Analysis of LTE Radio Load and User Throughput

Jari Salo[1] and Eduardo Zacarías B.[2]

Nokia Networks [1]Taguig City, Phillipines and [2]Ulm, Germany

Abstract

A recurring topic in LTE radio planning pertains to the maximum acceptable LTE radio interface load, up to which a targeted user data rate can be maintained. We explore this topic by using Queuing Theory elements to express the downlink user throughput as a function of the LTE Physical Resource Block (PRB) utilization. The resulting formulas are expressed in terms of standardized 3GPP KPIs and can be readily evaluated from network performance counters. Examples from live networks are given to illustrate the results, and the suitability of a linear decrease model is quantified upon data from a commercial LTE network.

Keywords

LTE, Traffic Model, Processor Sharing, Network Measurements

1. Introduction

A key topic in radio network planning concerns mapping statistics generated in the radio network layer, to the end-user experienced performance. In practical operation, a question relating planning and current conditions often arises: "what is the maximum LTE radio load, so that the user throughput is still at a given level?". Somewhat surprisingly, there are no practically useful engineering analyses available to this question in the literature. This is the main motivation for this study. A simple answer to the question is available by exploiting results from IP networking literature and basic queuing theory. In order to make those results practical, this paper states the existing theoretical results in terms of LTE radio utilization metrics, which can be easily computed from network performance statistics available in commercial LTE base station products.

Earlier work: The user throughput for elastic traffic transmitted over fixed-bandwidth, non-wireless links has been thoroughly investigated in engineering literature. An overview and references can be found in [1]. Wireless CDMA network user throughput has been analyzed in [2] and in other works of the authors thereof. Radio interface flow-level scheduler analysis from relatively theoretical viewpoint has been presented in [3] and [4]. An overview of LTE scheduling has been presented in [5]. The idea of modelling LTE radio scheduler using M/G/1 is obviously not unheard of, see for example [6]. However, none of the aforementioned contributions present any validation of the theoretical results against real-world network data or the results are presented in terms of statistics that are not available in real-world networks.

In this paper, the focus is on the average user throughput over the LTE radio interface. In particular, the so called M/G/1 Processor Sharing (PS) approach is used to express user throughput in terms of LTE radio interface Physical Resource Block (PRB) utilization, the statistics of which are available in every commercial LTE system. For details on the LTE system details, the reader is referred to the well-established literature, such as [7, 8, 9].

DOI: 10.5121/ijcnc.2017.9603

Contributions: The contributions of this paper include the following.

- Formulation of LTE user throughput in terms of the M/G/1 PS model by using PRB utilization as the load metric.
- Verification of the validity of the M/G/1 PS approach by comparing to live network measurements.
- Extension to rate-capped user throughput by using the M/G/R PS model for non-integer R.
- Application to load balancing between frequency layers, including closed-form formula for the optimal traffic balancing ratio for the two-layer case.

The results can be used for both FDD and TDD variants of LTE. Although in principle the general approach is applicable to both downlink and uplink, the uplink case is dependent on the power control and link adaptation implementation, which leaves more degrees of freedom to be considered. For this reason, throughout the paper ‘throughput’ refers to the downlink case, with the uplink applicability left for further study.

In Section 2, the average user throughput is formulated as a function of the number of active data flows sharing radio scheduler resources. In Section 3 these results are related to physical resource block (PRB) utilization by means of the well-known M/G/1 processor sharing formula, resulting in $\propto \frac{1}{1-\rho}$ degradation in user throughput with ρ denoting PRB utilization of the serving cell. An extension to the case where user throughput is externally rate-limited is also given. In Section 4 application of the result to traffic balancing between frequency layers is given. Section 5 discusses some aspects of the impact of cell interference coupling. In Section 6 live network measurement examples are provided to validate the results. Finally, conclusions are presented.

2. Throughput versus Number of Active Users

2.1 Basic Assumptions

The following assumptions are made:

- Full buffer traffic model so that, in the absence of other users, a single user obtains all available radio resources. For example, constant bit rate streaming traffic would not satisfy this condition.
- If there is more than one user, the scheduler shares radio resources equally, on average, between users. This fair sharing principle is assumed independently of the radio conditions of the UEs sharing the scheduler resources.

The first assumption will be relaxed in Section 3.3, and the second assumption will be discussed in Section 5. The notion of “radio resource” could be defined in various ways, but in case of the LTE radio interface, it is convenient to choose Physical Resource Blocks (PRBs) as the resource being shared. For LTE downlink, PRB utilization can be equated with transmit power utilization as long as physical layer and common channel overhead are properly taken into account.

With the assumptions above, instantaneous user throughput with x active users downloading simultaneously would be $\frac{1}{x}$ of the maximum throughput. A UE is said to be *active* if there are data remaining in the transmit buffer[1]. As different UEs in the cell start and finish their data transfers, the number of active UEs (x), and hence also the instantaneous user throughput, changes over time. An interesting metric is the average throughput experienced by UEs in the cell, which is discussed in the next section.

