Fundamental Tradeoffs among Reliability, Latency and Throughput in Cellular Networks

Alvarez, Beatriz Soret; Mogensen, Preben Elgaard; Pedersen, Klaus I.; Aguayo-Torres, Mari Carmen

Published in:
Proceedings of Globecom 2014

DOI (link to publication from Publisher):
10.1109/GLOCOMW.2014.7063628

Publication date:
2014

Document Version
Peer reviewed version

Fundamental Tradeoffs among Reliability, Latency and Throughput in Cellular Networks

Beatriz Soret*, Preben Mogensen*†, Klaus I. Pedersen*†, Mari Carmen Aguayo-Torres‡
*Nokia Networks, †Aalborg University, ‡Universidad de Málaga
beatriz.soret@nsn.com

***Abstract*—We address the fundamental tradeoffs among latency, reliability and throughput in a cellular network. The most important elements influencing the KPIs in a 4G network are identified, and the inter-relationships among them is discussed. We use the effective bandwidth and the effective capacity theory as analytical framework for calculating the maximum achievable rate for a given latency and reliability constraint. The analysis is conducted in a simplified LTE network, providing baseline - yet powerful - insight of the main tradeoffs. Guidelines to extend the theory to more complex systems are also presented, including a semi-analytical approach for cases with intractable channel and traffic models. We also discuss the use of system-level simulations to explore the limits of LTE networks. Based on our findings, we give some recommendations for the imminent 5G technology design phase, in which latency and reliability will be two of the principal KPIs.**

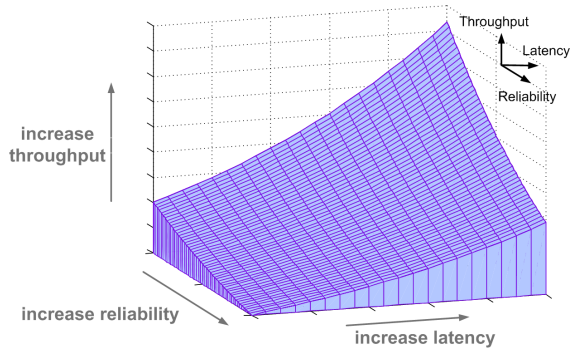

Fig. 1. Sketch of the tradeoffs among latency, reliability and throughput.

I. Introduction

The explosion of machine-to-machine (M2M) communications opens the possibility of implementing a myriad of applications requiring extremely low latency and ultra high reliability. LTE, the de facto standard for 4G cellular networks, is postulated as a candidate for the support of M2M [1]. One main concern for the use of the LTE network relates to its capability of meeting the stringent reliability and latency constraints without compromising the delivery of traditional applications.

In this paper we investigate three main Key Performance Indicators (KPIs) of LTE networks for M2M communications, namely latency, reliability and throughput. A sketch of the tradeoffs among the three KPIs is shown in Figure 1. In a wireless system it is challenging to fulfil simultaneously stringent reliability, latency and throughput requirements. LTE is to a large extent designed to maximize the system capacity for broadband data, where the most important KPI is the average user goodput. Average user goodput is typically maximized by aggressive use of retransmissions, as well as opportunistic radio channel aware scheduling techniques. The price to pay for this improvement in reliability and throughput is the degradation in terms of latency, with some packets with increased risk of experiencing potentially long delays. At the same time, retransmissions and other error control mechanisms implying some overhead have a cost in the form of increased latency. The question we want to answer is if the network is capable of delivering a payload of size A bits within B ms with a maximum latency of C ms and a reliability of D %, where typical values of ultra reliability are 99.999% or 99.99966% (a.k.a. six-sigma) [2]. We limit the scope of the paper to the downlink and PHY/MAC procedures, i.e. we do not include higher layer procedures such as Radio Link Control (RLC) and Transmission Control Protocol (TCP) retransmissions in the KPIs budget. The effect of higher layers can be added on top of it to get the end-to-end performance figures [3].

There are three main approaches to address the problem: analytical models, semi-analytical models and simulations. Analytical results provide us with valuable insight of the tradeoffs among parameters. However, current cellular systems are very complex with many different elements contributing to the KPIs under investigation. In this sense, striving to model until the very last detail often leads to mathematically intractable problems. The alternative is to enable assumptions, limiting the scope of the results if these assumptions throw away important aspects of realism. Then, it is advisable to conduct statistically reliable simulations, which complement the analtyical study and provide figures of the KPIs of interest. In between, semi-analytical models represent a middle ground making partial use of simulation results as an input to the analytical expressions.

