurement tools. We demonstrate that sound Internet traffic measurement requires a thorough understanding of the device, tools, and applications. Note that we do not aim to establish whether a particular tool is the best or worst. Our key observations are as follows.

1. Mobile systems employ CPU and WiFi transmit power optimization triggered by the battery level. We observe that the CPU optimization techniques, such as CPU hot-plugging and dynamic frequency scaling, mostly affect network I/O, while WiFi optimization, i.e., dynamic modulation scheme, affects the uplink throughput. These optimizations deteriorate application performance and network throughput. Charging the device, when the battery level is below 20%, does not improve the network performance. Therefore, one must be aware of the adaptive performance characteristics of mobile devices while conducting experiments (Section 2).

2. Although it is expected that VPN-based tools would provide degraded network performance as the packets spend more time on the device [10–12], we may estimate ambiguous latency and throughput in the presence of the VPN-based tools. For example, in the presence of PvG, SpeedCheck [1] estimates on-device latency instead of the network latency. Similarly, VoP doubles the uplink throughput estimates. The sources of these ambiguities are the implementation of the measurement tools, as we present in Section 3. VoP also delays the outgoing traffic, and PvG delays the incoming traffic. Therefore, to avoid such pitfalls in network and application performance measurements, one must have a good understanding of these applications and tools.

3. Furthermore, all these VPN-based applications fail to apply the application intended optimization through socket options and thus affect the application performance, as we demonstrate for the outgoing TCP traffic in Section 3.

Finally, we summarize the sources of the above ambiguous or imperfect measurement results (Section 4).

2 IMPACT OF SYSTEM OPTIMIZATION
Android devices may come with advanced CPU governors that save energy by hot plugging and unplugging of CPU cores, as supported by modern Linux kernels [8]. Apart from workload characteristics, the devices may also consider the status of the battery to employ the CPU cores. We look into the impact of such off-the-shelf system optimization on network latency and throughput on Nexus 6.
During our measurements with Nexus 6, we have found that two of the four cores remain offline when the battery discharges to below 20%, and the active cores operate at the maximum frequency of 1.73 GHz. When the battery level is above 20%, all the four cores become active, and their maximum operating frequency increases to 2.65 GHz. Therefore, the battery level also prompts dynamic frequency scaling.

We performed the following measurements to quantify the impact of this optimization on the network traffic characteristics. Specifically, we used SpeedCheck [1] (paid) and measured the latency and throughput on Nexus 6 (Android 7.0) when the battery levels were above 20% and below 20%. We performed the measurements using both WiFi and LTE. Each of the above four scenarios was repeated ten times, and the results are presented in Figure 1. Figure 1 shows that while hot unplugging of CPU cores on Android has a negligible impact on the latency, its impacts on throughput is significant. The availability of additional CPU cores, when the battery level is above 20%, improves the I/O performance across the two access technologies, WiFi and LTE. Furthermore, WiFi uplink throughput improves almost four times when the battery level is above 20% compared to when it is below 20%. The closer inspections of the MAC layer frames revealed that WiFi radio of the Nexus 6 switches from 802.11ac to 802.11g mode when the battery level drops below 20%. These performance limiting optimizations also affected the device responsiveness for various applications, such as browsing and streaming. This also implies that modern Android devices adapt the physical layer mechanisms similar to the iOS devices¹ to avoid unexpected shutdown of the devices [13] and to improve battery life.

Figure 1 also depicts that the downloading speed of SpeedCheck over LTE doubles when the battery level is higher than 20%. Similar to WiFi, we further looked into the physical layer modulation scheme used by the mobile device in the LTE network. We rooted Nexus 6 and installed Network Signal Guru [5] that samples LTE physical layer parameters after every 500 ms. Figure 2 shows that the modulation schemes were always 16QAM (Quadrature Amplitude Modulation) and 64QAM for uplink and downlink, respectively, during the throughput measurements. The other attributes in the figure are discussed in section 6. Nexus 6 employs three optimization techniques, triggered by the battery level, which affect the network and application performance. Charging the device, when the battery level is below 20%, does not improve the throughput either on WiFi or LTE and application performance. The optimization may vary from device to device.