2.2 Two user throughput metrics

Consider a UE located somewhere in a cell and experiencing certain radio quality. In the absence of any other users, the UE is allocated all available PRBs and receives some throughput T_1, where the subscript '1' emphasizes that there is one active user (i.e., $x = 1$). If there were $x > 1$ active UEs in the cell, the throughput of the user would be T_1/x instead. The maximum achievable user throughput, T_1, depends on the user location in the cell, radio conditions, number of transmit antennas, and so on. The *average* user throughput T_{ue} is defined as the expected value of T_1/x for positive integer x, in other words

$$T_{ue} = E\left[\frac{T_1}{x}\right], \quad x \geq 1, \tag{1}$$

where $E[\cdot]$ denotes expected value and T_1 and $x \geq 1$ are the random variables being averaged. The user radio conditions, and hence the maximum throughput T_1, can be assumed statistically independent of the number of active UEs in the cell, and (1) can thus be written as

$$T_{ue} = E[T_1]E\left[\frac{1}{x}\right], \quad x \geq 1. \tag{2}$$

The term $C = E[T_1]$ will be called cell capacity in this paper.

Unfortunately, the second term $E[1/x]$, which is the average of the inverse of the number of active UEs, cannot be always computed since it is not usually available as a radio counter. On the other hand, the average active UEs, $E[x]$, is a standardized KPI defined in 3GPP TS 32.425 and thus commonly implemented in commercial products. Therefore, a more practical metric results if $E[1/x]$ in (2) is replaced by $1/E[x]$, or

$$T_{sch} = \frac{E[T_1]}{E[x]}, \quad x \geq 1. \tag{3}$$

In 3GPP TS 32.425 this is called scheduled "IP throughput". It should be emphasized that $E[x] \neq E[1/x]$ and for this reason T_{sch} and T_{ue} are different throughput metrics and not equal in value. The scheduled throughput can also be written as [10]

$$T_{sch} = \frac{S}{W}, \tag{4}$$

where S is the average file size (bytes) and W is the average file transfer time. This form of the scheduled throughput is often used in fixed-line IP network throughput analysis where it goes by the name "flow throughput". An application to wireless network setting can be found [2] and many others published since.

The scheduled throughput T_{sch} is always lower than user throughput T_{ue}. This is a direct result of the concavity of $1/x$ and Jensen's inequality: $E[1/x] > 1/E[x]$ and subsequently $T_{ue} > T_{sch}$.

[1] The terms 'UE' and 'user' are used interchangeably. The number of active UEs is different from the number of RRC-connected UEs since a UE can be RRC-connected without having any data to receive or send.

While typically one would be more interested in the user throughput T_{ue}, in this paper the focus is on the scheduled throughput T_{sch} because T_{ue} is not usually computable from LTE base station performance counters. A discussion on the differences between different throughput metrics can be found in [10, 11].

3. Throughput versus PRB Utilization

In the previous section, it was shown that average scheduled throughput is inversely proportional to the average number of active UEs. However, the number of active UEs is perhaps not the most intuitive KPI for characterizing network load. A more commonly employed yardstick of network load is the fraction of utilized PRBs. It would, therefore be of practical interest to have some rule that relates PRB utilization to user throughput. Such is the topic of this section.

3.1 Definition of Radio Load

In fixed link throughput analysis, the link utilization ρ is usually expressed in terms of average flow size S (bits) and flow arrival rate λ (1/sec) as

$$\rho = \frac{\lambda S}{C}, \tag{5}$$

where C is the link bandwidth (bits per second). The resource shared between users, in this case, is the link bandwidth C. For the LTE radio interface, another option is to use LTE physical resource blocks (PRBs) instead. In this context, the LTE cell scheduler can be considered a “processor” that shares the PRBs equally between active UEs. Furthermore, the PRB utilization is available from system counters in any practical LTE system.

To map bytes to PRBs, the file size S needs to be converted to PRBs. For example, given an end user spectral efficiency of $\eta = 1$ bits per second per Herz, one PRB can fit 180 bits of end user data after channel coding and physical layer overhead. Hence a one MegaByte web page would generate scheduler PRB load of $10^6 \times 8/180 \approx 44000$ PRBs. More generally,

$$\rho = \frac{1}{180}\frac{\lambda S}{\eta R_{\text{prb}}}, \tag{6}$$

where η is the average cell spectral efficiency in bits per second per Hz, R_{prb} is the PRB rate (e.g., 10^5 PRBs per second for a 20MHz cell), and the scaling factor 180 comes from the standardized PRB bandwidth of 180kHz.

In practice, it is not necessary to know the traffic parameters λ and S since the PRB utilization ρ can be extracted from network statistics directly. This is in contrast to cell capacity C, which is less straightforward to estimate from counters.

3.2 Throughput versus PRB utilization via M/G/1 PS model

To relate the number of active UEs, $E[x]$ or $1/E[x]$, in (1)–(3) to PRB utilization, some statistical assumptions on the arrival rate λ of the user data flows need to be made. In the IP engineering literature, it is common practice to model TCP flow throughput in wireline links using the so called M/G/1 Processor Sharing model. The M/G/1 PS model assumes that λ, the number of data flow arrivals per time unit, has Poisson distribution. This is typically assumed well-justified, one reason being that it leads to simple formulas. In the sequel, the same approach is applied to LTE radio throughput.

