Impact of TCP Protocol Efficiency on Mobile Network Capacity Loss

Ke Liu
The Chinese University of Hong Kong
lk008@ie.cuhk.edu.hk

Jack Y. B. Lee
The Chinese University of Hong Kong
yblee@ie.cuhk.edu.hk

Abstract— **With the rapid increases in mobile Internet users, the need for mobile data services has grown tremendously over the years. Even with the deployment of advanced mobile data networks such as HSPA and LTE, many mobile operators around the world are still experiencing challenges in fulfilling the ever increasing bandwidth demands. While the obvious bottleneck is in the network infrastructure – limited number of cells and bandwidth, it is *not* the only bottleneck. Paradoxically, existing mobile networks may in fact be under-utilized from time to time. This seemingly contradicting observation is due to TCP's inability to utilize all the bandwidth available. This work investigates this problem by developing a stochastic model to relate the system's service response time to TCP's protocol efficiency, and to quantify the network's capacity loss due to TCP. Using real-world network parameters the model revealed that network capacity loss is surprisingly high at typical traffic loads. More interestingly, the results uncovered the inter-play between network capacity loss and protocol/channel bandwidth limits, which opens up a new dimension to the optimization of mobile cell bandwidth allocation.**

Index Terms—Markov Model, Capacity Loss, Mobile Data Network, Transmission Control Protocol (TCP)

I. Introduction

The need for mobile data services has grown tremendously with the rapid increases in mobile Internet users over the years. Even though some mobile operators have already deployed advanced networks such as High Speed Packet Access (HSPA) [1] and Long Term Evolution (LTE) [1], many are still experiencing challenges in fulfilling the ever increasing bandwidth demands. While the obvious bottleneck is in the network infrastructure – the limited number of cells and bandwidth, it is *not* the only bottleneck. Existing mobile networks may in fact be *under-utilized* from time to time. This seemingly contradicting observation is due to one important element in the mobile Internet – Transmission Control Protocol (TCP).

Specifically, a number of previous works [2-5] have clearly shown that TCP often fails to fully utilize the bandwidth available in the mobile network. For example, Liu and Lee [2] showed that TCP CUBIC [6] – the current TCP implementation in Linux kernel 2.6, can achieve a throughput of only 1.5 Mbps out of 5.6 Mbps over a 3G network, and 34 Mbps out of 81 Mbps over a LTE network respectively.

If TCP cannot fully utilize the bandwidth available, the unused bandwidth will be *lost* as bandwidth cannot be stored for use in the future. As a result, the TCP flow will last for a longer time, consuming future bandwidth to complete the transfer. In case the network becomes fully utilized later, the future users will suffer from longer service response time due to competition from the extended TCP flow. Effectively the network performs as if it has a *lower* capacity.

In this work we investigate this problem by developing a stochastic model to relate the system's service response time to TCP's protocol efficiency, and to quantify the network capacity loss due to TCP. The rest of the paper is organized as follows: Section II reviews some background and related work; Section III presents a stochastic model to quantify the impact of TCP protocol efficiency; Section IV analyzes such impact using numerical results computed from real-world parameters; Section V summarizes the study and discusses some future work.

II. Background and Related Work

A. TCP Performance over Mobile Data Networks

The performance of TCP over mobile networks has been studied by various researchers [2-5]. Common to these studies is the observation that TCP often cannot fully utilize all the bandwidth available in the network, even under good radio signal conditions [2]. This is due to a number of inter-related factors including (a) the presence of random packet losses; (b) rapid bandwidth fluctuations; (c) round-trip time (RTT) fluctuations; (d) large network queue; (e) small receiver's advertised window; and so on. It is beyond the scope of this paper to investigate the protocol dynamics of TCP in mobile networks and the interested readers are referred to the related studies [2-5] for more details.

In terms of protocol efficiency, which we defined as the ratio of achievable TCP throughput over the raw bandwidth available in the network, it can vary across (a) different TCP variants; (b) radio signal conditions; (c) types of networks; (d) link-layer configurations; and even (e) types of TCP client implementations. We summarize in the following the protocol efficiencies reported by previous studies.

In a recent study, Lin *et al.* [4] conducted extensive measurements over a 3G (CDMA 1xEV-DO) network using four TCP variants, namely TCP CUBIC [6], TCP Reno [7], TCP Westwood [8], and TCP Vegas [9]. They found that TCP CUBIC, TCP Reno, and TCP Westwood can achieve around 42%-80% protocol efficiency under different radio signal conditions. For higher-speed HSPA networks Ren and Lin [5] conducted simulations and found that TCP has less than 70% protocol efficiency. In another study Chan [3] evaluated the performance of TCP CUBIC, TCP Westwood, and TCP Vegas over HSPA networks using trace-driven simulations, and found protocol efficiencies ranging from 10.8% to 81.0% depending on packet loss rates and RTT.

