Understanding HTTP Flow Rates in Cellular Networks

Ying Zhang[1], Åke Arvidsson[1], Matti Siekkinen[2], Guillaume Urvoy-Keller[3]
[1]Ericsson Research, [2]Aalto University, [3]Laboratoire I3S UNS/CNRS UMR-7271

***Abstract*—Data traffic in cellular networks increased tremendously over the past few years and this growth is predicted to continue over the next few years. Due to differences in access technology and user behavior, the characteristics of cellular traffic can differ from existing results for wireline traffic. In this study we focus on understanding the flow rates and on the relationship between the rates and other flow properties by analyzing packet level traces collected in a large cellular network. To understand the limiting factors of the flow rates, we further analyze the underlying causes behind the observed rates, *e.g.*, network congestion, access link or end host configuration. Our study extends other related work by conducting the analysis from a unique dimension, the comparison with traffic in wired networks, to reveal the unique properties of cellular traffic. We find that they differ in variability and in the dominant rate limiting factors.**

I. Introduction

The volume of data traffic in cellular networks has been increasing exponentially for the past few years and it is predicted to continue increasing over the coming few years as well. Cellular systems operate under restrictive constraints of resources including radio channel capacity, network processing capability, and handset energy consumption. 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.

The operators and the community have already gained tremendous understanding of Internet traffic from various industry and research reports [1], [2], [3], [4], [5], which are obtained by analyzing the traffic from wired networks. Operators can use it to aid the design of better flow scheduling and performance optimization. However, given the prosperity of mobile Internet, it remains a question whether such practices can be reused in managing cellular networks. This paper aims at providing some answers to this question: examining similarities and distinctions between wireless and wireline traffic.

While there are many aspects of traffic properties, we selected one group of traffic, HTTP flows, and one metric, the flow rate, so that we can perform a more focused, detailed, and fair comparison. We selected HTTP flows because it is the dominant category of traffic in both wireless and wireline access. For example, TCP traffic in wireline may contain many more P2P flows than wireless, resulting in a biased comparison. The flow rate metric has been studied extensively a decade ago for wireline networks [6]. We believe that there is a need to re-examine the problem in the cellular context for the following reasons. First, new applications and new traffic patterns emerge. Web traffic has thus undergone a significant change, from the simple web page to complex applications, *e.g.,* media or social networking. Second, the appearance of new user devices and new network access technologies. In addition, the range of applications and operating systems significantly differs from the ones developed for wired networks. For example, application developers may have customized designs for cellular traffic due to the constraints on radio network resources and handset energy consumptions. There are few and limited studies for cellular networks and they do not reflect the recent cellular traffic surge [7] or have different focuses [8], [9].

We take two steps to investigate HTTP flow rates in cellular networks. From the macroscopic perspective, we examine the basic characteristics of HTTP flow rates, with an emphasis on their distribution and the correlation with other flow properties. To illustrate the unique characteristics of cellular data traffic, we perform a comparative analysis of wireline and wireless accesses. We seek answers to questions like "Are the flows in wireless network slower or faster in general?", "Do flow rates vary significantly with size? ", and "Which are the dominant limiting factors of flow rates?". To further answer the last question, from a microscopic perspective, we then study the causes that limit the rates at which flows transmit data using the algorithms proposed in [10]. Factors considered include congestion, sender window, receiver window, bandwidth, and application induced limitations. These analyses provide knowledge of the potential bottlenecks of data transfers.

We summarize our key findings below.

- From the macroscopic studies, we found that flow rates in both wireline and wireless data sets vary over several orders of magnitude and, as expected, wireless has smaller flow rates. For flows of the same size, the variance of rates in wireless is larger than that in wireline. It means that the impact of channel variance outweighs the regulation attempts, such as fair queuing, at the base station which constitutes the bottleneck.
- This observation is further confirmed in the microscopic root cause analysis. We observe that applications do not control the flow rates actively. To efficiently use the radio link resource, most applications over HTTP send most bytes as fast as possible. Interestingly, this observation is different from the recent findings in an LTE network [11], as the latter has much larger bandwidth which appears to shift the bottleneck back to the application.
- Finally, our results also suggest that the main rate limiting

ISBN 978-3-901882-58-6 © 2014 IFIP

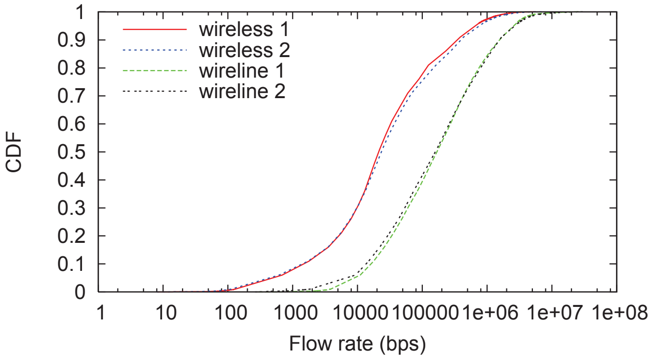

Fig. 1. CDF of flow rates.

