Understanding the Characteristics of Cellular Data Traffic

Ying Zhang
Ericsson Research
San José, CA., U.S.A.
Ying.Zhang@ericsson.com

Åke Arvidsson
Ericsson Research
Stockholm, Sweden
Ake.Arvidsson@ericsson.com

ABSTRACT

Because of rapidly growing subscriber populations, advances in cellular communication technology, increasingly capable user terminals, and the expanding range of mobile applications, cellular networks have experienced a significant increase in data traffic, the dominant part of which is carried by the http protocol. Understanding the characteristics of this traffic is important for network design, traffic modeling, resource planning and network control. In this study we present a comprehensive characterization study of mobile http-based traffic using packet level traces collected in a large cellular network. We analyze the traffic using metrics at packet level, flow level and session level. For each metric, we conduct a comparison between traffic from different applications, as well as comparison to traffic in a wired network. Finally, we discuss the implications of our findings for better resource utilization in cellular infrastructures.

Categories and Subject Descriptors

C.2 [**Computer-Communication Networks**]: Network Operations

General Terms

Measurement

Keywords

Traffic, measurement, comparison, wireline, wireless, cellular, fixed

1. INTRODUCTION

Because of the emergence of user-friendly smartphones and the advances in cellular data network technologies, the volume of data traffic carried by cellular networks has been experiencing a phenomenal rise. One large cellular operator has reported a growth of 8000% of cellular data traffic over the past four years [2] and it is expected to grow to 10.8 exabytes per month by 2016, an 18-fold increase over 2011. To cope with this explosive growth and best serve their customers, operators need to have a better understanding of the nature of traffic carried by cellular networks.

Among all the cellular data traffic, different "web services" (in the sense that they are carried over http) is the dominant category due to the increasing popularity of video streamimg (*e.g.,* YouTube), social networking (*e.g.,* Facebook) and file sharing (*e.g.,* Sourceforge). Web traffic in this sense has thus undergone a significant change over the last few years, from static displays of simple html pages to rich and complex media applications. As a consequence, the nature of the web traffic has changed as well.

Thus, although a large amount of work on traffic characteristics [4, 5, 14, 19] was done a decade ago, they apply techniques developed for static pages which may no longer be suitable. Moreover, there are few and limited studies for cellular networks. Prior work on understanding cellular data traffic is largely dated, *e.g.,* [11], or limited to flow level statistics, *e.g.,* [21, 18] with an emphasis on on understanding traffic patterns from different device types or applications.

In contrast, the goal of this paper is to provide an initial understanding and categorization of cellular data traffic, with emphasis on basic traffic statistics, differences between wireline and wireless accesses, and user behavior patterns. We seek answers to questions like "Are there any significant differences in traffic patterns across different services?", "What are the unique characteristics of cellular data traffic compared to fixed data traffic?", and "Are the usage pattern different across applications and across network access types?". This paper is the first study to provide answers to these types of questions about web traffic in large cellular networks. The answers to these and similar questions can be used for modeling traffic, improving performance, optimizing distribution and designing systems such as firewalls and intrusion detectors.

We report on a set of comprehensive studies of traffic collected in a large cellular network and in a small fixed network. We make the following three contributions. First, we study traffic characteristics for a wide range of services and find that different applications impose different demands on network resources on packet level, flow level and session level. Second, we compare these results to those in a fixed network with a focus on uncovering the unique patterns induced by the different network accesses. Third, we analyze the usage behavior with comparisons both across applications and across network types. Comparing wireless and wireline sessions, it is found that the former typically contain less data in more but shorter flows, and that they typically consist of smaller packets with burstier arrival patterns.

The remainder of this paper is organized as follows. We first introduce the metrics we use for characterizing traffic patterns in Section 2. We also present the methodology we have developed to identify these metrics as well as the analysis techniques. Section 3 describes our data sets and presents our results. Related work is

discussed in Section 4. Section 5 outlines some implications and conclude our study.