We recently conducted similar measurements over 3G (HSPA), 3.5G (HSPA+), and LTE networks under a variety of network conditions and mobile client configurations and found that TCP's protocol efficiency can range from 10% to 80%. In a separate work [2] we also developed an accelerated TCP which can improve TCP's protocol efficiency to 90% over HSPA and LTE networks. All these previous works confirmed two properties of TCP over modern mobile data networks: (a) TCP is unable to utilize all the physical bandwidth available; and (b) TCP's protocol efficiency can vary significantly, and even under good radio conditions, can still vary significantly.

B. *Modeling of Mobile Data Networks*

A number of previous works [10-12] have studied the modeling of 3G/HSPA and 4G/LTE networks using queuing theory, continuous-time, and discrete-time Markov chains. These studies all modeled packet arrivals to a cell as a Poisson process with exponentially distributed per-packet service time.

In the model developed by Ghaderi *et al.* [10] they assumed TCP flows to share the cell capacity equally. Another study by Johansson *et al*. [11] employed queueing models to analyze and compare the throughput performance of single and multi-carrier HSDPA networks. However neither work considered TCP's throughput limit over mobile networks which can be substantially lower than the available bandwidth as discussed earlier.

In another study Bodrog *et al.* [12] developed an equivalent queuing network model for a mobile cell implementing HSDPA UTRAN. They derived the congestion loss probability and round trip time for feeding into TCP Reno's throughput formula to estimate the resultant TCP throughput performance. However, recent HSPA/LTE network measurements [3-5] revealed that packet losses are more likely non-congestion-related as the link buffer is often larger than TCP's receiver advertized window size. Moreover, the modeled TCP Reno has long been replaced by other TCP variants (e.g., TCP CUBIC in Linux) and thus the results may not be applicable to today's networks. By contrast, our study adopted TCP throughput parameters from measurements of real systems and networks, and thus can more accurately reflect their performance impacts.

Our study differs from the existing works in three major ways. First, to our knowledge the impact of protocol-limited throughput on cell capacity utilization has not been studied before. Our work develops a model to relate protocol efficiency to cell utilization to quantify its performance impact. Second, our study reveals that channel bandwidth can also degrade cell capacity utilization substantially and this calls for a new look on the interplay between cell capacity and channel bandwidth in planning network infrastructures. Third, our study incorporates measured (as opposed to modeled) protocol properties obtained from production TCP implementations (as opposed to obsolete implementations) in real mobile networks which enables us to evaluate the practical impact of protocol and channel throughput limits.

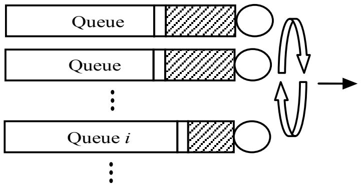

Fig. 1. Round-robin scheduling in the Node-B of HSPA/LTE networks.

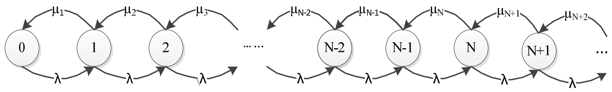

Fig. 2. State transition diagram for modeling a mobile cell.

III. System Model

In this section we present a system model to incorporate the effects of resource allocation in modern mobile networks. The system model also incorporates TCP protocol efficiency and channel bandwidth limit in a Markovian birth-death process with state-dependent service rates. We derive the mean service response time as the performance metric for mobile Internet services. Finally we use the model to obtain the cell capacity loss due to protocol efficiency and channel bandwidth limit.

A. *Mobile Cell Bandwidth Allocation*

Cell bandwidth allocation in modern HSPA and LTE networks is done by the Node-B [1,11]. At the Node-B, transmission scheduling is typically done using the round-robin (RR) scheduler as depicted in Fig. 1. Each user is allocated a dedicated buffer at the link layer. Packets arriving at the Node-B destined to a user will queue up at the user's buffer awaiting transmission. The Node-B then scan through the buffers in a round-robin manner to retrieve a packet for transmission if (a) the queue is non-empty; and (b) the link layer channel to the user has bandwidth available.