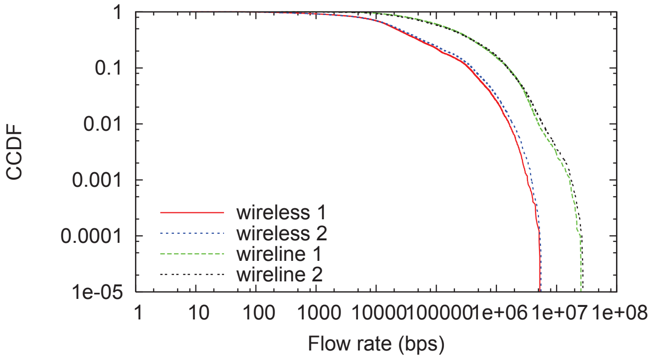

Fig. 2. CCDF of flow rates.

cause differs between wireless and wireline traffic. The wireless access link seems to be the bottleneck in most cases. This means that to reduce response times in a wireline network, increasing the initial congestion window might pay off, while this should have a detrimental effect in the wireless network.

In the rest of this paper, we first introduce the methodology in Section II. Section III presents various statistics for flow rates and the root cause analysis results are shown in Section IV. We discuss related work in Section V and conclude in Section VI.

II. Methodology and Data Set

We divide our analysis in two steps. The goal of the first step is to report basic flow statistics, including flow rate, size and duration. Once these macroscopic statistics are discovered, we investigate the causes behind these flow rates in our second step using our algorithm in [10] which is based on the understanding of TCP dynamics and congestion control mechanisms.

We have used two data sets collected in Europe during 2011 without packet sampling but with automatic anonymisation of IP addresses. 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 2 weeks of data with 3.9 million sessions from 50k IP addresses. The wireline data set was collected at the head end of a municipal access network. It contains 31 hours of data with 19 million sessions from 21k IP addresses. The HTTP traffic accounts for 67.4% of the total wireline traffic and 75.8% of all the wireless traffic in bytes. We are aware that the network conditions vary across countries and regions. We however expect that our findings are representative of developed countries where both types of access are popular and well deployed.

III. Macroscopic: flow rate characterization

We drive our analysis based on a number of hypotheses regarding possible differences that are related to different access technologies. In the following analysis, the flow rates are computed as the flow size divided by the flow duration, where the flow size is defined as the total number of bytes transferred and the flow duration is the time elapsed between the first and last packet. Noting that some flows only contain a few packets, which makes flow rate less meaningful, we focus on flows with significant data transfers by removing flows shorter than 100 milliseconds [6] or smaller than 100 packets. To verify that the conclusion is not biased by a particular data set or time period, we separate the two-week wireless data into two sets of one-week each, and the 31-hour wireline data to two sets of 15.5-hours each.

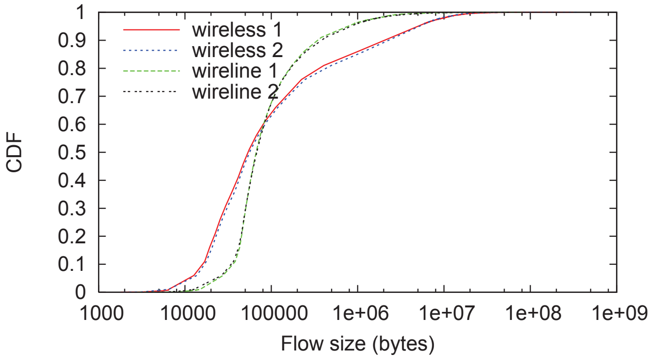

Fig. 3. CDF of flow sizes.

A. Flow rate distribution

Hypothesis 1: It is suspected that wireless flows are slower because of constraints related to, *e.g.,* longer RTTs (*e.g.,* due to link layer error corrections in the wireless access network) and constrained bandwidth (*e.g.,* because of sharing between users in the same cell or sector). Note that it is unlikely to be caused by errors or losses because they are corrected by the link layer recovery mechanisms in wireless access networks.

To examine this hypothesis, we first show the CDFs (cumulative distribution functions) of flow rates in Figure 1. First, we observe that the flow rates in both data sets vary over several orders of magnitude. Second, flow rates tend to be slower for wireless access than for wireline access. We also show the tails of the flow rates in log-log scale in Figure 2. It is seen

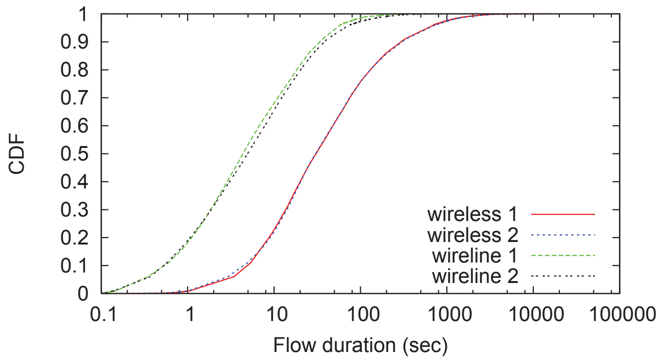

Fig. 4. CDF of flow durations.