2. METHODOLOGY

In this section, we will describe the metrics we use and the data analysis methodologies.

2.1 Metrics

The list of possible features one could consider is very large. We can broadly classify these into three categories:

Packet level statistics: Examples of packet level features include sizes and inter-arrival times of packets. Such characteristics are simple to compute as they can be gleaned directly from a trace. Moreover, they offer a characterization of the application that is independent of the notion of flows, sessions or other higher level aggregations.

Flow level statistics: A flow is defined as a unidirectional sequence of packets that have some common property, typically the values contained in the 5-tuple (source IP, destination IP, source port, destination port, IP Protocol type). Examples of flow level features include the distributions of durations, data volumes and packet numbers of flows. We can also compute the packet level statistics of packets within flows. In particular, we can characterize the burstiness of a flow by the relative variance of the inter-arrival times between packets of the flow in question.

Session level statistics: Some interesting characteristics can be captured only by considering statistics across multiple flows. For example, many http objects embed other objects with different URLs also from different sites which may be downloaded over different TCP connections. Therefore, considering a set of flows belonging to the same session allows a closer representation of user experience.

2.2 Extracting Sessions

Sophisticated web services today often include many static and dynamic objects. For example, in response to a request, a browser typically first downloads a main html object that defines the structure of the page, followed by a cascading style sheets object that describes the presentation of the page and a number of embedded objects such as images or advertisements which in turn also may contain embedded objects.

To extract complete sessions ("web pages") from a packet trace we first parse the dump and extract the http related packets to construct http flows. Then we group flows that belong to the same page download process to form sessions.

In more detail we first parse the packet payload to obtain information related to the http protocol itself such as the url, agent, objects, and response code. We then apply two rules to group flows that belong to the same service request.

- The first rule is based on content. By parsing html pages, we gather lists of objects embedded in the current page. Then, we search for these objects among subsequent flows and, once a match is found, the flow is grouped with the other flows of that request.

- The second rule is based on time. In some cases, requests are generated by locally executed Java scripts and in other cases http pages are encrypted. To handle these and other, similar scenarios, we use a heuristic timeout to group flows as follows. If two flows between two specific IP addresses are seen within a short time interval, it is likely that they belong to the same http request. We have tested different thresholds and

found that the results are not that sensitive to the exact value and settled for 0.5 seconds in our experiments.

Our algorithm was compared to another, independently designed and implemented one, and it was found that the two produced very similar results. Nevertheless, it is clear that the results will contain errors due to, *e.g.,* encryption and dynamically generated requests.

2.3 Clustering Applications

Next, we identify similarities and differences across different applications.

Identify similarities using clustering: In order to group applications with similar traffic patterns, we choose clustering, an unsupervised learning mechanism wherein unlabelled training data is grouped based on similarity. We selected the K-means algorithm [12] because it is one of the quickest and most simple. The algorithm partitions objects in a data set into a fixed number K of disjoint clusters. A small distance between two objects implies a strong similarity whereas a large distance implies a low similarity.

Identify difference using statistical tests: To determine differences between distributions we use MWU, the Mann-Whitney U-test [13]. MWU is a non-parametric method that measures whether two distributions are equal based on the ranks of the samples, *i.e.,* the relative location in a merged, ordered vector. It determines if the two distributions are equal or there exists significant differences. We choose the U-test because it is efficient and accurate when the number of samples is relatively small.

2.4 Analysis Steps

In summary, our analysis contains three dimensions: access media types, comparison metrics, and applications, cf. Figure 1. To conduct our analysis systematically, we are first guided by the two goals, *i.e.,* comparing access media types and identifying differences between cellular and wireline web traffic. To this end, we apply the same filter on both cellular and wireline data. The filter first selects a target metric M_i to evaluate, and then picks an application or a group of applications A_i. Finally, we use statistical tests to detect differences between the two distributions.