Let C be the total data bandwidth of the cell and n be the number of users in the cell. Assuming all the queues are non-empty, then the transmission rate for each user, denoted by r, is equal to

$$r = C / n \,. \tag{1}$$

However this transmission rate may not be achievable under two cases. First, each user's radio channel is limited to a maximum bandwidth dictated by the mobile standard. For example, HSPA channels have a maximum per-channel bandwidth of 7.2 Mbps. Thus in a cell with a capacity of 78Mbps, a user will still be limited to a maximum of 7.2 Mbps even if it is the only user in the cell. We use r_{max} to denote this *channel bandwidth limit*.

Second, existing TCP may reach its throughput limit before reaching either the bandwidth limit in (1) or r_{max}. To model this we use r_{tcp} to represent TCP's throughput limit when operating under ideal conditions in a cell with unlimited capacity. TCP's protocol efficiency as discussed in Section II-A is then equal to r_{tcp}/r_{max}. As r_{tcp} cannot be larger than r_{max} we only need to consider r_{tcp} in the rest of the derivations.

B. *Markov Chain Model*

We model a mobile cell using the Markov chain model depicted in Fig. 2. The state, denoted by k, represents the number of users in the system. Similar to previous works [10-12] we assume users to arrive at the system according to a Poisson process of rate λ. The cell service rate is also Poisson with mean μ_k. Note that the cell service rate is state-dependent due to the throughput limit of TCP. With a cell capacity of C, each user will have a throughput of

$$r = \min\left\{ C/k, r_{tcp} \right\} \tag{2}$$

Intuitively, when there are few users, i.e., k is small thus $r_{tcp}>C/k$, the per-user throughput is then limited by TCP's maximum throughput under the mobile standard (e.g., r_{tcp}=5Mbps under r_{max}=7.2Mbps 3G standard in a C=100Mbps cell). By contrast, when there are many users, i.e.,

$$k \geq N = \left\lceil C/r_{tcp} \right\rceil \tag{3}$$

the share of bandwidth each user has will then fall below TCP's upper limit r_{tcp}, and is then limited by the cell's capacity. Therefore the state-dependent service rate, aggregated from all k users, can be computed from

$$\mu_k = \begin{cases} k r_{tcp}, & 0 \leq k < N \\ C, & k \geq N \end{cases} \tag{4}$$

Let p_k be the limiting probability that the system is in state k. In the steady-state we have

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda}{\mu_{i+1}} \qquad k = 1,2,... \tag{5}$$

As the service rate μ_k is state-dependent, we formulate it into two cases:

$$p_k = \begin{cases} p_0 \left(\dfrac{\lambda}{r_{tcp}} \right)^k \dfrac{1}{k!} & k \leq N-1 \\ p_0 \left(\dfrac{\lambda}{r_{tcp}} \right)^{N-1} \left(\dfrac{\lambda}{C} \right)^{k-N+1} \dfrac{1}{(N-1)!} & k \geq N \end{cases} \tag{6}$$

Since $\sum_{i=0}^{\infty} p_i = 1$, we can solve for p_0 from

$$p_0 = \left(\sum_{k=0}^{N-1} \left(\frac{\lambda}{r_{tcp}} \right)^k \frac{1}{k!} + \sum_{k=N}^{\infty} \left(\frac{\lambda}{r_{tcp}} \right)^{N-1} \left(\frac{\lambda}{C} \right)^{k-N+1} \frac{1}{(N-1)!} \right)^{-1} \tag{7}$$

Rearrange terms we have

$$p_0 = \left(\sum_{k=0}^{N-1} \left(\frac{\lambda}{r_{tcp}} \right)^k \frac{1}{k!} + \left(\frac{\lambda}{r_{tcp}} \right)^{N-1} \frac{1}{(N-1)!} \frac{\lambda}{C-\lambda} \right)^{-1} \tag{8}$$

Substituting (8) back into (6) we can then obtain the limiting probabilities.

C. Performance Metric for Mobile Internet

Unlike voice services, the performance of mobile Internet cannot be adequately measured using traditional call-based metrics such as blocking probability. For services such as web browsing and file download, the service response time metric would be more representative of the user's experience.

Consider the states $\{k \mid 0 \leq k < N\}$. For these states the bottleneck is TCP's throughput limit as the cell has more than sufficient capacity. Assuming each user download one unit of data, then the mean service response time will be equal to $1/r_{tcp}$. For the states $\{k \mid k > N\}$, the bottleneck is shifted to the cell capacity as each user has a throughput limited to C/k, and the corresponding mean service response time is equal to k/C.