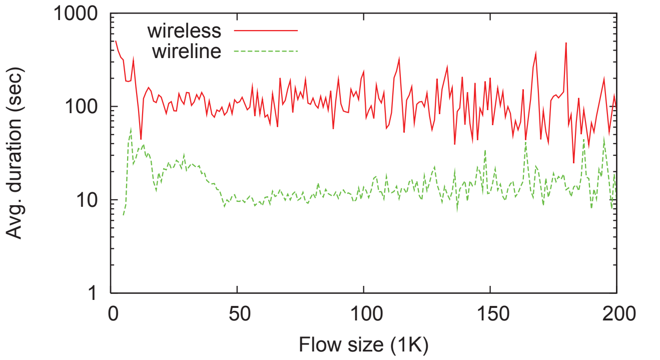

Fig. 5. Flow duration for the same flow size.

that both data sets exhibit long tail distributions and, again, that flow rates for wireless tend to be lower than for wireline access. Both figures thus confirm *Hypothesis 1*, *i.e.,* wireless networks feature smaller rates in general as compared to wired networks.

We further investigate the reasons for this difference. Since the flow rate is computed from flow size and flow duration, we next examine their distributions. This analysis is driven by ***Hypothesis 2:*** because of the limitations of the devices, *e.g.,* screen size, computing capability and power supply, one may expect that sizes and durations for content differ between mobile devices and PCs.

Figure 3 and Figure 4 show the CDFs of flow sizes and flow durations. It is seen that (1) the range of flow size distributions are similar from the curves of both access, and (2) wireless flows are more “extreme” than wireline ones with larger fractions of small and large flows respectively. Thus, the assumption of content being adapted to mobile devices with smaller resulting sizes is not generally observable. On the other hand, the flows last much longer in wireless, which is likely to be one reason for the lower flow rates. We further confirm this in Figure 5 where we group flows according to their sizes and compute the average duration of each group on y-axis and which clearly shows longer flow durations for files of the same size. Lower rates in wireless are thus the result of technological constraints or more resource contention, leading to longer duration rather than difference in sizes of the transfers.

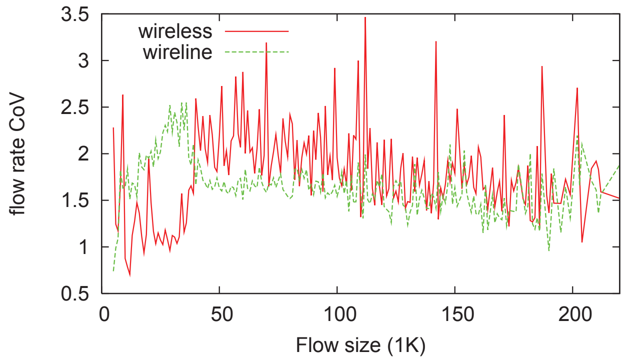

Fig. 6. CoV of flow rates with same flow sizes.

B. Inter-flow rate variability

While the above figures present the distribution of all flow rates, we next focus on the stability/variability of flow rates, because of the following two conjectures. On the one hand, wireless flows may exhibit smaller variability because their bottleneck typically is the air interface which is shared and subject to fair scheduling while wireline flows may be constrained anywhere in the network and fair scheduling seldom is present at these bottlenecks (***Hypothesis 3***). On the other hand, wireless networks offer variable radio conditions. Moreover, the air interface schedulers transmit more traffic for users with better radio conditions in order to achieve higher overall throughput. The variability in radio conditions may thus be translated to a higher variability with wireless access (***Hypothesis 4***).

We examine the variability by grouping flows by size. Figure 6 shows the coefficients of variation (CoV) for different, size-based flow groups with at least 30 flows in each group. It is seen that the throughput is much more variable in wireless networks than in wireline ones, especially for larger flows. This is confirmed by Figure 7 which shows the CDFs of flow group CoVs.

However, we realize that grouping by size alone may result in biases due to the variability of channel conditions experienced by different users. For instances, two users with very different yet stable wireless connections, *e.g.,* 500 kbps and 1 Mbps may be assigned to the same size group in Figure 6, resulting in very large CoV. To remove such bias, we also group by source and also by destination addresses. To overcome the dynamic IP assignment issue in wireless networks, we first identify an active transfer window for each IP address (*i.e.,* periods during which packets are continuously sent/received from/to this IP), and include flows within these windows. The distributions of their CoVs are also shown in Figure 7 and we conclude that, for all three grouping methods, wireless flows consistently show higher variability. These results thus suggest to reject *Hypothesis 3* and accept *Hypothesis*

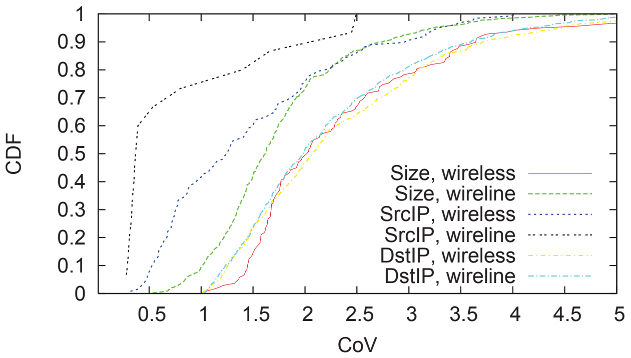

Fig. 7. CoV of flows grouping by source IP, destination IP, and flow size.