It is well-known from basic textbooks [12], that under the assumption of Poisson distributed λ the proportion of time with x active UEs is $\pi_x = (1-\rho)\rho^x$, where x is a non-negative integer and ρ is the load defined in (6). With this information, the expected values $E[1/x]$ and $E[x]$ in (2) and (3) can be computed. The details can be found in a number of references and the result for user throughput is [12, 10, 11, 2]

$$T_{ue} = \frac{C(1-\rho)}{\rho} \ln \frac{1}{1-\rho}, \tag{7}$$

while the scheduled throughput is simply

$$T_{sch} = C(1-\rho). \tag{8}$$

As mentioned, the measurement of user throughput T_{ue} is not very straightforward and not typically implemented in commercial base station products, while the scheduled throughput T_{sch} is simpler to compute and has been standardized by 3GPP [13].

3.3 Throughput versus PRB utilization via M/G/R PS model

Consider the case where, due to some throughput limiting mechanism, a UE is able to use only a $1/R$th portion of the cell PRB resources even when there are no other active UEs. A typical example is the mobile operator capping the rates according to contractual conditions. A suitable model, in this case, is the M/G/R PS, where R defines the fraction of PRBs the UE is allocated. The M/G/R processor sharing version of the scheduled throughput in (8) can be shown [14] to become

$$T_{sch} = \frac{C}{1 + \frac{E_2(R,R\rho)}{R(1-\rho)}}, \tag{9}$$

where $E_2(R, y)$ is Erlang's second formula:

$$E_2(R,y) = \frac{\frac{y^R}{R!}\frac{R}{R-y}}{\sum_{i=0}^{R-1}\frac{R^i}{i!} + \frac{y^R}{R!}\frac{R}{R-y}}. \tag{10}$$

Here R is a positive integer and $y = R\rho > 0$ is used for brevity. Setting $R = 1$ results in the special case of (8).

It can be observed that the seemingly innocent constraint of external throughput limitation results in a considerably more involved formula, that can no longer be calculated using pencil-and-paper. Another unpleasant finding is that R is forced to be an integer which forces the single-user PRB utilization $1/R$ to be $\frac{1}{2}, \frac{1}{3}, ...$ which is unnecessarily coarse for practical use. Fortunately, it is possible to generalize (10) to real-valued R. For example, [15] gives the formula

$$E_2(R,y) = \frac{RK(R,y)}{R - y\left[1 - K(R,y)\right]}, \tag{11}$$

where

$$K(R,y) = \frac{g(R+1,y)}{1 - G(R+1,y)}, \tag{12}$$

with $g(\gamma, x)$ and $G(\gamma, x)$ denoting the probability density function and cumulative density function of the gamma distribution, respectively. In (12) $R \geq 1$ is a real number.

Fig. 1 illustrates the formula (9). In the figure, the user throughput on the vertical axis has been normalized with the cell throughput $C = E[T_1]$. It can be seen that with increasing R, the user throughput becomes increasingly limited by the external constraint and less impacted by the load from other users. For example, for $R = 5$ the user throughput is one-fifth of the cell throughput until it starts to decrease at around $\rho \approx 0.5$.

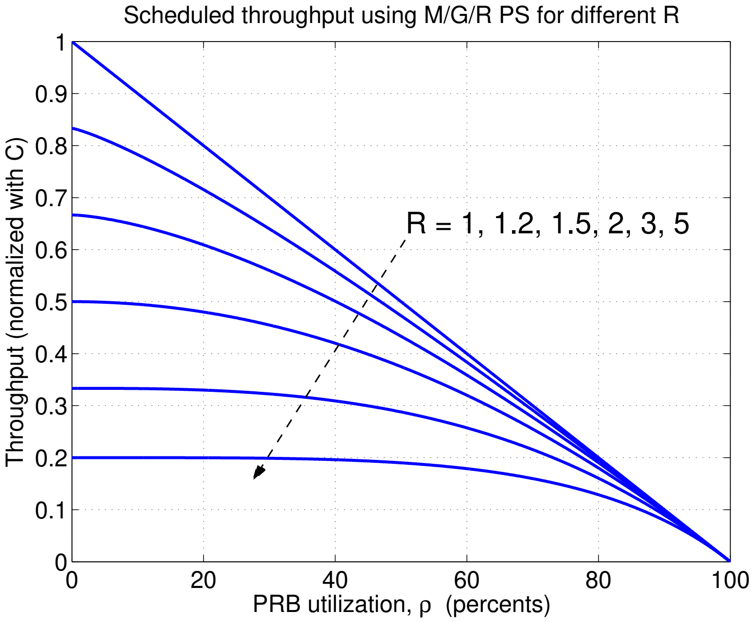

Figure 1: Normalized scheduled throughput versus PRB utilization, M/G/R PS model with different rate capping settings.