To detect difference across wireless apps, we first cluster applications into groups using both an automatic clustering algorithm and manual grouping based on service type. The purpose is to examine if there exist any groups of applications with similar performance properties. After creating groups, we select one pair of applications from the same group and one from a different group. Next we examine each metric with respect to distribution differences. The logic behind this analysis is to first identify high level differences such as performance differences. Then we search for explanations by detailed low level traffic characteristics using different metrics. At the last step, we also verify the explanation by checking that other cases where these low level differences occur exhibit similar high level differences.

3. EXPERIMENTAL RESULTS

3.1 Data Sets

We have used two data sets collected during approximately the same period. The *wireless* data set was collected on a Gn interface between a Gateway GPRS Support Node (GGSN) and a Serving GPRS Support Node (SGSN) in a cellular network. It contains 168 hours of data with 3.9 million sessions from 50k IP addresses. The *wireline* data set was collected at the head end of a municipal network. It contains 31 hours of data with 19 million sessions from 21k IP addresses. Both sets contain full-size packets
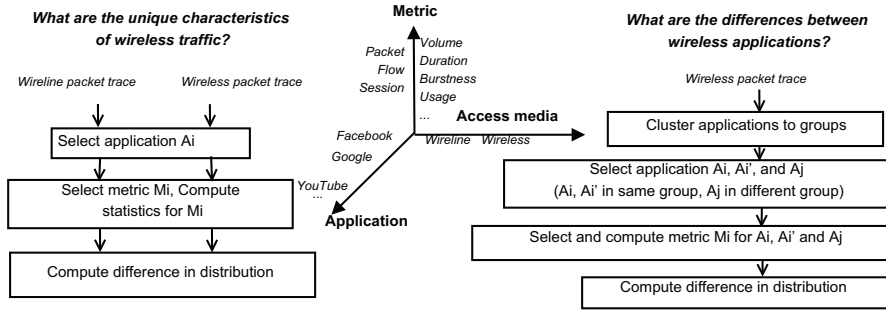
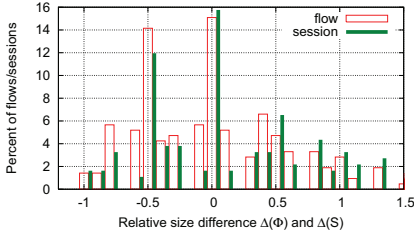Figure 1: Analysis logic.

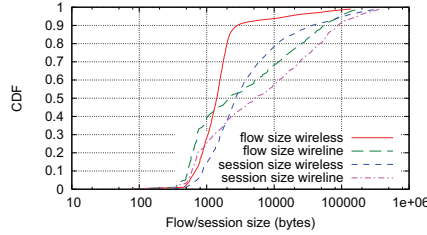Figure 2: Distributions of flow/session sizes for wireless and wireline.

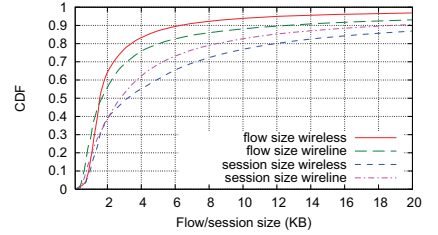Figure 3: CDF of flow/session sizes for Flickr.

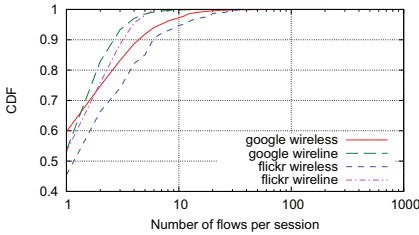Figure 4: CDF of flow/session sizes for Google.