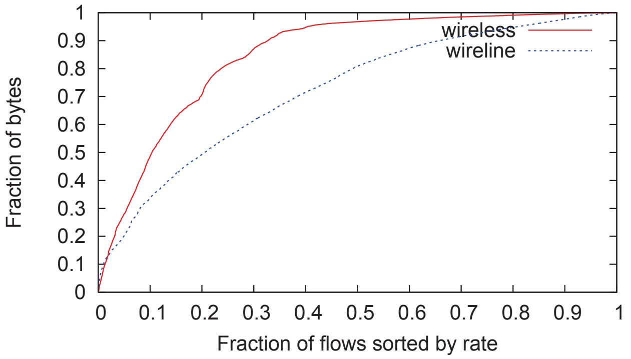

Fig. 8. Traffic volume distribution across flows.

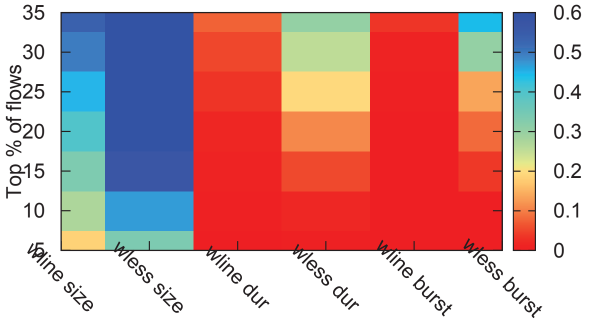

Fig. 9. Fraction of overlapping top % flows ranked by rate and size, duration, burstiness.

4. It means that the regulation effect of the scheduler at the base station apparently has less impact than the variability of the channel conditions. It could also be caused by different subscription plans with different caps on data rates.

C. Correlation with flow size

We further look into other properties of flows with high rates by examining the correlations between flow rates and other factors. From Figure 1 and 2 we observe that most flows are slow, *i.e.,* 50% of rates are smaller than 20 kbps for wireless, but that there also are quite a few fast flows which reach rates above 1 Mbps. Thus, our ***Hypothesis 5*** is that these fast flows also are the "elephant flows", *i.e.,* the ones which carry most of the traffic. To verify it, we compute the accumulated fraction of bytes carried by a subset of the flows sorted by rate in descending order. The result is shown in Figure 8 with fastest flows at the right end of x-axis. The results confirm our expectation; the top 20% fastest flows account for 45% of the traffic in wireline data set and 70% of the traffic in the wireless data set. This suggests that flow rates are determined by radio optimizations, *e.g.,* sending data as fast as possible by, *e.g.,* switching from common channels to dedicated channels for "persistent flows". This result not only confirms the high correlation between flow sizes and rates from previous work [6] but also suggests that the correlation is more pronounced in wireless.

Next we consider the relationship between flow rate and other key parameters. To this end, we first select the top $k\%$ of fastest flows (measured as throughput), and then check the fraction of these flows also belonging to the top $k\%$ largest flows (measured in flow size), longest flows (measured in flow duration) and most bursty flows (measured in the average packet inter-arrival time for packets in each flow). This analysis is similar to [5] but with a slightly different definition. Figure 9 shows the k values in y-axis and different comparison metrics in x-axis. The colors indicate different degrees of overlap. For example, dark blue indicates large overlap and red means small overlap. The figure shows that rates of wireless flows exhibit a strong correlation with flow size. On the other hand, we observe a weak correlation with burstiness while rates of wireline flows exhibit medium correlation with flow size and no correlation with burstiness. It is also seen that flow rate exhibits no correlation with flow duration in any of the data sets. It is consistent with Figure 8, *i.e.,* fast flows are also elephant flows, and this correlation is more pronounced in wireless. Interestingly, some of these fastest flows are also the most bursty ones in wireless, which is not significant in wireline. One likely reason is that fast flows fully utilize the radio access link by sending packets as fast as possible, thus, are more affected by the dynamic conditions on the radio links, resulting in large jitter.

IV. Microscopic: The Origins of Flow Rates

From the above analysis of flow rates, we observe that wireless flows are generally slower and more variable, and that they exhibit higher correlation with flow sizes. This suggests that the application level just sends data as fast as possible (*i.e.,* with little control), and the limiting factor of flow rates is the wireless access link. Next, we conduct a deeper analysis on the limiting factors behind the observed rates and seek for additional evidences of the lesser impact of the application as compared to the access link characteristics in wireless.

A. Method

In this section, we apply a root cause analysis (RCA) tool on the same dataset, to understand the factors affecting the rates of flows in the access networks we consider. The *RCA* technique has been introduced and validated in [10][1]. First, the tool divides each TCP connection into two kinds of periods, application limited periods (*ALP*) and bulk transfer periods (*BTP*). ALPs correspond to time intervals during which the application limits the throughput achieved, which happens, *e.g.,* when a server applies rate limitation or when users pause to read a web page before clicking on another link and the browser uses persistent HTTP. During BTPs, the transport or network layer (of the end hosts involved in the transfer or on the path between those hosts) is the limiting factor. Within a single TCP connection, multiple such phases may exist interleaved with each other, *i.e.,* a connection is transformed into a series of ALPs and BTPs in the first phase.