As for analytical study, we use the effective bandwidth and effective capacity theories. The effective bandwidth function of a given time-varying source process is defined as the minimum service rate necessary to deliver the data by fulfilling certain latency requirements, expressed in the form of a statistical delay constraint [4]. Analogously, the effective capacity [5] is defined as the maximum constant arrival data rate that a given time-varying service process can support while meeting the delay constraint. Both concepts can be jointly used to analyze a wireless system with random traffic and channel fluctuations

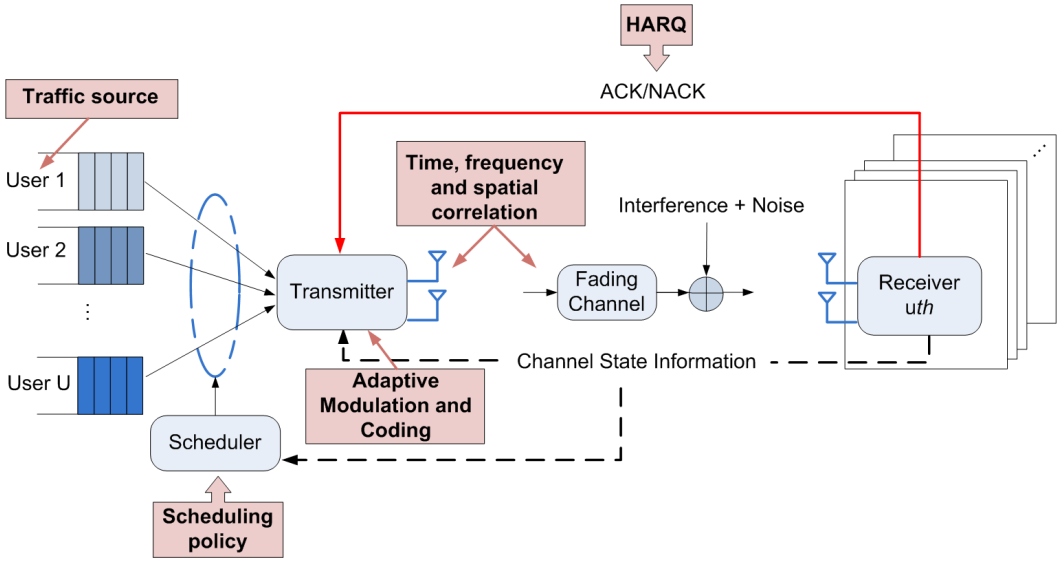

Fig. 2. LTE system model.

[6]. Since their introduction, the effective bandwidth and effective capacity have been attracting much interest, and several studies have shown that the models are capable of estimating the QoS metrics in different scenarios, see e.g. [6]- [11].

The remainder of the paper is organized as follows. We first give an overview of the PHY/MAC procedures impacting the latency and reliability performance in LTE. In Section III we propose the joint use of the effective bandwidth and the effective capacity theory as analytical or semi-analytical framework to investigate the tradeoffs among the KPIs. Section IV discusses the use of system level simulations and several sources of imperfections in the system. Based on our findings, we give some recommendations for the design of 5G networks in Section V. Conclusions in Section VI close the paper.

II. System Aspects

A sketch of the LTE system model is shown in Figure 2, which we use to discuss the various sources contributing to the KPIs of interst in a 4G network [12]. Data from different users arrive from a higher layer application and are stored in user's transmission buffers. A very simple and commonly used model for the variable traffic assumes a finite payload per user and user arrivals according to a Poisson process. When a user finishes transmitting the payload, the call is terminated. More realistic self-similar traffic models for a data network assume bursts of packets alternating with silences, and some correlation between packets within a burst. In general, correlation and burstiness in the traffic source are harmful for the latency performance [13].

Periodically, the scheduler allocates the channel resources to users based on the buffer occupancy, the Channel State Information (CSI) reported by the users, the pending retrasnmissions, and other factors. LTE supports both QoS- and radio channel aware scheduling disciplines [14]. The QoS-aware mechanisms can be used to prioritize the scheduling of time-critical messages, with the aim of reducing the transmitter queuing delay of such messages. However, avoiding queueing delays cannot be guaranteed at high offered loads approaching the cells capacity limit (or even exceeding it).