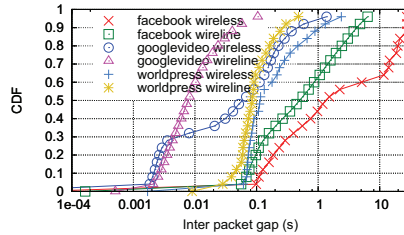Figure 5: Number of flows per session for Google and Flickr (wireless/wireline)

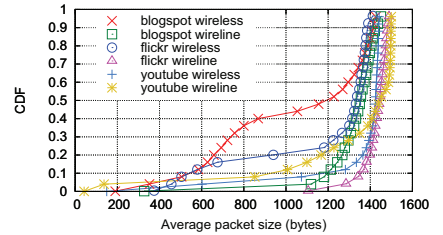Figure 6: Average inter-packet gaps for Facebook, Google Video and Worldpress.

Figure 7: Average packet size packet sizes for Blogspot, Flickr and YouTube

to enable reconstruction of all http transactions, but we have only had direct access to anonymized results where individual users cannot be traced. While both sets are very large, the results may not be representative of other networks, countries or time period, depending on, *e.g.,* the combinations of devices, applications, tariffs and preferences. We start by using the two-level domain identifiers to identify an origin; *e.g.,* the two URLs *profile.facebook.com* and *www.facebook.com* are considered to be the same two-level domain identifier *facebook.com*.

3.2 Differences between Cellular and Wireline Traffic

3.2.1 Content Differences

Because of the limited bandwidth of wireless links and the capacity of mobile devices, one may expect that the content sent to mobile devices is different from the content sent to PCs; mobile content may, *e.g.,* fit a more compact device, consume less bandwidth or require less computation. We first examine the correctness of this hypothesis.

For each application, we compute the average flow size Φ in bytes for both types of access, Φ_{wire} and Φ_{cell} respectively, after which we compute their relative difference:

$$\Delta(\Phi) = \frac{\Phi_{\text{wire}} - \Phi_{\text{cell}}}{\Phi_{\text{wire}}}.$$

Similarly, we compute the average session size S in bytes per application for both types of access, S_{wire} and S_{cell} respectively, and then their relative difference $\Delta(S)$.

Figure 2 shows the distributions of the relative differences. We observe that for the majority of applications, about 70%, the amount of data indeed is larger for wireline than for wireless. However, for around 15% of the applications we note that the amount of data is the same for both accesses (the relative differences are zero) and for the last 15% of the applications the amount of data volumes is, in fact, larger for wireless than for wireline. Wireline traffic tends to use larger *flows* but this may be deceiving since the differences are smaller when measured over complete *sessions*.

To further highlight these observations, we consider two example applications *viz.*. the photo sharing site Flicker in Figure 3 and

the search engine Google in Figure 4. Both figures show the distributions of the amount of data per flow and session. Flicker shows expected results: more data is delivered per flow and per session with wireline access than with wireless access, and this difference is more obvious for flows. For Google, however, we observe two opposite trends; more data per flow with wireline access but more data per session with wireless access. This is because Google uses more flows per session in wireless, as can be seen in Figure 5. Among the 100 applications we test, 82 have similar observations, suggesting that it is quite common to use more flows in parallel for wireless access. This may be an attempt to provide resilience against single TCP time outs caused by variable delays due to link layer retransmissions.

3.2.2 Traffic Differences

Besides customizing the content and the presentation for mobile users, service providers can also optimize other traffic characteristics for mobile access. Below we focus on packet inter-arrival times and packet sizes.

For each flow, we compute the average of the inter-arrival time between any two consecutive packets, and then plot the distribution per application. Three examples are given in Figure 6, *viz.* Facebook, Google Video, and Worldpress, as they are among the most popular web sites and at the same time exhibit different traffic characteristics. Google Video and Facebook both exhibit much larger packet gaps over wireless access compared to wireline access, but this difference is negligible for Worldpress. In particular we note that the gaps for Facebook are surprisingly large (50% of them are larger than 1 second). We calculated the network delay from TCP SYN/ACK pairs at the same time as these delays were noted and found that this indeed is significantly less. The cdf of inter-packet gaps would be step shaped for "smooth" traffic (packets being equally spaced) but gently sloping for "bursty" traffic (some packets close to each other and other packets far apart) and conclude that traffic in cellular networks tends to be more bursty. This may be an inevitable result of the radio access but it has consequences for the required margins in the transmission network.