In the second phase, the tool drills down into the BTPs and tries to identify the main rate limiting cause. A number of metrics (so called limitation scores) are computed for each BTP which then are used to classify each BTP into one of the following limitation causes:

- **Unshared bottleneck link:** The target flow alone utilizes all the capacity of the bottleneck link. The latter is likely to be the access link at the sender or receiver side.
- **Shared bottleneck link:** There is cross traffic competing for the bandwidth.
- **TCP receiver:** The receiver buffer constrains the maximum number of bytes TCP can send without acknowledgement when there is still available spare bandwidth. This limitation means that the bandwidth delay product is larger than the receiver buffer or the receiver is not consuming the data fast enough.
- **Transport limited:** When there is no packet loss, and no limits from the TCP receiver, the TCP congestion control mechanism unnecessarily limits the increase of the congestion window. In such cases, increasing the initial congestion window of TCP will reduce the total transfer time [13].
- **Mixed/unknown:** A mixture of the causes described above or unidentified causes.

The algorithms are implemented using a DBMS-based approach called InTraBase [12][2]. In the following study, we use four hypotheses based on a priori knowledge and test them against our RCA results. From our experience with wired networks [14], we came with up:

- ***Hypothesis 5:*** Most of the bytes are sent as fast as possible by TCP.
- ***Hypothesis 6:*** Application limited flow rates are smaller than those limited by other causes.
- Moreover, according to typical network provisioning strategies, we have ***Hypothesis 7:*** The last mile is the network bottleneck in wireless.
- Finally, the recent discussion on increasing the initial congestion window size of TCP to improve the Web performance[15], [13] gives rise to ***Hypothesis 8:*** Increasing the size of the initial congestion window would improve the performance.

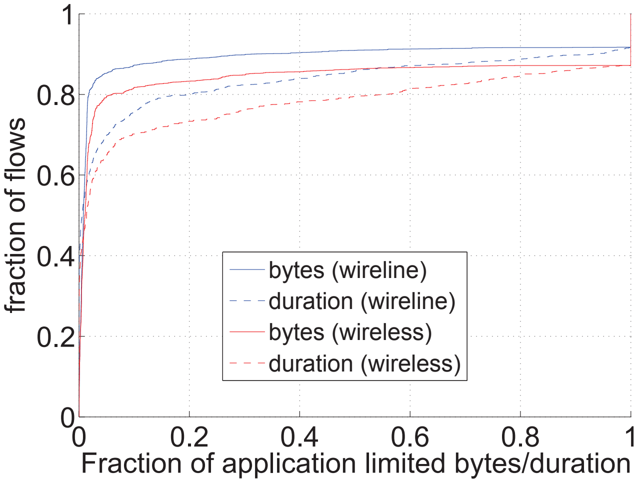

Fig. 10. Connections tend to have either all or no bytes application limited.

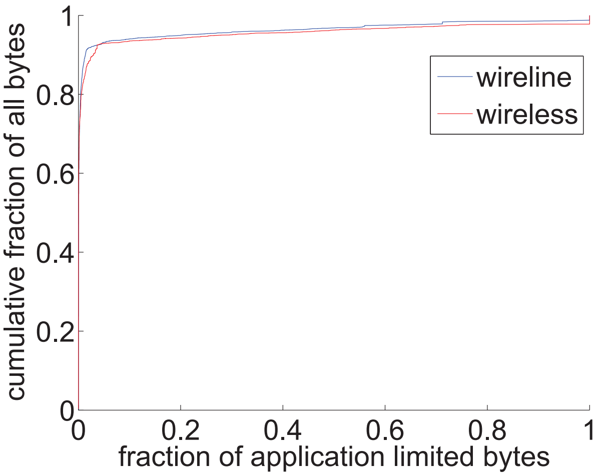

Fig. 11. Majority of bytes are not rate limited by applications.

Hypothesis 5 and 3 are important to back the results in Section 3. More precisely, we need that (i) application having little impact on the rate for most flows and (ii) the last mile being the bottleneck, to confirm that the difference between wireless and wireline traffic stems from the radio access link.

B. *The role of applications*

While using RCA, connections shorter than 130 packets are not analyzed at all but are directly classified as opportunistic flows, similarly to [6]. Consequently, we end up analyzing a minority of flows but a majority of bytes because of the heavy-tailed nature of Internet traffic.

To enrich the analysis with application-level knowledge, we inferred the Web service used, based on keywords in the URL. The results show that only a few applications were responsible for the majority of the bytes in both wireless and wireline traces. More precisely, the top 20 most popular sites aggregate about 80% of the bytes in the wireline case and over 90% of

[1] We implement the RCA algorithms using a DBMS-based approach called InTraBase [12]. The tool is available upon request to matti.siekkinen@aalto.fi.

[2] The tools are available by request to matti.siekkinen@aalto.fi.