3 IMPACT OF MEASUREMENT TOOLS

3.1 In-situ Traffic Measurement Tools

The forwarder and the packet inspector are two components of the VPN-based in situ traffic measurement tools exemplified by VoP, Lumen, and PvG, as shown in Figure 3.

The primary role of the forwarder is to forward (i) the packets received from Android applications to the Internet, and (ii) the packets received from the Internet to the Android applications. The forwarder also copies those packets to the inspection queue to isolate traffic analysis from the path of the packet.

The forwarder essentially creates a new TCP socket on seeing a TCP SYN packet from the VPN interface. The forwarder in Lumen and VoP establish a socket connection with the remote server using connect() API before sending SYN-ACK to the application. PvG, on the other hand, establishes socket connection after replying with

---

¹https://www.forbes.com/sites/ewanspence/2017/12/20/apple-iphone-kill-switch-ios-degrade-cripple-performance-battery/
SYN-ACK. Later, we demonstrate how these implementations affect network performance measurements. The forwarder creates a new UDP socket when it detects a new UDP flow. These newly created sockets are protected so that packets from the newly created flows do not loop the tun interface [3].

A packet inspector is responsible for inspecting the packets in its queue. In the case of Lumen and PvG, the packet inspector performs the privacy analysis on the packets, whereas the VoP’s inspector sends packets to the desktop application.

In the later sections, we quantify the impact of VoP, Lumen, and PvG on (a) the network performance, and (b) the network characteristics of applications.

3.2 Addressing Biases

We took the following steps to ensure that the measurement results presented in the upcoming sections are not the artifacts of misconfigured tools and the measurement setup.

(i) Battery level. For the upcoming measurements, we ensured that the devices had more than 80% charge. This is because mobile devices might restrict resources based on the battery level, as we have shown in section 2.

(ii) Throughput throttling. VoP also offers to throttle downlink and uplink traffic. All the measurements in this paper were conducted without any throughput throttling.

(iii) Software Auto Update. During the experiments, application and the auto system updates were disabled on mobile devices.

(iv) Advertisements. We have purchased without ad subscriptions of SpeedCheck and SpeedTest to avoid any biases caused by the free versions.

3.3 Impact on Network Performance

This section explores the network performance using SpeedCheck [1] and SpeedTest [2]. These two applications work as the traffic load generator without any VPN-based tools and in the presence of the listed VPN applications. Without any VPN scenario gives the baseline performance. SpeedCheck connects to its servers in Germany, and SpeedTest connects to the servers in the LTE operator network within a few kilometers from the mobile device. The measurements were repeated ten times.

Figure 4: Impact on LTE network latency and throughput. We used SpeedCheck and SpeedTest on Nexus 6 in the presence of Lumen (Lum.), VoP, PvG, and Baseline, i.e., without any localhost VPN.

(1) Latency. Figure 4 (left) compares the network latency reported by two applications in the presence of the VPN-based tools. From the tcpdump traces, we have identified that SpeedTest uses 10-12 requests/responses of few bytes (less than 100 Bytes) over a TCP connection to estimate the latency. SpeedTest estimates the baseline latency of 16-18 ms. This is expected, as the server was located at the operator’s network. It experiences 3-5 ms additional latency in the presence of Lumen and PvG, whereas VoP increases the latency by three-fold. This is due to the energy optimization strategy adopted by VoP, which we discuss in the upcoming sections.

In contrast, SpeedCheck reports the median baseline network latency of about 45 ms. From the corresponding tcpdump traces, we have identified 10 empty and consecutive TCP flows (without any data exchange) for each latency measurements. These flows suggest that SpeedCheck uses TCP connect() API to measure the latency. Both VoP and Lumen increase the median latency significantly. We speculate that these two take more time to set up new TCP flows. However, SpeedCheck underestimates the latency in the presence of PvG, which is the consequence of the sending SYN-ACK by the PvG forwarder before the connection is established with the remote server, as discussed in Section 3.1.