As for packet sizes, we note that links are associated with maximum transfer units (MTUs) and packets which exceed the MTU of a link will be fragmented or dropped by the router depending on the setting of the IP fragmentation flags. Noting that traffic in cellular networks is tunnelled between GGSNs and SGSNs/RNCs using the GTP protocol, and that this adds overhead (*i.e.,* the GTP layer and the outer IP layer) to the packets, it may be suspected that packets the size of which is close to the MTU of, *e.g.,* Ethernet, could exceed this size when tunnelled in cellular networks. We examined this problem and found that 87% of the applications have the "Don't Fragment" bit set in the IP header. Moreover, Figure 7 shows the packet size for three applications, *viz.* Blogspot, Flickr and Google Video. For the two former, wireline packets are much larger than wireless packets while for the later packet sizes are about the same. We noted that Blogspot does not set the "Don't Fragment" bit while Google and Flickr does, but found no fragmentation at all irrespective of these settings and we conclude that this may be the desired result of a deliberate choice by service providers to use smaller packets in wireless networks.

3.2.3 Usage Differences

Understanding the diurnal patterns of cellular traffic is important for the operators to better provision both network and computational resources. In this analysis, we compare the traffic volumes and access times at different hours of the day, aggregated over all

applications and separated for the two most popular applications, Facebook and Google.

We investigate the diurnal patterns by aggregating all applications. For this analysis, we map each flow to the local time, and compute the traffic volume seen at each time interval and normalize the result by the total volume. Figure 8 shows clear diurnal patterns of traffic volumes. In general, users are more active during the daytime than at night but this pattern is strikingly less pronounced for wireless traffic compared to wireline traffic. Noting that the wireline peak co-insides with office hours, a possible explanation for this is that wireline devices are used for working and that web based services are more popular for work related issues. (The wireline devices may be used for other services at other times.) Wireless devices, on the other hand, are used more outside work and relatively more often for web based services than for other services with possibly higher data volumes. In Figure 9 we examine the most popular web sites and see similar patterns over the day for both applications. An important conclusion from these findings is that cellular networks may offer less potential for handling growth by moving traffic in time through, *e.g.,* nightly pre-caching of popular content or low night time tariffs.

3.3 Differences across Wireless Applications

Next, we identify the differences between wireless applications by examining if applications can be categorized to different groups, if they are different along certain metrics, and if the grouping implies different QoS requirements.

3.3.1 Resource Utilization of Different Categories

Two important optimizations are the times to keep a bearer up and the time until idle state. These are set by silence period timeouts hence the more they vary, the more difficult it is to choose timeouts. We quantify this resource utilization by examining the period between any two consecutive packets. We consider the metric of the average packet inter-arrival time for each flow (inter-packet gap).

To examine the difference, we group applications based on their types based on the domain knowledge. In particular, we study three groups: video streaming (YouTube, GoogleVideo, RTVE, Vimeo, Youku), social web (Tuenti, Facebook, Twitter, LinkedIn, MySpace, LiveJournal) and news (CNN, BBC, washingtonpost, wikipedia). For each application group, we show the distribution of inter-packet gap per flow in Figure 10. We indicate the spread within each group (in terms of the standard deviation) by vertical bars. As can be seen in the figures, the different groups exhibit different packet inter-packet gaps and these differences could be used to support advanced optimizations, *e.g.,* application dependent handling of bearer and terminal states.

Finally, we demonstrate the difference using two examples. Figure 11 shows the corresponding CDFs for the two most popular web sites. It is seen that, in general, the next packet of a flow will arrive no more than a few seconds after its predecessor but that this varies between applications. This means that, *e.g.,* silence periods are easier to detect for apple than for facebook; for apple a few seconds of silence means that the flow is over with a probability of about 0.9 while the same figure for facebook is about 0.5.