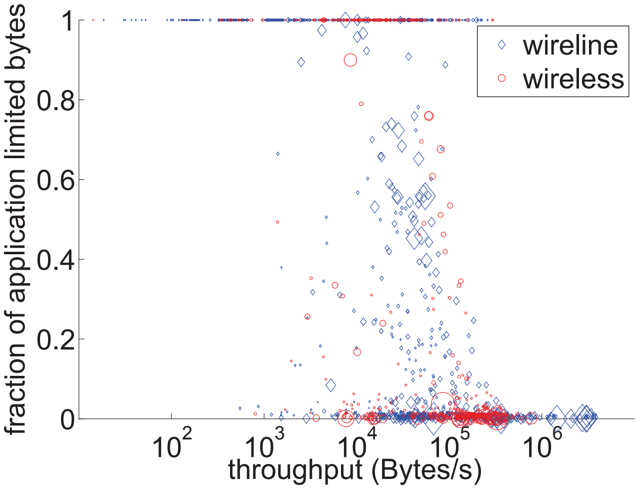

Fig. 12. Throughput distributions for different ALP fractions. The fastest flows carry no application limited bytes.

the bytes in the wireless case. Facebook, YouTube and other video streaming are the dominant applications, generating over 40% of the traffic.

We first examine the role of the application in limiting the flow rates. Figure 10 plots the CDF of the per-flow fraction of bytes carried by and time spent in ALPs. We observe that a clear majority of both wireless and wireline connections experience no rate limiting by application at all. The plot does not, however, tell the difference between small and large flows and this is why we show in Figure 11 the overall fraction of bytes (over the whole set of flows) that are classified as application limited. An overwhelming majority of bytes are not rate limited by applications, which means that the most prominent root causes for the flow rates lie in the network and transport layers as was also suggested by our analysis in Section III. Hence, **Hypothesis 5** is confirmed. Note that this is different from the findings in LTE [11] where the available bandwidth is much higher and thus the application becomes the limit.

To better understand the role of the applications, we further study how much applications slow down transfer rates. We compare the throughput of a flow against the fraction of application limited bytes in Figure 12. Each symbol corresponds to one flow (circles indicate wireless and diamonds indicate wireline). In addition, the size of the symbol is proportional to the size of the flow where larger symbols correspond to larger flows. We observe that the fastest flows are those that carry no application limited bytes and application limited flows are likely to be those with very low rates, confirming **Hypothesis 6**. In line with the interpretation of Figure 8, we observe that flows that feature high rates tend also to be large flows.

Though the above analysis shows that the application is not the dominant limiting factor for *all* traffic, it may play a different role for different types of applications. Thus, below we investigated separately for a few popular applications. The results of YouTube transfers are shown in Figure 13, which shows a dual distribution. This is because some of the flows correspond to download of the web pages, while some others, usually larger, flows correspond to the actual video content download. The former type of flows is likely sent at full rate by the server, whereas the video transfers are rate throttled by servers resulting in ALPs in our classification. However, it is known that the YouTube content delivery strategy varies depending on many factors [16]. While servers often apply rate throttling, sometimes the whole video is downloaded in one shot bypassing the server rate throttling, resulting in a single BTP. Furthermore, a so called Fast Start period is always present in the beginning of the video delivery regardless of the strategy. During that period, all the available bandwidth is typically used to deliver a sizeable part of content in order to quickly fill the client's playback buffer. That phase corresponds also to a BTP.

Overall, Figure 13 reflects the known complex nature of the traffic patterns of YouTube well. We also looked separately at Facebook traffic. It turned out that, similarly to YouTube traffic, it has a dual nature where roughly half of the connections were completely application rate limited, and the other half not application limited at all.

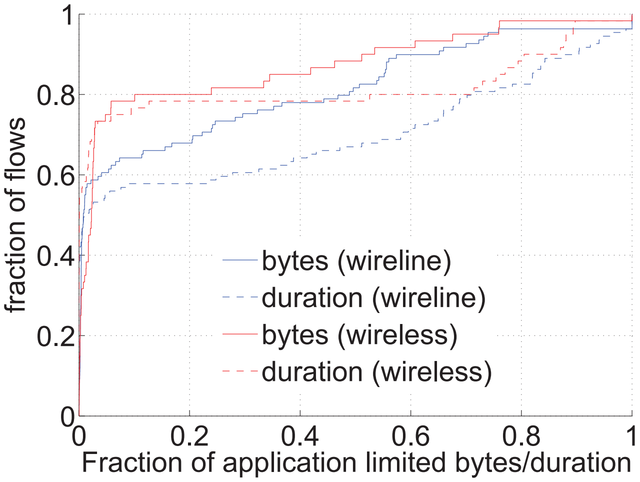

Fig. 13. Some YouTube transfers are server rate limited.

Root cause	Un-shared	Shared	Receiver	Trans-port	Mixed/ unknown
Wireline	1%	12%	0%	38%	49%
Wireless	58%	16%	0%	3%	23%

TABLE I
RCA RESULTS FOR BULK TRANSFER PERIODS.

C. A closer look at bulk transfer periods

We now focus on the BTPs (periods where the rate is not driven by the application). We already know that BTPs dominate in our traffic, but to further back the results of Section III we need more evidences that confirm the observed rates are explained by the access link in the wireless case. We report the fractions of bytes per limitation cause in Table I. While the impact of the application on the throughput achieved is similar in both wireless and the wireline networks, the bulk transfer analysis shows two main differences:

1) Unshared bottleneck link limitation is dominant among the causes observed in the wireless network in contrast to being almost non-existent in the wireline network. For the wireless case, it is likely that the bottleneck is the client's access link as servers are usually hosted in well-provisioned networks. This observation suggests that there is sufficient capacity in the wireline network under study to sustain the current users' HTTP traffic. The transfers reach the throughput limit set by the wireless network much more often than the wireline network, which supports **Hypothesis 3**, *i.e.,* the last mile is the bottleneck.