Thus the overall mean service response time can be computed from the conditional expectation with c=C and r=r_{tcp}:

$$S(c,r) = \sum_{k=0}^{N-1} p_k \frac{1}{r} + \sum_{k=N}^{\infty} p_k \frac{k}{c} \tag{9}$$

Substituting (6) into (9) and noting that

$$\sum_{k=N}^{\infty} k \left(\frac{\lambda}{C} \right)^{k-N+1} = \lambda \left(\frac{\lambda}{(C-\lambda)^2} + \frac{N}{C-\lambda} \right),$$

we have

$$S(c,r) = \frac{\displaystyle\sum_{k=0}^{N-1} \left(\frac{\lambda}{r} \right)^k \frac{1}{k!} \frac{1}{r} + \left(\frac{\lambda}{r} \right)^{N-1} \frac{1}{(N-1)!} \frac{\lambda}{c} \left(\frac{\lambda}{(c-\lambda)^2} + \frac{N}{c-\lambda} \right)}{\displaystyle\sum_{k=0}^{N-1} \left(\frac{\lambda}{r} \right)^k \frac{1}{k!} + \left(\frac{\lambda}{r} \right)^{N-1} \frac{1}{(N-1)!} \frac{\lambda}{c-\lambda}} \tag{10}$$

which relates the mean service response time to TCP's protocol efficiency via r=r_{tcp}. We can also compute the mean number of users in the cell from

$$U(c,r) = \sum_{k=0}^{\infty} p_k k \tag{11}$$

Substituting (6) into (11), we have

$$U(c,r) = \frac{\displaystyle\sum_{k=0}^{N-1} \left(\frac{\lambda}{r} \right)^k \frac{1}{k!} k + \left(\frac{\lambda}{r} \right)^{N-1} \frac{1}{(N-1)!} \lambda \left(\frac{\lambda}{(c-\lambda)^2} + \frac{N}{c-\lambda} \right)}{\displaystyle\sum_{k=0}^{N-1} \left(\frac{\lambda}{r} \right)^k \frac{1}{k!} + \left(\frac{\lambda}{r} \right)^{N-1} \frac{1}{(N-1)!} \frac{\lambda}{c-\lambda}} \tag{12}$$

By comparing the mean number of users in the cell to N in (3) we can see the relation between traffic load and operating regime of the cell (c.f. Section IV-A).

D. Protocol-limited Capacity Loss

Using (10) we can proceed to investigate the impact of TCP protocol efficiency on network capacity loss. The idea is that if TCP's protocol efficiency is 1, i.e., r_{tcp}=r_{max}, then one can in fact use a mobile cell with *less* capacity and yet still be able to achieve the *same* mean service response time.

Formally, let $C_{tcp}(s)$ be the cell capacity required to achieve a mean service response time no longer than s:

$$C_{tcp}(s) = \min\left\{ c \mid S(c, r_{tcp}) \leq s \right\} \tag{13}$$

Now if TCP is 100% efficient then r_{tcp}=r_{max} so the cell capacity needed to achieve the same mean service response time will be given by

$$C_{opt}(s) = \min\left\{ c \mid S(c, r_{max}) \leq s \right\} \tag{14}$$

It can be shown that $S(c,r)$ is a decreasing function of r, and thus $C_{opt}(s) < C_{tcp}(s)$ if $r_{tcp} < r_{max}$. We define L_{tcp} to be the ratio of cell capacity lost due to protocol inefficiency:

$$L_{tcp}(s) = \left(C_{tcp}(s) - C_{opt}(s) \right) / C_{tcp}(s) \tag{15}$$

which quantifies the extent of protocol-limited capacity loss.

Table 1. Throughput limits for various TCP variants in 3G networks [3][14].

TCP Variants	0.1% packet loss	0.5% packet loss	1.0% packet loss
TCP Cubic (Mbps)	2.89	1.15	0.73
TCP Westwood (Mbps)	3.67	2.03	1.44
TCP Vegas (Mbps)	1.41	0.64	0.49
Accelerated TCP (Mbps)	4.25	4.23	4.15
Optimal (Mbps)	4.53	4.53	4.53

Table 2. Throughput limits of TCP Cubic and Accelerated-TCP measured from a production LTE network [2].

{RSRP, SINR}	{-64, 27}	{-81, 27}	{-90, 27}	{-110, 15}
TCP Cubic (Mbps)	34	36	30	27
Accelerated TCP (Mbps)	74	73	53	47
Maximum UDP Goodput (Mbps)	81	80	55	50

E. *Channel-limited Capacity Loss*

The previous analysis reveals another interesting bottleneck – the channel bandwidth as imposed by the mobile standard. For example, 3G has a maximum channel bandwidth of 7.2 Mbps. Thus even if a mobile cell with 78 Mbps cell capacity has only one user, that user is still limited to 7.2 Mbps. Applying the same reasoning as protocol-limited throughput suggests that a mobile operator can achieve the same mean service response time with *less* cell capacity if one can *raise* the channel bandwidth limit.