4. Application: Traffic Balancing between Frequency Layers

Considering two cells on different carrier frequencies that cover the same physical area (e.g. a 'radio sector'), an interesting question is that of how the cell loads are related to the average user throughput of the sector. From earlier discussion, the throughput in case of a single cell is defined as $T_{sch} = \frac{S}{W}$ where S is the average file size and W is the average time to transmit the file to the UE. To average the user throughput over two cells, a traffic splitting ratio can be introduced. The fraction of arriving data flows assigned to the first cell is denoted with γ which leaves the portion of $1 - \gamma$ for the second cell. For a given average file size S, the average sector user throughput (8) can be written as a weighted sum

$$
\begin{aligned}
T_{sec} &= \gamma T_{sch,1} + (1-\gamma)T_{sch,2} && (13)\\
&= \gamma \frac{S}{W_1} + (1-\gamma)\frac{S}{W_2} && (14)\\
&= \gamma C_1(1-\rho_1) + (1-\gamma)C_2(1-\rho_2), && (15)
\end{aligned}
$$

where C_i and ρ_i are the average cell throughput and load of the ith cell, respectively. The traffic splitting ratio is assumed to apply to all traffic and thus the cell loads are[2]

$$\rho_1 = \gamma\frac{\lambda S}{C_1}, \tag{16}$$

$$\rho_2 = (1-\gamma)\frac{\lambda S}{C_2}. \tag{17}$$

All told, the average user throughput across the two cells of the sector becomes

$$\begin{aligned} T_{sec} &= \gamma C_1\left(1-\gamma\frac{\lambda S}{C_1}\right) \\ &\quad + (1-\gamma)C_2\left(1-(1-\gamma)\frac{\lambda S}{C_2}\right). \end{aligned} \tag{18}$$

Fig. 2 illustrates this result in case of 10Mbps offered traffic and $C_1 = 20$Mbps. The average user throughput is shown for different traffic split between layers, for a few selected values of C_2. It can be seen that an optimum traffic split maximizing user throughput exists. Interestingly, for $C_2 = 40$Mbps all traffic should be carried by the second layer in order to maximize average user throughput.

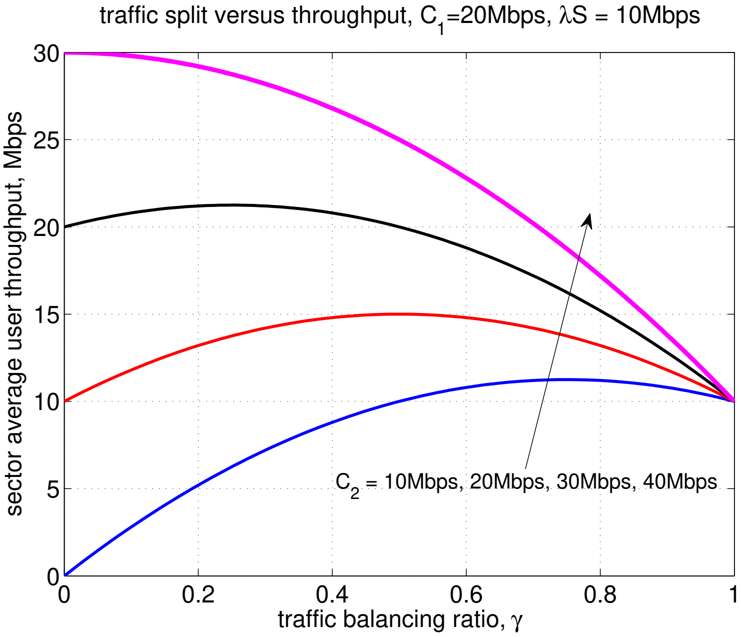

Figure 2: Average sector user throughput as a function of traffic split. $C_1 = 20$Mbps, sector total offered traffic is 10Mbps.

Fig. 2 invites the following question: if the sector offered traffic λS and the cell capacities are fixed, what is the optimum traffic balancing factor that maximizes the sector user throughput T_{sec}

[2] In this section it is more convenient to express load in terms of traffic volume and cell traffic capacity, rather than PRB utilization.

in (18)? Skipping some straightforward calculations, the optimum splitting ratio turns out be

$$\gamma_{\text{opt}} = \frac{1}{2} + \frac{1 - \frac{C_2}{C_1}}{4\hat{\rho}_1}, \tag{19}$$

where $\hat{\rho}_1 = \frac{\lambda S}{C_1}$, i.e., the load of the first cell if the second cell was non-existent. The optimum traffic split depends only on the ratio of cell capacities, not on their actual values. When $C_1 = C_2$, the even split, $\gamma_{\text{opt}} = \frac{1}{2}$ is optimum, as expected.

Fig. 3 illustrates the optimum γ for different values of $\frac{C_2}{C_1}$ and $\hat{\rho}_1$. It can be seen that if the capacity ratio $\frac{C_2}{C_1}$ is higher than a certain threshold, that depends on total sector traffic via $\hat{\rho}_1$, no positive γ_{opt} exists and to maximize user throughput the first cell should not carry any traffic at all.