2) Rate limiting by the transport protocol has much stronger presence in wireline traffic compared to wireless traffic. This finding indicates that the practice of using a larger initial congestion window size could speed up transfers in the wireline networks as advocated in [13].

Note also that while newer versions of TCP, e.g., TCP Cubic under Linux, are able to achieve higher rates on paths with high bandwidth-delay product, they are apparently not operating in this regime for the wireline case. For the case of TCP Cubic, this means that the transport protocol is operating in the TCP-mode and not the Cubic-mode [17]. It thus suggests that the correct solution to speed up transfers is to increase the initial congestion window.

The situation is apparently completely different in the wireless network under study. In this case, transfers are rarely limited by congestion windows that could have opened faster. In such a scenario, the better strategy is not to inflate the initial congestion window, especially when considering the observation that applications often operate close to the limit of the access link capacity. Increasing the initial congestion window could instead lead to losses or buffer bloat [19], or to an increased RTT if a buffer can absorb the resulting burst, but not to any reduction in the response time. Hence, **Hypothesis 8** holds true in wireline network but should be rejected in wireless network.

Table I also shows commonalities between wireline and wireless. We first observe that there is no obvious receiver window limitation, which suggests that the transport layer of the operating systems of the end users (Windows, Linux, Android, iOS) is correctly configured and does not constitute a bottleneck. Surprisingly, we do not observe small receiver window induced bufferbloat problem as reported in [18]. The lack of receiver limitation in the wireless also suggests that the processing capacity of the mobile devices is not a bottleneck either. There are also similar fractions of shared bottleneck links, which suggests that congestion within the network path (not at the edges) occurs similarly irrespective of the type of access. It indicates that the access networks are well-provisioned in both networks.

D. Bufferbloat problem

We further checked for evidence of bufferbloat which has been discussed a lot recently by the research community [19], especially in the cellular case. We computed the ratio of the average RTT to the SYN/ACK-RTT as an indicator of bloated buffers during the transfer. We found some cases where that ratio was large, going in extreme cases even beyond 100. The RCA results for flows with ratios greater than or equal to 5 revealed that a great majority of such wireless flows were limited by an unshared bottleneck link which suggests that the wireless access link is the bottleneck. A closer look at the number of applications used in parallel would be required to assess the impact of bufferbloat on the end users. Indeed, if the user runs a single application at a time, the increase of RTT should have little impact on the perceived performance. We leave this study for future work.

E. Summary

We summarize the key findings of the limiting factors study below. First, the application behaviour explains the rate of only a minority of the large flows (**Hypothesis 5** accepted). Second, if a flow rate is limited by the application, the achieved rate is in general low, confirming **Hypothesis 6**. Third, the end host transport layer appears correctly parametrized and does not constitute a bottleneck. Finally, the solutions to achieve higher rates in the two network seem at odds: increasing the initial congestion window looks like a good strategy for wireline access. In contrast, such action could have an adverse impact, leading to losses and increased delays in the wireless network, which rejects **Hypothesis 8**. The solution to achieve higher rates in our wireless network is thus much more costly as it requires increasing the capacity for the (wireless) access link, agreeing with **Hypothesis 7**.

V. Related Work

Our work builds on top of many attempts to understand the web traffic properties in the past [1], [2], [3], [4], [5]. In particular, the closest work to ours, T-RAT [6] analyzes packet-level TCP dynamics and infers the cause of the rates at which flows transmit data. Lan*et al.* [5] found strong correlations between flow size, rate and burstiness. Our work revisits and re-validates these findings in wireline networks, which were done decades ago and, we focus on identifying the distinctions between wireless and wireline traffic.

Recently there has been a number of research in cellular traffic. They can be classified as device based analyses or network based analyses. Papers in the former group study wireless usage by instrumenting mobile devices [20], [21], [22] while papers in the later group, which includes our work, relies on cellular data traffic collected by operators [8], [9], [11], [23], [24], [25], [26]. Some of them focus on understanding the user mobility patterns and resource utilization. For instance, Paul *et al.* [25] analyzed the radio resources usage with respect to the user mobility. 3GTest [20] measures the network performance of popular smartphone platforms, and ARO [21] characterizes the radio resource usage of mobile apps. These studies are orthogonal to ours, as mobility is a unique property in wireless networks.

Within the second group, a large body of existing work, like ours, focus on finding differences in traffic characteristics related to different users, applications, and device types. For example, [8] focuses on the diverse usage of smartphone apps in cellular networks, [9] uses flow-level data to study traffic pattern differences due to mobile device types and applications and [26] examines the performance of over-the-top video in wireless. Though we also analyze different application patterns, our focus is on the comparison to wireline traffic for the same application. Recently, Huang *et al.* [11] finds that flow rates are limited by applications in LTE networks, which has much larger bandwidth in radio access link compared to 3G networks that we consider in this study. Overall, our study differs from related work by the *explicit* comparison between wireless and wireline access, and by *directly* comparing results. Though some similar analyses have been done in related work separately, none of them drew any conclusion from the comparison perspective.