In the ideal case users can fully utilize all cell bandwidth, with each user receiving a bandwidth of *C*/*n*, where *n* is the number of users. To model this case we can simply replace the service rate in (4) by

$$\mu_k = C \textit{ for } k \geq 0 \tag{16}$$

and the Markov chain reduces to the classic *M*/*M*/1 queue with mean service response time given by

$$S_{min}(c) = \frac{\lambda}{(c-\lambda)c} \tag{17}$$

Hence we can compute the capacity needed from

$$C_{min}(s) = \min\{c \mid S_{min}(c) \le s\} \tag{18}$$

and the corresponding channel-limited capacity loss from

$$L_{ch}(s) = \left(C_{opt}(s) - C_{min}(s)\right) / C_{opt}(s) \tag{19}$$

This channel-limited capacity loss will be useful to mobile operators for evaluating the potential performance gains from upgrading to a faster mobile standard (e.g., from 3G to 3.5G), even without increasing the cell capacity.

IV. Performance Evaluation

In this section we apply the system models developed in Section III to compute numerical results using real-world system parameters to study the performance impact of TCP's protocol efficiency. For mobile cell bandwidth we adopted *C*=78 Mbps [13] to represent a HSPA network cell and *C*=980 Mbps [1] to represent a LTE network cell.

As discussed in Section II, TCP protocol efficiency varies depending on many factors. We adopted the HSPA-based simulation results by Chan [3] for TCP CUBIC, TCP Vegas, and TCP Westwood, and the experimental results from Liu and Lee [14] for the Accelerated TCP. Table 1 summarizes the various TCP variants' throughput under three packet loss rates, as well as the optimal throughput – equal to the average raw network bandwidth available. For the LTE case we only adopted experimental results from Liu and Lee [2] to compare TCP CUBIC and Accelerated TCP. Simulation results are not available as accurate modeling of LTE behavior remains on-going research. Table 1 and 2 show the throughput results of a single TCP connection. Multiple concurrent TCP connections may result in higher aggregate throughput and the effect can be incorporated by corresponding adjustments to the protocol efficiency values.

A. *Service Response Time*

We first consider the service response time for a user downloading a file (e.g., web object or a photo) using TCP. We assume the file size to be exponentially distributed with a mean file size of 0.126 MB [15]. We first consider the HSPA case in Fig. 3, which plots the mean service response time (in seconds) versus traffic load (defined as *λ/C*). We observe that the mean service response time in all cases stayed above zero but did not increase appreciably until reaching a high traffic load (e.g., above 0.75). The same observation also applies to the LTE case in Fig. 4. This is unlike ordinary queuing systems where the mean service response time typically increases from zero as the traffic load increases.

To see why, recall from Section III that if the system operates in state $k<N$, then there is more per-user bandwidth than can be utilized by TCP (i.e., $C/k>r_{tcp}$) and in this case the system throughput is operating in the *protocol-limited regime*. By contrast, if $k \geq N$, then the per-user available bandwidth is lower than the TCP throughput limit, and thus the system becomes cell capacity limited, i.e., operating in the *capacity-limited regime*. A stochastic system will spend more time at the lower states (i.e., small *k*'s) at light traffic load and the service response time will then be primarily limited by protocol efficiency. The parameter *N* thus represents a break-even point, at or beyond which the system will not suffer from protocol deficiencies. Using (3) we can compute this break-even point for HSPA and LTE networks for all TCP variants. Comparing to the mean number of users in the system according to (12) these are equivalent to a traffic load of over 0.9 for all TCP variants. This result suggests that in practice protocol efficiency will be a significant factor to service response time unless the mobile cell is operating at close to full load (i.e., > 0.9).

B. *Network Capacity Loss*

Mobile network infrastructure is extremely costly. Thus an interesting question is to what extent the deployed network capacity is lost due to TCP's protocol efficiency. To answer this question we apply the equations from Section III-D to compute the protocol capacity losses in Fig. 5 and 6 for HSPA and LTE networks respectively. Note the vertical lines mark the protocol efficiencies for the various TCP variants.