3.3.2 Differences between Apps in the Same Category

Applications of the same type, *e.g.,* social networks, may be differently implemented and optimized. Next, we apply K-means clustering to special subsets of applications and add Principal Component Analysis (PCA) [16] to investigate the impact of the differ-
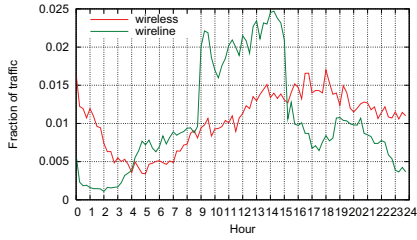
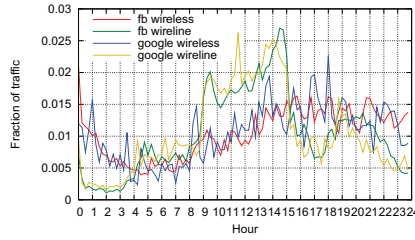**Figure 8: Usage over time for aggregated over all traffic.**

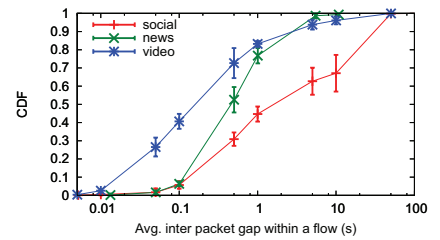**Figure 9: Usage over time for the two most popular sites.**

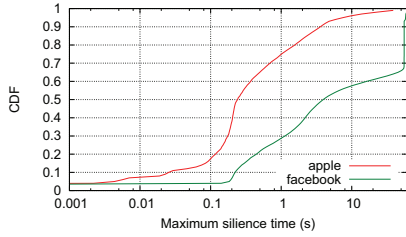**Figure 10: Average inter-packet gap per flow for different groups of applications.**

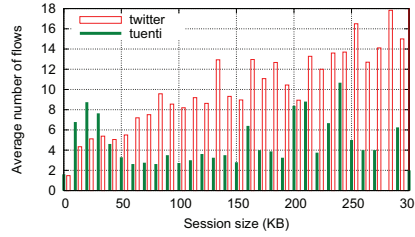**Figure 11: Maximum packet gap per session for two popular sites.**

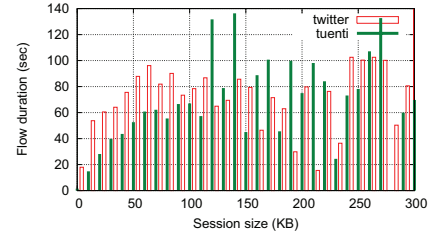**Figure 12: Average number of flows for sessions of the same size.**

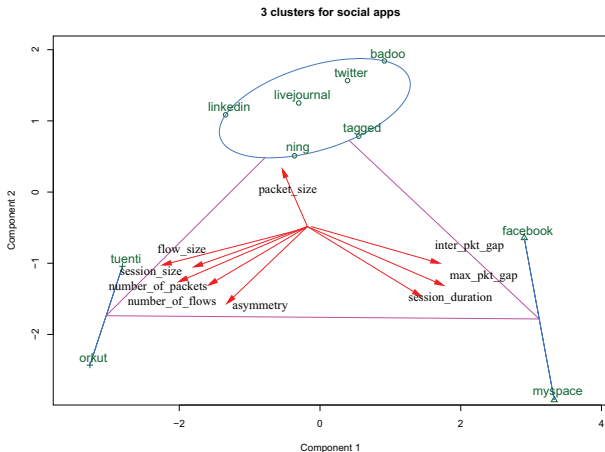**Figure 13: Session duration with same session size.**

Figure 14: Clustering the top 10 most popular social websites.

ent variables on the clusters. Figure 14 shows the results for the top 10 most popular social websites. The first component is dominated by *flow size*, *session size* and *inter-packet gap*, while the second component reflects different *packet size*. For example, Facebook and Tuenti are in different clusters because their flows and sessions are quite different hence they are far apart in the first component. Similarly, Tuenti and twitter are in separate clusters as their packet sizes differ significantly.