The most relevant work is [27] which compares the two accesses from the general traffic properties. Our work takes that idea one step further to examine the flow rates and the underlying causes.

VI. Conclusions

In this paper, we have analyzed the flow rate characteristics of cellular data traffic and compared the results to those obtained for wireline networks. We relied on two large data sets collected from a wired and a wireless access network under the control of the same ISP in Europe, to carry out this comparison. We found that flow rates in wireless networks are smaller and that they exhibit higher variability (apparently due to the high variance of the radio access links). In terms of the limiting factors of flow rates, we observed that applications have limited control on flow rates for both wireless and wireline. However, the wireless network is operated close to its limits as the access link most of the time is the bottleneck. This is in clear contrast with the wireline network where there is spare capacity at the edge. Using larger initial congestion windows to achieve high flow rates, which is widely used in wireline networks, is thus not sufficient in wireless. Our findings thus provide insights for resource management and optimizations in cellular networks.

References

[1] B. A. Mah, "An empirical model of http network traffic," in *Proc. IEEE INFOCOM*, 1997.

[2] P. Barford and M. Crovella, "Generating representative web workloads for network and server performance evaluation," *SIGMETRICS Perform. Eval. Rev.*, vol. 26, no. 1, 1998.

[3] 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.*, vol. 29, no. 1, 2001.

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

[5] K.-c. Lan and J. Heidemann, "A measurement study of correlations of internet flow characteristics," *Computer Networks*, vol. 50, pp. 46–62, January 2006.

[6] Y. Zhang, L. Breslau, V. Paxson, and S. Shenker, "On the characteristics and origins of internet flow rates," *SIGCOMM Comput. Commun. Rev.*, vol. 32, pp. 309–322, August 2002.

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

[8] 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*, pp. 329–344, 2011.

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

[10] M. Siekkinen, G. Urvoy-Keller, E. W. Biersack, and D. Collange, "A root cause analysis toolkit for tcp," *Computer Networks*, vol. 52, no. 9, pp. 1846–1858, 2008.

[11] J. Huang, F. Qian, Y. Guo, Y. Zhou, Q. Xu, Z. M. Mao, S. Sen, and O. Spatscheck, "An in-depth study of LTE: Effect of network protocol and application behavior on performance," in *Proc. ACM SIGCOMM*, August 2013.

[12] M. Siekkinen, E. W. Biersack, V. Goebel, T. Plagemann, and G. Urvoy-Keller, "InTraBase: Integrated traffic analysis based on a database management system," in *Proceedings of IEEE/IFIP Workshop on End-to-End Monitoring Techniques and Services*, May 2005.

[13] J. Chu, N. Dukkipati, Y. Cheng, and M. Mathis, "Increasing TCP's Initial Window." RFC 6928 (Experimental), April 2013.

[14] M. Siekkinen, D. Collange, G. Urvoy-Keller, and E. W. Biersack, "Performance limitations of ADSL users: A case study," in *Proceedings of the PAM 2007*, April 2007.

[15] N. Dukkipati, T. Refice, Y. Cheng, J. Chu, T. Herbert, A. Agarwal, A. Jain, and N. Sutin, "An argument for increasing tcp's initial congestion window," *SIGCOMM Comput. Commun. Rev.*, vol. 40, pp. 26–33, June 2010.

[16] M. Hoque, M. Siekkinen, J. K. Nurminen, and M. Aalto, "Dissecting mobile video services : An energy consumption perspective," in *Proceedings of the 14th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, WoWMoM'13, 2013.

[17] S. Ha, I. Rhee, and L. Xu, "Cubic: a new tcp-friendly high-speed tcp variant.," *Operating Systems Review*, vol. 42, no. 5, pp. 64–74, 2008.

[18] H. Jiang, Z. Liu, Y. Wang, K. Lee, and I. Rhee, "Understanding bufferbloat in cellular networks," in *Proceedings of CellNet'12*, 2012.

[19] H. Jiang, Y. Wang, K. Lee, and I. Rhee, "Tackling bufferbloat in 3g/4g networks," in *Proceedings of the 2012 ACM conference on Internet measurement conference*, IMC '12, (New York, NY, USA), pp. 329–342, ACM, 2012.

[20] 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*, pp. 165–178, 2010.

[21] 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.

[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.

[23] A. Gember, A. Akella, J. Pang, A. Varshavsky, and R. Caceres, "Obtaining in-context measurements of cellular network performance," in *Proc. ACM SIGCOMM IMC*, 2012.

[24] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "A first look at cellular machine-to-machine traffic: large scale measurement and characterization," in *Proc. ACM SIGMETRICS*, 2012.

[25] 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.

[26] 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*, pp. 127–136, 2011.

[27] Y. Zhang and Å. Arvidsson, "Understanding the characteristics of cellular data traffic," in *Proceedings of the 2012 ACM SIGCOMM workshop on Cellular networks: operations, challenges, and future design*, 2012.