(2) Uplink Throughput. Figure 4 (center) depicts that SpeedTest estimates higher uplink baseline throughput, as the server is in the LTE operator network. It uses multiple parallel TCP connections to estimate the throughput. Both Lumen and PvG reduce the throughput of SpeedTest/SpeedCheck by half compared to the baseline measurements. However, Lumen severely affects the uplink throughput measurements of the SpeedCheck. It uses a single TCP connection and sends a large amount of data. From an exception in the debug log, we characterized that Lumen’s forwarder cannot handle such volume. Interestingly, VoP doubles the uplink throughput of both applications.

(3) Downlink Throughput. Figure 4 (right) demonstrates that SpeedTest measures similar downlink throughput in the presence of the VPN tools to the baseline. Lumen aids the highest throughput measurements with SpeedCheck. However, VoP and PvG degrade the throughput of SpeedCheck significantly.

The typical network measurement tools, such as SpeedCheck and SpeedTest, can have different methods to estimate the latency
and throughput. While their baseline estimates are reasonable, their estimates vary according to the implementation of the VPN tools.

3.4 Impact on Realtime Application (UDP)

In this section, we investigate the traffic from three realtime applications; IMO, WhatsApp, and Skype. The versions of the apps used are presented in Table 1. While these applications fall into the broad category of messaging applications, their varying traffic characteristics help us to study the impact of the design of VoP and Lumen. We could not use these applications in the presence of PvG in several trials. We used a rooted Nexus 6 (Android 7.0) and a non-rooted LG G5 (Android 8.0) for these measurements.

These apps exchange bi-directional encrypted UDP traffic. The conversations were two minutes long over LTE, and we ran 3 iterations in each of the following scenarios. We investigate their inter-packet gaps and bitrates. As the baseline, we initiated conversations between Nexus 6 and LG G5 using these apps without VoP or Lumen and captured traffic using tcpdump on Nexus 6. We then repeated the experiments with VoP running on Nexus 6 and collected traffic from VoP. Finally, we used Lumen. Since Lumen does not store traffic, we captured traffic with tcpdump on Nexus 6.

**Baseline Results.** Figure 5(a) shows that IMO has the highest inter-packet gaps, and Skype packets have the smallest gaps. These apps also have distinct data rates with Skype having the highest data rate, as shown in Table 1.

**Impact of VoP.** Compared to the baseline packet-gaps in Figure 5(a), VoP significantly alters the inter-packet gaps of outgoing UDP packets, as shown in Figure 5(b). Most of the outgoing packets across all applications have an inter-packet gap of about 100 ms. In contrast, the incoming packets have had similar distributions to the baseline. This delay is similar to the latency measurements with VoP discussed earlier. Table 1 shows that the outgoing data rates of Skype and Whatsapp reduce significantly, which we speculate to be a consequence of the delays introduced by VoP.

**Impact of Lumen.** Figure 5(c) shows that with Lumen the inter-packet gaps of the outgoing packets are similar to the baseline measurements. Besides, the applications experience similar bitrates to the baseline and when using Lumen as shown in Table 1.

3.5 Impact on Realtime Application (TCP)

We used Periscope (v1.24) to study the impact of VoP and Lumen on realtime TCP flows. Periscope’s live broadcast did not work in the presence of PvG. Periscope broadcasts over LTE across three different scenarios. We capture traffic on Nexus 6 using tcpdump for baseline and Lumen scenarios.

Similar to our observations for UDP traffic, we observed 100 ms inter-packet gap, as shown in Figure 6 (left). From the distribution of packet size in Figure 6 (right) (collected by VoP), we notice that more than 70% packets captured by VoP are larger than 1500 bytes. From Traffic traces, we have identified that VoP creates packets of a maximum of 65549 bytes for Periscope, and the uplink throughput measurements flow from SpeedCheck.

From the source code in Github, we have identified that VoP forwarder implements the maximum segment of 65535 bytes for the TCP flows. It accumulates traffic from the client application, and the segments reach the maximum size very quickly with very
Moorthy et al. - Mobile System Optimizations Affect Downlink and Uplink Throughput Measurements, Infocom 2017

3.6 Analysis with Socket Options

In this section, we investigate the performance of the VPN-based tools in processing the flows with TCP_NODELAY (Nagle’s algorithm) socket option on Nexus 6. We specifically look into this option, as it has a direct impact on the local delay and thus affects the performance of web browsing and other realtime applications, such as live broadcasting, crypto/stock exchange applications, on mobile devices. We developed a separate traffic generating application that creates two blocking TCP sockets enabled and disabled Nagle’s algorithm. The application sends 1300 bytes data over LTE after every 20 ms to a remote server at the university campus. The application also receives data from the remote server after every 20 ms in separate TCP sessions.