We examined all metrics of social applications but, due to the limited space, we focus on one metric, the number of flows in sessions of the same size, using Tuenti and Twitter from different clusters as examples. To deliver the same amount of data, some applications may use a few long flows while others may use multiple small flows. Figure 12 shows that for Twitter the number of flows

tend to grow with the session size whereas for Tuenti the number of flows appears to be relatively independent of the session size. That is, larger sessions correspond to *more* flows in Twitter but to *longer* flows in Tuenti.

The same volume of data can be delivered by a variable number of flows, similar volumes of data can also be delivered with different delays. Figure 13 compares the duration of sessions of similar sizes from the same example, and it is seen that the former tends to produce longer delays with shorter sessions. The delay can be very high, probably because of different application implementation designs.

4. RELATED WORK

There are numerous studies of web workloads from the perspective of proxies [5, 20], browsers [7, 3], and servers [1]. Different models have been studied to capture web traffic dynamics [4, 14, 19, 6]. Most studies are, however, limited to wired networks while there are few and limited studies related to wireless networks. A group of studies attempted to understand cellular traffic from different perspectives. [21] focuses on the diverse usage of smartphone apps in cellular networks. [18] studies the traffic dynamics using flow-level data. They propose a Zipf-like model to capture the volume distributions of application traffic and a Markov model to capture the volume dynamics of aggregate Internet traffic. Paul *et al.* [15] analyzed the radio resources usage. Other studies focus on a specific application such as over-the-top video [8]. Also related are studies that proposed measurement tools for smartphone devices characterizing either the device performance or the performance of certain apps [10, 17, 22]. Our study is complementary, as it focuses on the traffic pattern of web traffic, and it has implications on resource consumption. The tool BLT [9] has some similarities to our web page extraction but appears to handle neither embedded documents nor HTML 1.1.

5. CONCLUSIONS

In this paper, we have analyzed the characteristics of cellular data traffic with respect to different applications and compared the results to those obtained in fixed networks. We summarize our main findings below.

1) We found that, for a majority of applications, wireless sessions contain less data than wireline ones, but that for some others there is either no difference or even larger volumes for wireless networks. 2) We have also noted that a few long flows in fixed networks often translate to many short flows in cellular networks. Flow sizes can thus be deceiving since the number of flows varies between the two accesses and between different applications. An important conclusion is that traffic volumes only can be compared on a session level.

3) Next we noted that packet arrivals are burstier in wireless networks than in wireline ones. This may be because of the radio access (*e.g.,*, it may encourage batches of acknowledgements which in turn may trig batches of user data *etc.*). In any case we note that this means that higher margins are needed in, *e.g.,* cellular backhaul sthan in xDSL access networks. 4) We found the packet sizes differ between the two accesses: wireless packets are smaller. We believe that this is a deliberate optimization to prevent the fragmenting that otherwise could occur as packets are subject to additional tunnelling overhead in cellular networks.

5) The demand for these services vary less over time in cellular networks than in fixed ones. We believe that, to some extent, this may be because the services we are looking at are more attractive during office hours whereas other services are favoured at other times. Another reason may be that the wireline data may include corporate traffic. In any case, suggests that cellular (core) networks reach a more even utilization than wired ones which means that there is less room to handle growth by moving traffic in time through, *e.g.,* nightly pre-caching of popular content or low night time tariffs. 6) Finally, we have seen that inter packet gaps differ significanlty between different service types and we note that this can be used to support advanced optimizations like, *e.g.,* application dependent handling of bearer and terminal states, without DPI.

6. REFERENCES

[1] M. F. Arlitt and C. L. Williamson. Internet web servers: workload characterization and performance implications. *IEEE/ACM Trans. Netw.*, 5(5):631–645, October 1997.