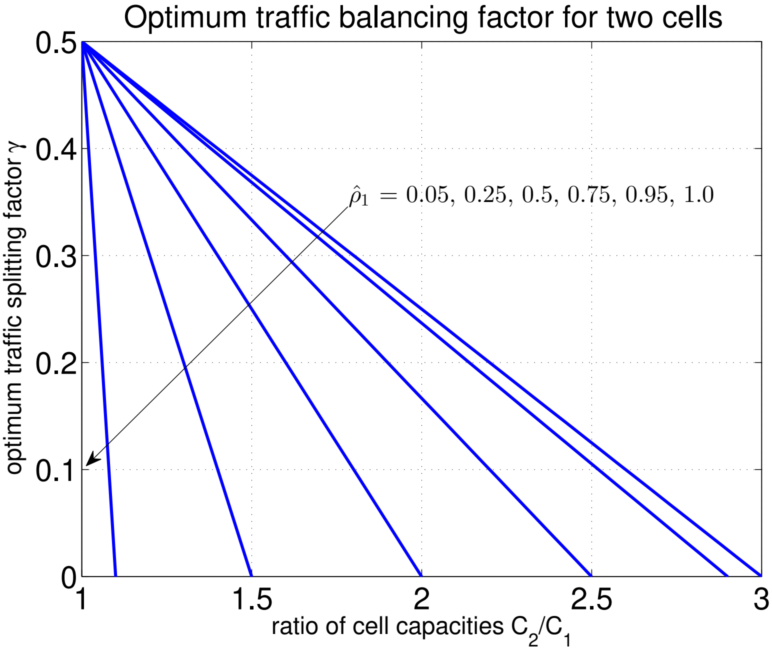

Figure 3: Optimum traffic splitting factor γ for two cells of the same sector.

5. Discussion

In (8) the ρ and T_{sch} denote the PRB utilization and user throughput of the serving cell. Nothing is said about the load of the surrounding cells. This raises the question of how $T_{sch} = C(1 - \rho)$ behaves when the neighbor cell load also increases at the same time with ρ. Interference received from neighbor cells degrades the spectral efficiency η in the serving cell, hence it is expected that the cell capacity C decreases as neighbor cell load increases. Decreasing η also increases the PRB utilization (6) since the number of bits per PRB decreases (for a fixed traffic volume S in bytes). However, such increase in ρ due to other cell interference is indistinguishable from an increase due to traffic volume, and therefore in this sense it does not directly impact throughput calculation. However, cell capacity C is expected to decrease with neighbor cell interference and hence the user throughput should decrease superlinearly. Such phenomena have however not been observed in measurements of several live networks.

Increased load can also have positive impact on spectral efficiency, namely in the form of multi-user diversity gain, where the scheduler exploits frequency selectivity of the wideband radio channel. UEs are opportunistically scheduled on the frequency subband that has the highest relative channel gain for that UE. The multiuser diversity increases with the number of UEs scheduled per TTI, which was related to PRB utilization in Section 3. Cell capacity gains of up to 50% have been simulated [16]. The impact on the present discussion is that if the serving cell load increases at the same time with the neighbor cell load, the degradation in spectral efficiency is partially compensated by multiuser scheduling gain.

Depending on the implementation the radio scheduler may also trade off spectral efficiency for latency by using free PRBs to transmit with lower channel coding rate. This improves latency since the probability of retransmission is reduced, but it also reduces spectral efficiency at low load. On the other hand, at high load, most PRBs tend to be in use and the packet scheduler is thus forced to operate at higher spectral efficiency. Other variations of the scheduling policies include prioritization according to radio quality or quality of service considerations.

6. Live Network Examples

6.1 Average Active UEs versus User Throughput

Fig. 4 shows an example of two cells serving a large number of smart-phone users. The horizontal axis is the average number of active UEs, $E[x]$, that has been extracted from hourly operations support system (OSS) performance counters over a period of two weeks. The hourly averages of x have been binned to integers and for each bin, the average flow throughput is plotted. The flow throughput shown on the vertical axis is also obtained from the OSS counters and computed according to the scheduled throughput (T_{sch}) definition in 3GPP TS 32.425. Each dot in the figure is the average UE throughput for the horizontal binned value of $E[x]$. It can be seen that the scheduled throughput scales approximately inversely proportional to the average active UEs, as predicted by theory.

6.2 PRB Utilization versus User Throughput, M/G/1 PS model

Fig. 5 illustrates scheduled throughput T_{sch} for six cells from three different networks. PRB utilization is extracted from hourly counter measurements collected over a period of two weeks and binned to 2PRB granularity. Each dot presents the average scheduled throughput of the PRB bin computed based on the 3GPP method [13]. It can be seen that the theoretical model (8) fits measurements fairly well, and throughput falls approximately linearly as a function of radio utilization. For example, at 50% utilization user throughput has dropped to half of the cell throughput while for 75% radio load a single user receives on average only 25% of maximum throughput. Such simple rule of thumbs can provide useful capacity management guidance for LTE networks, including traffic steering between different frequency layers.