Intuitively, protocol-limited capacity loss measures the amount of cell capacity that is deployed but *fails* to contribute to reducing the mean service response time. A protocol capacity loss of 0.4 means that one can achieve the same mean service response time by using a network cell with only 1–0.4=0.6 (i.e.,

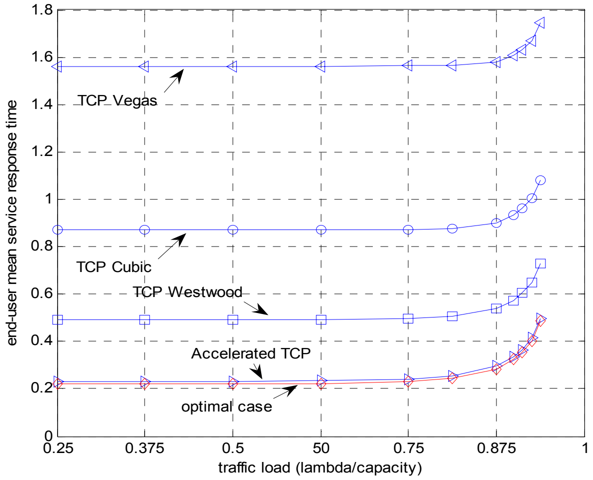

Fig. 3. Comparison of end-user mean service response time of four TCP variants in a HSPA network with 0.5% loss rate and *C*=78 Mbps capacity.

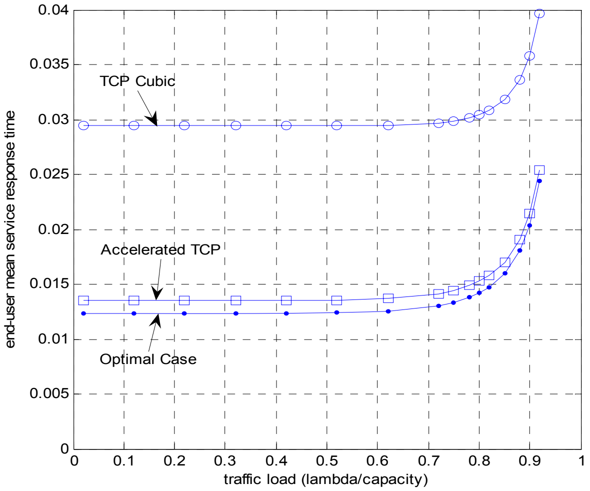

Fig. 4. Comparison of end-user mean service response time of TCP Cubic and Accelerated-TCP in a LTE network with *C*=980 Mbps capacity.

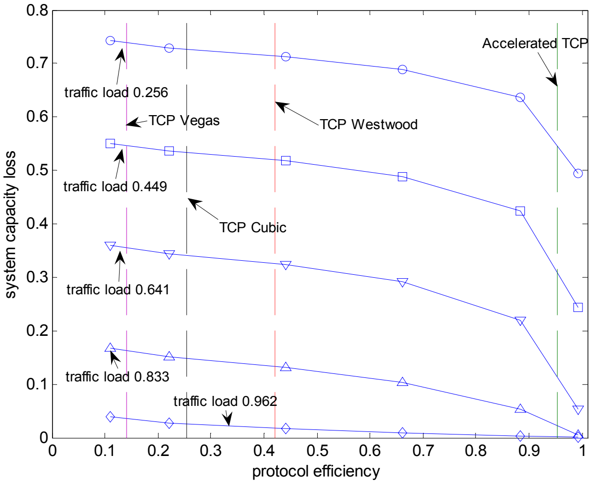

Fig. 5. Comparison of network capacity loss due to TCP protocol efficiency at various traffic loads.

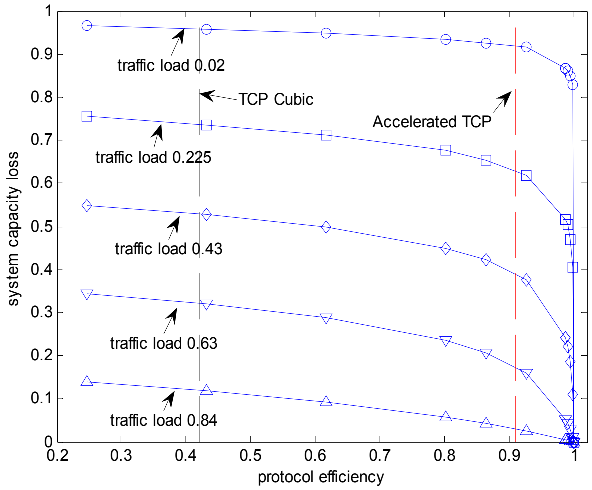

Fig. 6. Comparison of network capacity loss due to TCP protocol efficiency at various traffic loads.