[2] AT&T. AT&T SXSW Press Release. www.att.com/Common/docs/SXSW_Network%20 Fact_Sheet.doc, 2011.

[3] P. Barford, A. Bestavros, A. Bradley, and M. Crovella. Changes in web client access patterns: Characteristics and caching implications. *World Wide Web*, 2(1-2):15–28, January 1999.

[4] P. Barford and M. Crovella. Generating representative web workloads for network and server performance evaluation. *SIGMETRICS Perform. Eval. Rev.*, 26(1), 1998.

[5] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker. Web caching and zipf-like distributions: Evidence and implications. In *INFOCOM'99*, pages 126–134, 1999.

[6] H.-K. Choi and J. O. Limb. A behavioral model of web traffic. In *Proc. International Conference on Network Protocols*, 1999.

[7] C. Cunha, A. Bestavros, and M. Crovella. Characteristics of www client-based traces. Technical report, Boston, MA, USA, 1995.

[8] J. Erman, A. Gerber, K. K. Ramadrishnan, S. Sen, and O. Spatscheck. Over the top video: the gorilla in cellular networks. In *Proc. ACM SIGCOMM IMC*, pages 127–136, 2011.

[9] A. Feldmann. Blt: Bi-layer tracing of http and tcp/ip. In *Proceedings of the 9th international World Wide Web conference on Computer networks: the international journal of computer and telecommunications networking*, page 321–335, Amsterdam, The Netherlands, The Netherlands, 2000. North-Holland Publishing Co.

[10] J. Huang, Q. Xu, B. Tiwana, Z. M. Mao, M. Zhang, and P. Bahl. Anatomizing application performance differences on smartphones. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pages 165–178, 2010.

[11] R. A. Kalden. *Mobile internet traffic measurement and modeling based on data from commercial GPRS networks*. PhD thesis, Enschede, 2004.

[12] S. Lloyd. *K*-means clustering. http://en.wikipedia.org/wiki/K-means_clustering, 1957.

[13] T. W. MacFarland. Mann-Whitney U-Test. http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U, 1947.

[14] B. A. Mah. An empirical model of http network traffic. In *Proc. IEEE INFOCOM*, 1997.

[15] U. K. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das. Understanding traffic dynamic in cellular data networks. In *Proc. IEEE INFOCOM*, 2011.

[16] K. Pearson. Principal Component Analysis. http://en.wikipedia.org/wiki/Principal_component_analysis, 1901.

[17] F. Qian, Z. Wang, A. Gerber, Z. Mao, S. Sen, and O. Spatscheck. Profiling resource usage for mobile applications: a cross-layer approach. In *Proceedings of the 9th international conference on Mobile systems, applications, and services*, 2011.

[18] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang. Characterizing and modeling internet traffic dynamics of cellular devices. In *Proc. ACM SIGMETRICS*, pages 305–316, 2011.

[19] F. D. Smith, F. H. Campos, K. Jeffay, and D. Ott. What tcp/ip protocol headers can tell us about the web. *SIGMETRICS Perform. Eval. Rev.*, 29(1), 2001.

[20] A. Wolman, M. Voelker, N. Sharma, N. Cardwell, A. Karlin, and H. M. Levy. On the scale and performance of cooperative web proxy caching. *SIGOPS Oper. Syst. Rev.*, 33(5), December 1999.

[21] Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, and S. Venkataraman. Identifying diverse usage behaviors of smartphone apps. In *Proc. ACM SIGCOMM IMC*, pages 329–344, 2011.

[22] L. Zhang, B. Tiwana, Z. Qian, Z. Wang, R. P. Dick, Z. M. Mao, and L. Yang. Accurate online power estimation and automatic battery behavior based power model generation for smartphones. In *Proceedings of the eighth IEEE/ACM/IFIP international conference on Hardware/software codesign and system synthesis*, 2010.