Adaptive Modulation and Coding (AMC) (a.k.a. link adaptation) techniques are applied at the transmitter, such that the constellation size and the coding rate are dynamically modified to exploit the channel diversity and with a reliability target, in the form of BLock Error Rate (BLER) [12]. Thus, more robust modulation and more aggressive coding are applied when the reliability is jeopardized. Naturally, reducing the BLER has a cost in terms of reduced spectral efficiency.

LTE exploits time, frequency and spatial diversity to maximize the spectral efficiency. At the same time, the system suffers from time, frequency and spatial correlation, and the three ingredients degrade the latency performance. The users granted access to the wireless channel are allocated transmission resources on a subframe (1 ms) and physical resource block (PRB) resolution, where one PRB consist of 12 subcarriers, corresponding to a bandwidth chunk of 180 kHz. The wireless channel is time-variant, and can be represented by a stochastic model capturing certain time-frequency correlation properties. Uncertainties related to the variability of the radio channel are reduced by using closed loop codebook MIMO techniques (e.g. spatial diversity). Todays LTE implementations use mainly 2x2 MIMO (i.e. 4th order diversity), although the standard in principle supports up to 8x8 MIMO. Furthermore, the wideband characteristics of LTE (up to 20 MHz per carrier and 100 MHz with carrier aggregation) also offer frequency-diversity that helps in reducing the variability of the effective radio channel.

At the receiver, the desired signal for the UE is subject to both additive thermal noise as well as time-variant and frequency selective interference. Among others, the experi-

enced interference depends on the scheduling activity of the neighboring cells (i.e. other-cell load dependent) as well as the location of the user. Use of advanced receiver algorithms with interference mitigation capabilities improves the post detection SINR, reducing the sensitivity of the performance to the randomness in the interference. Today, LTE mainly relies on linear minimum mean square error (MMSE) receivers with interference rejection combining (IRC), while standardization of more advanced network assisted receivers with non-linear interference cancellation capabilities is ongoing [15]. Additionally, some techniques supported in LTE for inter-cell interference coordination (ICIC) [16] and coordinated multi-point (CoMP) [17] have a tendency to increase the interference variability, which challenges the traditional link adaptation framework.

HARQ is an error correction mechanism based on retransmissions. The receiver produces either an ACK, for the case of error-free transmission, or a NACK if some errors are detected. Upon reception of a NACK message, the desired packet will be sent again. LTE supports hybrid automatic repeat request (HARQ) with soft combining. The delay between two HARQ transmissions on the same stop-and-wait channel is 8 ms. This 8 ms delay is mainly a result of having 1 ms subframe and certain terminal processing requirements for LTE [12].

III. Analytical Framework

A. Setting the Scene

In the analytical framework, two random processes model the bursty traffic source (source process) and the fading channel with the associated link adaptation procedures (channel process). For simplicity, we assume only one user, but the theory can be generalized for multiuser systems [7]. The physical time divided into units referred to as symbol periods and represented by the transmission discrete time unit, n. On the one hand, the traffic source has instantaneous source rate $a[n]$, meaning that each symbol the source generates $a[n]$ bits, that are stored in the user queue. On the other hand, each symbol, $c[n]$ bits are removed from the queue and transmitted to the air, being $c[n]$ the instantaneous channel rate. No packages are considered in this fluid model.

Transmission over certain channels (e.g. Rayleigh channels) cannot accomplish any deterministic delay requirement. In this case, it is more convenient to express the latency requirement in terms of a probabilistic delay constraint (D^t, ϵ), where the target delay is D^t, and the probability of exceeding D^t is denoted by ϵ.

As for reliability, $c[n]$ is capturing the adaptive transmission rate and all the PHY/MAC procedures related to minimize the probability of loosing packages, such as AMC and HARQ. Thus, a more dense constellation and higher coding rate are selected when the channel condition of the user is good, increasing $c[n]$. Upon arrival of a NACK package, the corresponding retransmitted packages have priority, lowering $c[n]$.

B. The effective bandwidth and effective capacity functions

The effective bandwidth of a source expresses the minimum constant service rate required by a given arrival process to guarantee a probabilistic delay constraint [4]. Mathematically,

$$E_A(\nu) = \lim_{n\to\infty} \frac{1}{n\cdot\nu} \log \mathrm{E}\left[e^{\nu A[n]}\right] \quad \forall \nu \geq 0\,, \tag{1}$$

with $A[n]$ being the accumulated source rate, i.e. the amount of bits generated by the source from 0 to $n-1$, $A[n] = \sum_{m=0}^{n-1} a[n]$; and $\mathrm{E}[\cdot]$ is the expectation.