6.3 Applicability of the M/G/1 PS model across cells

We now study the suitability of a linear decrease model between the scheduled throughput and the cell load (such as, for example, the one in eq. 8), across the cells of one example live network. The following results are based on hourly, cell-level performance measurement counters (PMC) for one

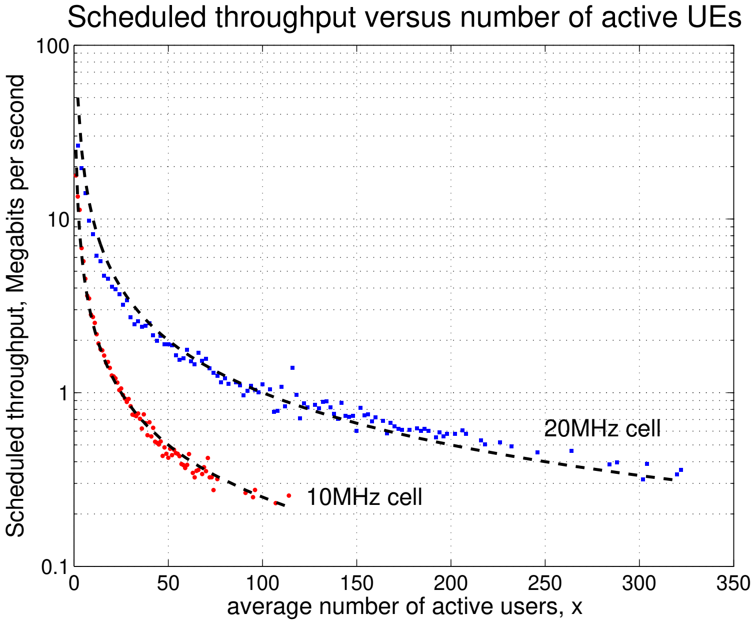

Figure 4: Scheduled throughput T_{sch} versus average number of active UEs, two examples from live network.

week.

Let ψ denote the correlation coefficient for a given cell, calculated upon the hourly scheduled throughput (T_{sch}) and PRB utilization (ρ) samples from the PMC. Let also ρ_{max} denote the maximum load observed for the cell. The values of ρ have been quantized to a resolution of 2 PRBs, and T_{sch} computed by aggregating the per-bin samples.

At first glance, 72.4% of the cells in this example network have ψ < -0.5, and could be therefore considered as having a linear decrease for T_{sch} as a function of ρ. This can be further sliced according to ρ_{max} since lowly-loaded cells often exhibit bursty traffic patterns that do not fit with the M/G/1/PS assumptions. Indeed, 90% of cells with ρ > 0.7 have ψ < -0.5. This is about 37% of the total cells.

The two-dimensional histogram of the per-cell tuples (ρ_{max}, ψ) is shown in fig. 6. We observe that most of the cells with good linearity (e.g., ψ < -0.5) also exhibit high maximum load (e.g., ρ_{max} > 0.7). In contrast, lowly-loaded cells (for example, ρ_{max} < 0.2) show a somewhat uniform looking distribution of ψ, suggesting that the model does not apply to them because one or more assumptions are violated.

Finally, we evaluate the mean absolute error when fitting T_{sch} and ρ to the linear model in eq. 8. For simplicity, the cell capacity C is estimated via simple least squares (LS) fit:

$$C_{LS} = \frac{\sum_i (1-\rho_i) T_{sch,i}}{\sum_i (1-\rho_i)^2} \tag{20}$$

The error associated to fitting the model for a given cell, on the other hand, is measured as an average of the absolute value of the relative error over the N_s throughput samples (after binning ρ