Performance of VPN-based Tools. Figure 7(a) compares the outgoing inter-packet gap of the application flows; having Nagel’s algorithm enabled and disabled. When Nagel’s algorithm is enabled, more than 70% of the packets sent from the application have more than 20 ms delays at the network layer. In the presence of VPN applications, disabling Nagel’s algorithm by the application does not improve the delay compared to the baseline (Figure 7(b)). Interestingly, VoP’s packet gap reduces, as it receives packets from the local TCP/IP stack without delay. From traffic traces, we have identified that these VPN-based tools do not disable Nagel’s algorithm while establishing socket connections.

Figure 8 shows the performance of the VPN applications for incoming traffic. The application receives data at almost similar gaps observed at the network interface. However, in the presence of PPG, the application receives 40% packets at late. The packet-gaps patterns suggest that it uses a fixed interval to read the VPN interface similar to VoP.

The investigations in this section reveal that the VPN-based tools do not set the TCP/IP socket options as intended by the other user applications. Consequently, they can misguide the developers and degrade application performance. For example, SpeedTest disables Nagel’s algorithm or sets the TCP_NODELAY socket option to send tiny packets to measure the network latency. Findings in this section explain the higher latency experienced by SpeedTest in Section 3.3.

4 SOURCES OF IMPERFECTION

Mobile system optimizations affect downlink and uplink throughput, whereas the VPN-based tools mostly affect the uplink throughput and latency, i.e., they mostly affect the outgoing traffic. In this section, we summarize the sources of such measurement results.

Energy-Aware Optimization. Energy-aware system optimization can affect the network performance by limiting the network I/O and by applying adaptive modulation schemes. Therefore, it is wise to perform such measurements when the battery is fully charged. VoP, Lumen, and PPG rely on different sleeping techniques to optimize their energy usage. The additional latency introduced by VoP on outgoing packets is the artifact of using a fixed sleep interval of 100 ms in the main VPN thread. This delay further contributes to the large outgoing packets for higher bitrate uplink traffic and energy consumption for fragmentation. PPG also introduces a fixed delay for the incoming traffic. Regardless, these delays affect not only the quality of the measurements but also the quality of experience when using other user applications.

Forwarder. In situ VPN-based measurement tools are middle-boxes that tap the packets using the VPN interface. These applications, therefore, implement a forwarder which primarily consists of three threads: the main VPN thread, and two-socket reader/writer threads. The reader/writer threads continuously iterate through a list of live sockets, which contributes to the delays. The forwarder also implements a flow state machine for each flow and constructs/de-constructs the packets. The implementation of the forwarder affects the latency and throughput measurements. We have also shown that the characteristics of the newly created flows and their packet headers might not be the same as those generated by the applications. The reason is that the socket options must be set before the connection establishment.
6 LTE RADIO RESOURCE ALLOCATION

In LTE networks, Physical Resource Block (RB) is considered as the unit of the radio resource. With 5 MHz bandwidth, there are 25 RBs. In an RB, there are 12 sub-carriers in the frequency domain. Each of the RBs can have either $7 \times 12$ or $14 \times 12$ resource elements (REs), where 7 and 14 are the symbols, in the time domain, over 0.5 and 1 ms respectively using normal cyclic prefix (CP) [15].

Now the amount of bits an RB can carry depends on the channel quality indicator (CQI) notification from the UE. Essentially, each CQI maps to a modulation and coding scheme according to Table 2. CQI indicates not only the channel quality but also a device’s capability whether the device can receive data of a particular modulation and coding scheme or not.