Dual to the effective bandwidth, the effective capacity of a channel process expresses the maximum arrival rate that the channel can support by fulfilling the delay constraint [5]. Mathematically,

$$E_C(\nu) = \lim_{n\to\infty} \frac{1}{n\cdot\nu} \log \mathrm{E}\left[e^{\nu C[n]}\right] \quad \forall \nu \geq 0\,, \tag{2}$$

where $C[n] = \sum_{i=0}^{n} c[i]$ is the accumulated transmission rate.

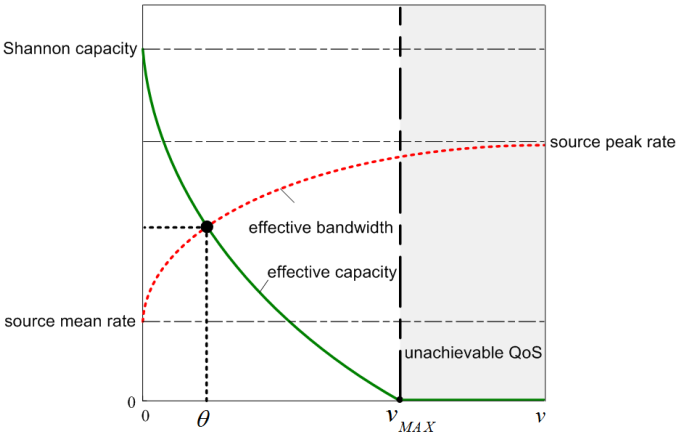

Fig. 3. Effective bandwidth and effective capacity function.

C. The intersection of the two curves

The effective bandwidth and effective capacity curves are depicted in Figure 3. In both cases, a high value of the parameter ν indicates a more severe delay requirement – lower D^t or ϵ –, and a small value of ν symbolizes loose delay requirements. Therefore, the effective bandwidth curve of a traffic source increases with ν, starting always at the source mean rate and tending towards the peak rate of the source as $\nu \to \infty$.

On the other hand, the effective capacity of the channel starts at Shannon's capacity when $\nu = 0$, with no delay constraints imposed, and decreases asymptotically with ν. In the case of Rayleigh channels, the curve reaches zero at a certain point denoted in the figure as ν_{MAX}. Higher values of ν imply QoS requirements that are not achievable by the channel, regardless of the traffic source. If we combine both curves, a working point of the system can be defined corresponding to the intersection of the two curves. This point is called the QoS exponent θ.

The QoS exponent θ captures the statistical delay guarantees (D^t, ϵ), and it symbolizes the point in which both the traffic source and the channel are able to meet the delay requirement.

The connection between θ and the explicit latency requirements is given by

$$\epsilon = Pr(D \geq D^t) \asymp e^{-\theta \cdot E_A(\theta) \cdot D^t} \qquad D^t \to \infty, \tag{3}$$

The above equation states that the probability of the delay-bound decays exponentially as the target D^t increases. See [5], [6] for additional details.

D. Computation of the two functions

The computation of equations (1) and (2) is in general difficult. For the effective bandwidth function, several traffic models have been investigated in the literature; see e.g. [18] where, among others, periodic, Gaussian and ON/OFF processes are treated, as well as multiplexing of several sources. For mathematically intractable sources, there are also methods to estimate the effective bandwidth, see e.g. the Dembo estimator in [19] or the simulation-based approach in [20].

As for the effective capacity, let us assume a single channel and single antenna system, i.e. only time-correlation is present. In this case, the accumulated transmission rate can be divided into blocks of k symbols, assuming intra-block correlation but not inter-block correlation,

$$C[n] = \sum_{i=0}^{\frac{n}{k}-1} C_i[k] \ , \tag{4}$$

with $C_i[k] = \sum_{m=0}^{k-1} c[k \cdot i + m]$.

To neglect the inter-block correlation, the choice of k has to be closely related to the correlation of the channel. If the channel is strongly correlated, longer blocks have to be defined in order to assume independent blocks. If the size of the blocks k is large enough, then $C[n]$ is the sum of a sufficiently large number of independent random variables, and the Central Limit Theorem applies. The effective capacity function of the channel is then written [6] [18]

$$E_C(\nu) = \frac{\mu_k}{k} - \frac{\nu}{2}\frac{\sigma_k^2}{k}, \tag{5}$$

where μ_k and σ_k^2 are the mean and the variance of the transmission rate in a block,

$$\begin{aligned} \mu_k &= \mathrm{E}\left[C_i[k]\right] \\ \sigma_k^2 &= \mathrm{E}\left[C_i^2[k]\right] - \mu_k^2 \ . \end{aligned} \tag{6}$$

With C[n] a stationary and ergodic process, the mean of each block equals the mean value of the process without correlation, and can be easily obtained for classical channel models like Rayleigh or Nakagami. The evaluation of the variance of the blocks is more challenging as it results from a multivariate distribution. See [6] for more details.