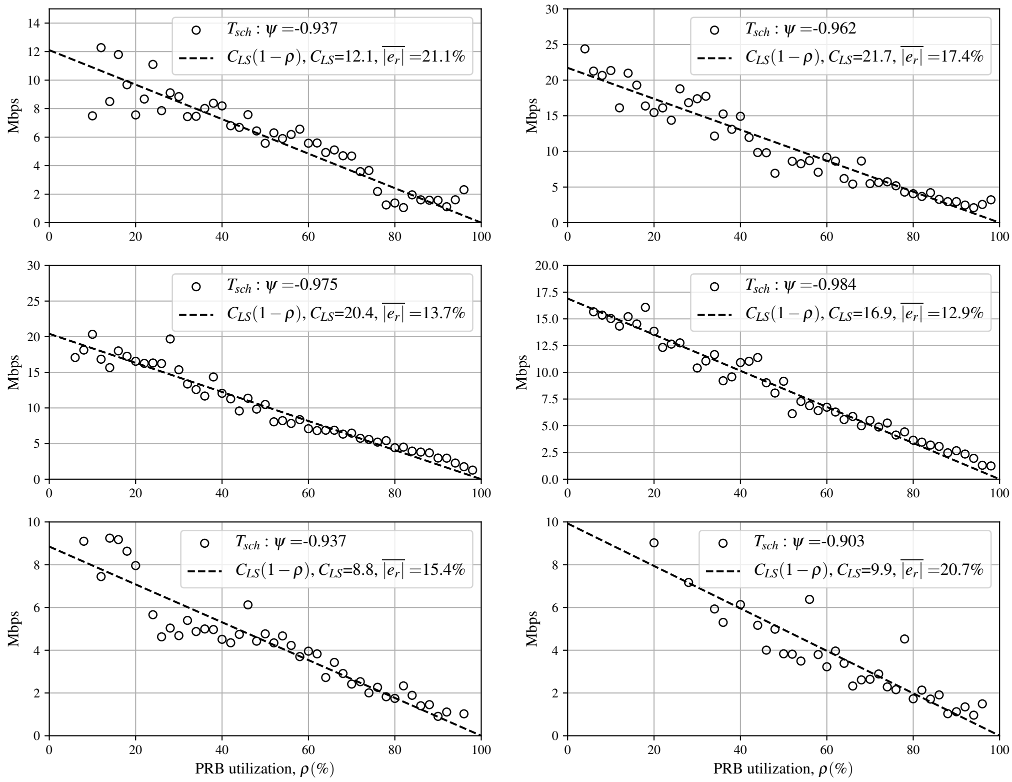

Figure 5: Scheduled throughput versus PRB utilization, six cells from three different networks. The correlation coefficient ψ and mean absolute error of the fit are given for each cell.

to 2 PRB resolution):

$$e_r = \frac{1}{N_s} \sum_i \frac{|\hat{T}_{sch,i} - T_{sch,i}|}{T_{sch,i}} \tag{21}$$

where $\hat{T}_{sch,i} = C_{LS}(1 - \rho_i)$ is the value calculated based on the linear model in (8).

Figure 7 shows that the average of e_r among cells for the high-load, high linearity network slice is between 5 and 27%.

7. Conclusion

This paper discussed the mapping of LTE user throughput to radio utilization. The average scheduled throughput, measured using the 3GPP method, was shown to be fairly accurately predicted by the cell capacity divided by the average number of active UEs. Adding the usual assumption that user data flow arrivals are Poisson distributed this result was expressed in terms of Physical Resource Block utilization, the outcome being the well-known M/G/1 Processor Sharing formula that predicts the linear decrease of user throughput with PRB utilization. An extension to external rate limitation was given, as well as an application to load balancing between frequency layers was discussed. Measurement data from real-world networks indicate that the scheduled throughput degrades about linearly with the serving cell PRB utilization, and therefore the linear degradation predicted by $C(1 - \rho)$ is a useful approximation for practical network operations purposes.

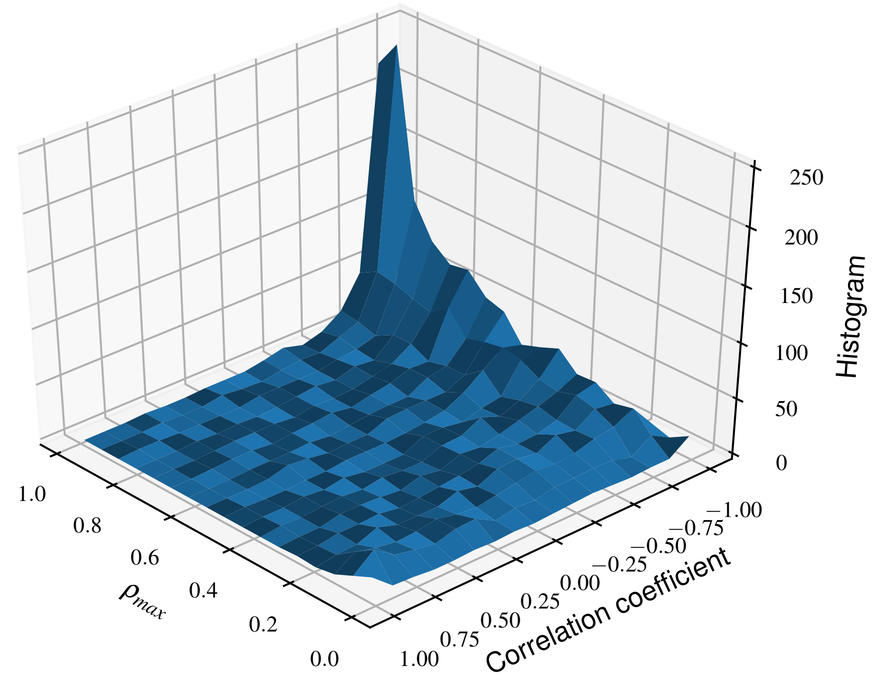

Figure 6: Joint distribution of (ρ_{max}, ψ) across cells in a live network.