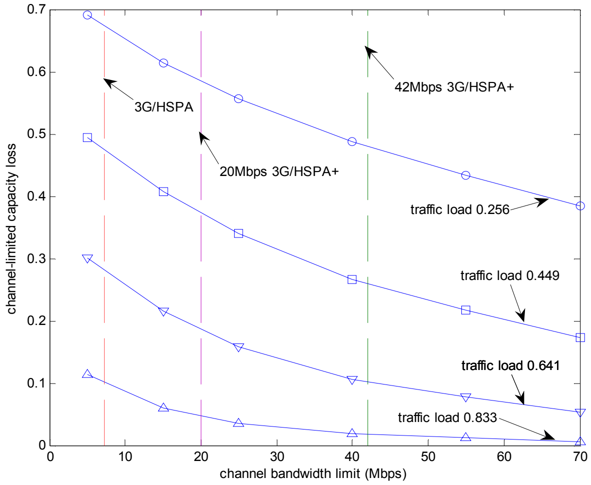

Fig. 7 Comparison of 3G/HSPA network capacity loss due to channel bandwidth limit at various traffic loads.

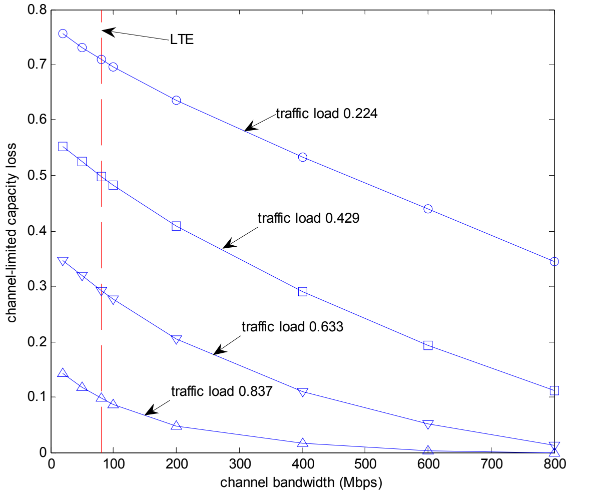

Fig. 8 Comparison of LTE network capacity loss due to channel bandwidth limit at various traffic loads.

60%) of the original cell capacity if users can utilize *all* the channel bandwidth available, i.e., $r_{tcp}=r_{max}$.

The results in Fig. 5 and 6 show that the protocol-limited capacity losses are surprisingly high. For example, TCP CUBIC with a protocol efficiency of 25.4% has a protocol-limited capacity loss over 0.5 at a traffic load of 0.449 and below. More surprisingly, even the most efficient TCP variant, Accelerated TCP with a protocol efficiency of 93.4%, exhibited a protocol-limited capacity loss over 0.3 at the same traffic load. These results can be explained by the observation that below the break-even point *N*, the cell is operating in the protocol-limited regime where the cell bandwidth cannot be fully utilized. Thus increasing the cell capacity will have no impact to the service response time below the break-even point. While increasing cell capacity does reduce service response time at and above the break-even point, the proportion of time the system stays at those states is proportional to the traffic load. Hence unless operating at a high traffic load, the capacity loss will become far more significant than what the protocol efficiency suggests.

Next we compute the channel-limited capacity losses versus the different channel limiting throughput in Fig. 7 and 8 using the definitions from Section III-E. In contrast to protocol-limited capacity loss, here we assume users can fully utilize the channel bandwidth and thus any capacity lost is solely due to underutilization of cell capacity due to the channel bandwidth limit. In contrast to protocol-limited capacity loss, the network capacity losses in Fig. 7 and 8 decrease linearly for higher channel bandwidth limits. This strongly suggests that a mobile operator could improve service response time without increasing cell capacity simply by upgrading to a mobile standard with higher channel bandwidth. For example, upgrading from 7.2Mbps 3G/HSPA to 42Mbps 3G/HSPA+, one could reduce the capacity loss from 0.28 to 0.12 at a traffic load of 0.64. Similar conclusions can be drawn for LTE networks as 100Mbps+ LTE standards are already in the works. This interesting result suggests that there is still room for optimization in the existing mobile infrastructure, even without costly cell bandwidth upgrades

V. Summary and Future Work

This work offers a first look into the impact of TCP protocol efficiency on mobile network capacity. Given the extremely high cost of mobile network infrastructure, the loss of even a few percent of the network capacity can be very costly. Yet our analysis revealed that unless the cell operates at high traffic load, both HSPA and LTE networks would suffer from significant capacity losses. This calls for the need to further optimize TCP for use over mobile data networks. In addition, our analysis of channel-limited capacity loss revealed that upgrading the mobile base station to higher-speed mobile standards is an effective way to improve cell bandwidth utilization, even if the underlying cell capacity is kept unchanged. In addition, the analysis also revealed an inter-play between cell capacity and protocol/channel bandwidth limit. In particular, too large a cell capacity may not necessary improve service quality significantly as the bottleneck would be shifted to the protocol/channel bandwidth limit. This opens up a new problem/opportunity in the allocation of bandwidth to mobile cells, as a cost-effective allocation will need to incorporate the impact of protocol efficiencies and channel bandwidth limits, in addition to traffic load and other network parameters. These have substantial economic significance to mobile operators and thus warrant further investigations.