The equations to compute the number bits an RB can hold for a certain CQI, and the number of RBs is required by an eNB to transmit a packet can be expressed as the followings.

$$\text{RB}_{\text{bits}} = \text{RE}_{\text{bits}} \times n \times t_s = C_{\text{CQI}} \times M_{\text{bits}} \times n \times t_s \quad (1)$$

In equation 1, $M_{\text{bits}}$ is the bits for a modulation scheme, $n$ is the number of usable REs, and $t_s$ is the duration of time slot (0.5 or 1 ms).

$$\text{RB}_n = (\text{PacketSize}_{\text{bits}} + \text{RLC}_{\text{bits}} + \text{MAC}_{\text{bits}}) / \text{RE}_{\text{bits}} \quad (2)$$

| CQI | Modulation | Real Bits ($N_m$) | $C_{\text{CQI}} = N / 1024$ |
|-----|------------|-------------------|-------------------------------|
| 1   | QPSK       | 78                | 0.0762                        |
| 2   | QPSK       | 120               | 0.1171                        |
| 3   | QPSK       | 193               | 0.1884                        |
| 4   | QPSK       | 308               | 0.3                            |
| 5   | QPSK       | 449               | 0.4384                        |
| 6   | QPSK       | 602               | 0.5879                        |
| 7   | 16QAM      | 378               | 0.3691                        |
| 8   | 16QAM      | 490               | 0.4785                        |
| 9   | 16QAM      | 616               | 0.6015                        |
| 10  | 16QAM      | 466               | 0.4550                        |
| 11  | 64QAM      | 567               | 0.5387                        |
| 12  | 64QAM      | 666               | 0.6503                        |
| 13  | 64QAM      | 772               | 0.7539                        |
| 14  | 64QAM      | 873               | 0.8525                        |
| 15  | 64QAM      | 948               | 0.9258                        |

Table 2: Channel Quality Index (CQI), Modulation Scheme, and Coding Rate mapping [7].

Figure 9 shows the usage of the Modulation Scheme and the number of resource blocks for a large file download on Nexus 6 with CQI11. LTE supports QPSK, 16QAM, and 64QAM, i.e., each RE can carry a maximum of 2, 4, and 6 bits accordingly. Let us consider the duration of 1 RB is 1 ms ($t_s$), and there are 168 REs. Nevertheless, mostly 120 REs ($n$) are available for carrying data traffic. For CQI11, the modulation scheme is 64QAM and the effective code rate $C_{\text{CQI}} = N_m / 1024 = 0.55$. Therefore, an RE can hold only, $\text{RE}_{\text{bits}} = C_{\text{CQI}} \times M_{\text{bits}} = 0.55 \times 6 = 3.32$ bits and an RB can hold $n \times \text{RE}_{\text{bits}} = 398$ bits.

The number of RBs required for a packet in a downlink can be computed using equation 2 by considering the additional bits for...
RLC and MAC headers. However, the network may not allocate the RBs according to the CQI. It may have other complex resource scheduling algorithms, as it has to deal with various types of traffic and users. The number of uplink RBs also may vary.

7 APPLICATION

```c
int val = 1;
// Disabling Nagel’s Algorithm
setsockopt(sockfd, SOL_TCP, TCP_NODELAY, &one, sizeof(one));
if (connect(sockfd, &servaddr, sizeof(servaddr)) < 0)
    LOGE('"***Server Connect Error***");
for (int i = 0; i < 5000; i++) {
    usleep(20000);
    char *dat = rand_string(1300);
    gettimeofday(&tv, NULL);
    times[i] = (tv.tv_sec*1000000LL+tv.tv_usec)/1000;
    n = write(sockfd,dat, 1300);
    if (n < 0){
        LOGE("Error sendto %s", strerror(errno));
        break;
    }
}
```

Listing 1: TCP sending code with/without Nagel’s algorithm.

```c
int BUFSIZE = 4096;
if (connect(sockfd, &servaddr, sizeof(servaddr)) < 0)
    LOGE('"***Server Connect Error***");
while (true) {
    bzero(buf, BUFSIZE);
    n = read(sockfd, buf, BUFSIZE);
    if (n > 0) {
        gettimeofday(&tvo, NULL);
        times[i]=(tvo.tv_sec*1000000LL+tvo.tv_usec)/1000;
        i = i+1;
    } else 
        break;
    if (i==5000)
        break;
}
```

Listing 2: TCP receiving code.