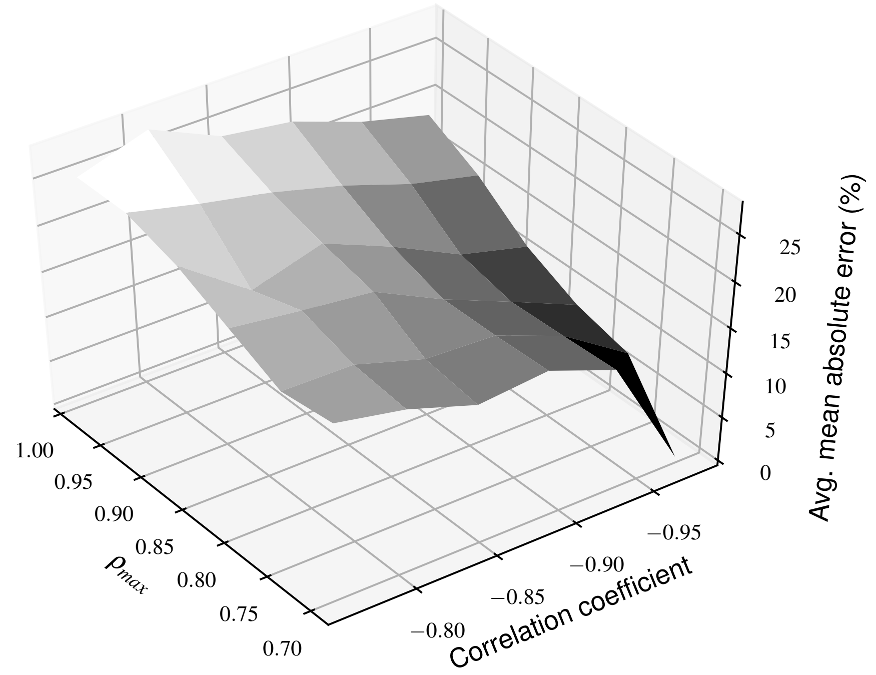

Figure 7: Average of e_r from eq. 21, in case of the linear model in eq. 8 and the LS fit for C from eq. 20.

References

[1] J. W. Roberts, "A survey on statistical bandwidth sharing," *Comput. Netw.*, vol. 45, no. 3, pp. 319–332, Jun. 2004. [Online]. Available: http://dx.doi.org/10.1016/j.comnet.2004.03.010

Cited on page(s): 33

[2] T. Bonald and A. Proutière, "Wireless downlink data channels: user performance and cell dimensioning," in *Proc. ACM MOBICOM*, 2003, pp. 339–352. [Online]. Available: http://doi.acm.org/10.1145/938985.939020 *Cited on page(s):* 33, 35, 37

[3] J. Melasniemi, P. Lassila, and S. Aalto, "Minimizing file transfer delays using SRPT in HSDPA with terminal constraints," in *4th Workshop on Network Control and Optimization, Ghent, Belgium*, 2010. *Cited on page(s):* 33

[4] G. Arvanitakis and F. Kaltenberger, "PHY and MAC layer modeling of LTE and WiFi RATs," Eurecom, Tech. Rep. EURECOM+4879, 03 2016. [Online]. Available: http://www.eurecom.fr/publication/4879 *Cited on page(s):* 33

[5] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in lte cellular networks: Key design issues and a survey." *IEEE Communications Surveys and Tutorials*, vol. 15, no. 2, pp. 678–700, 2013. *Cited on page(s):* 33

[6] X. Li, U. Toseef, T. Weerawardane, W. Bigos, D. Dulas, C. Görg, A. Timm-Giel, and A. Klug, "Dimensioning of the LTE S1 interface," in *Proc. IFIP WMNC*, 2010. *Cited on page(s):* 33

[7] H. Holma and A. Toskala, *LTE for UMTS - OFDMA and SC-FDMA Based Radio Access*. Wiley Publishing, 2009. *Cited on page(s):* 33

[8] F. Khan, *LTE for 4G Mobile Broadband: Air Interface Technologies and Performance*, 1st ed. New York, NY, USA: Cambridge University Press, 2009. *Cited on page(s):* 33

[9] S. Sesia, I. Toufik, and M. Baker, *LTE, The UMTS Long Term Evolution: From Theory to Practice*. Wiley Publishing, 2009. *Cited on page(s):* 33

[10] N. Chen and S. Jordan, "Throughput in processor-sharing queues," *IEEE Trans. Automat. Contr.*, vol. 52, pp. 299–305, 2007. *Cited on page(s):* 35, 36, 37

[11] A. A. Kherani and A. Kumar, "Stochastic models for throughput analysis of randomly arriving elastic flows in the Internet," in *Proc. IEEE INFOCOM*, 2002. *Cited on page(s):* 36, 37

[12] L. Kleinrock, *Queueing Systems*. Wiley Interscience, 1975, vol. 1,2. *Cited on page(s):* 37

[13] 3GPP, "Performance measurements Evolved Universal Terrestrial Radio Access Network (E-UTRAN)," 3rd Generation Partnership Project (3GPP), TS 32.425, 2016. *Cited on page(s):* 37, 41

[14] K. Lindberger, "Balancing quality of service, pricing and utilisation in multiservice networks with stream and elastic traffic," in *Proc. ITC 16*, 1999, pp. 1127–1136. *Cited on page(s):* 37

[15] V. Naumov and O. Martikainen, "Queueing systems with fractional number of servers," The Research Institute of the Finnish Economy, Discussion Papers 1268, 2012. [Online]. Available: https://EconPapers.repec.org/RePEc:rif:dpaper:1268 *Cited on page(s):* 37

[16] A. Pokhariyal, T. E. Kolding, and P. E. Mogensen, "Performance of downlink frequency domain packet scheduling for the UTRAN Long Term Evolution," in *Proc. IEEE PIMRC*, Helsinki, Sep. 2006. *Cited on page(s):* 41