Acknowledgment

The authors would like to thank the anonymous reviewers for their comments and suggestions in improving this paper. This work was funded in part by the Innovation and Technology Fund (ITS/014/10 and ITS/146/11) provided by the Innovation and Technology Commission, HKSAR. Any opinions, findings, conclusions or recommendations expressed in this material/event (or by members of the project team) do not reflect the views of the Government of the Hong Kong Special Administrative Region, the Innovation and Technology Commission or the assessment panel of the Innovation and Technology Support Programme of the Innovation and Technology Fund.

Reference

[1] E. Dahlman, S. Parkval, J. Skold and P. Beming, "3G Evolution: HSPA and LTE for Mobile Broadband," 1st edition, 2007.

[2] K. Liu and J. Y. B. Lee, "Mobile Accelerator: A New Approach to Improve TCP Performance in Mobile Data Networks," in *Proc. 7th IEEE International Wireless Communications and Mobile Computing Conference (IWCMC)*, Istanbul, Turkey, July 5-8, 2011.

[3] Stanley C.F. Chan, "A Novel Link Buffer Size and Queue Length Estimation Algorithm and its Application on Bandwidth-Varying Mobile Data Networks," M.S. Thesis, Dept. Information Engineering, The Chinese University of Hong Kong, Hong Kong, 2011

[4] X. Lin, A. Sridharan, S Machiraju, M, Seshadri and H. Zang, "Experiences in a 3G Network: Interplay between the Wireless Channel and Applications," in *Proc. ACM MobiCom*, 2008, pp. 211-222.

[5] F. Ren and C. Lin, "Modeling and Improving TCP Performance over Cellular Link with Variable Bandwidth," *IEEE Trans. Mobile Computing*, Vol. 10, No. 8, Aug, 2011

[6] S. Ha, I. Rhee and L. Xu, "CUBIC: A New TCP-Friendly High-Speed TCP Variant," in *Proc. International Workshop on Protocols for Fast and Long Distance Networks*, 2005.

[7] M. Allman, V. Paxson and E. Blanton, "TCP Congestion Control," *Request for Comments 5681,* September 2009.

[8] S. Mascolo, C. Casetti, M. Geria, M. Y. Sanadidi and R. Wang, "TCP Westwood: Bandwidth Estimation for Enhanced Transport over Wireless Links," *in Proc. ACM SIGMOBILE*, July 2001.

[9] L. S. Brakmo, S. W. O'Malley, and L. L. Peterson, "TCP Vegas: New techniques for congestion detection and avoidance," in *Proc. ACM SIGCOMM*, London, U.K., Oct 1994, pp.24-35.

[10] M. Ghaderi, R. Boutaba, and G. W. Kenward, "Stochastic Admission Control for Quality of Service in Wireless Packet Networks," in *Proc. 4th IFIP Networking Conference*, pp. 1309-1320, May 2005.

[11] K. Johansson, J. Bergman, D. Gerstenberger, M. Blomgren, and A. Wallen, "Multi-Carrier HSPA Evolution," *Vehicular Technology Conference,* vol. 1, no. 5, pp. 26-29, April 2009.

[12] L. Bodrog, G. Horvath, and C. Vulkan, "Analytical TCP throughput model for high-speed downlink packet access," *Software, IET*, vol.3, no.6, pp.480-494, December 2009.

[13] C. Cheevallier, C. Brnner, A. Garavaglia, K. P. Murray and K.R. Baker, "WCDMA (UMTS) Deployment Handbook: Planning and Optimization Aspects," 1st edition, 2006.

[14] K. Liu and J. Y. B. Lee, "Improving TCP Performance over Mobile Data Networks with Opportunistic Retransmission,". in *Proc. Of Wireless Communication Network Conference (WCNC)*, April 7-10, 2013.

[15] D. P. Heyman, T. V. Lakshman, A. L. Neidhadt, "A New Method for Analysing Feedback-Based Protocols with Applications to Engineering Web Traffic over the Internet," in *Proc. ACM SIGMETRICS*, pp. 24-38, 1997.