E. Generalization and semi-analytical approach

For the sake of simplicity, the discussion in the previous subsection does not consider frequency or spatial correlation, but several studies in the literature have addressed these aspects (see e.g. [8] and [9]). Besides, works including different PHY/MAC elements relevant for the transmission rate are also available, like encoding/decoding in [10] or the HARQ process and its associated contribution to the delay in [11]. The challenge of integrating all these elements into a unified analytical model, however, is overwhelming. Instead, a semi-analytical approach for the channel process makes use of empirical statistics extracted from simulations. The sampled values are plugged into (5), yielding

$$E_C(\nu) = \frac{\hat{\mu}_k}{k} - \frac{\nu}{2}\frac{\hat{\sigma}_k^2}{k}, \tag{7}$$

where $\hat{\mu}_k$ and $\hat{\sigma}_k^2$ are the sampled mean and variance. To obtain them, the accumulated transmission rate is splitted into blocks of length k and the statistics are measured over a long realization of the channel process via simulations.

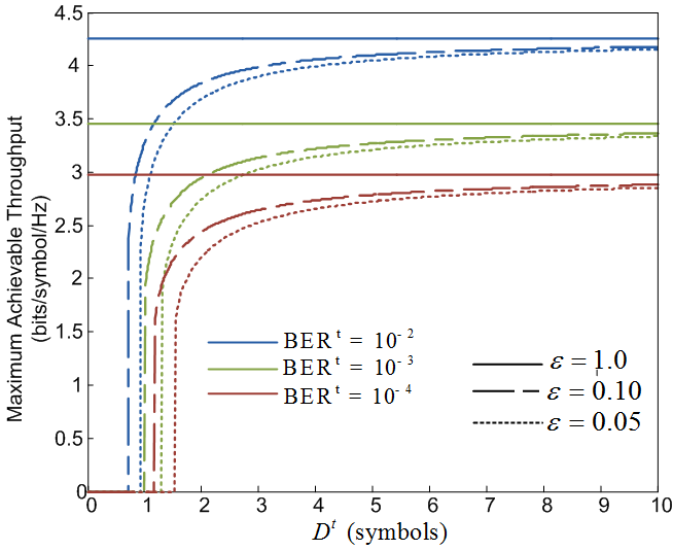

Fig. 4. Numerical example of the tradeoff among throughput, latency and reliability in a wireless system.

F. Numerical Example

An example of the tradeoff between throughput, latency and reliability by means of the effective bandwidth and effective capacity framework is illustrated in Figure 4. The traffic source is constant (which implies constant effective bandwidth), and the channel is an uncorrelated Rayleigh process with single carrier and single transmit/receive antenna. We assume adaptive modulation with QAM schemes up to 64QAM and a given Bit Error Rate (BER) requirement[1]. It means that the throughput is maximized while maintaining the target BER by selecting the proper QAM modulation for a given instantaneous channel quality. Neither encoding nor HARQ are used. The maximum achievable rate is plotted as a function of the target delay (latency requirement) in the constraint. Different lines represent different values of ϵ, ranging from 1.0

[1] In accordance with the adopted fluid model, we talk here about BER and not BLER.

TABLE I. SUMMARY OF TYPICAL IMPERFECTIONS FOR SPECTRAL EFFICIENCY OPTIMIZED DOWNLINK TRANSMISSION.

Component	Remarks	Imperfections
Channel State Information (CSI)	UE measures the experienced channel quality. Used by the eNB for link adaptation and scheduling.	UE measurement imperfections, often modeled with zero mean Gaussian error of 1dB standard deviation in the SINR dB domain. Uplink decoding error of CSI on the order of less than 1%
Physical Downlink Control Channel (PDCCH)	Sending scheduling grant to the UE	BLER of 1%-2%
Physical Downlink Data Channel (PDSCH)	Data transmission to the UE	1st transmission BLER: 10%-30% 2nd transmission BLER: 1%-5% 3rd transmission BLER: 0.1%
MAC-layer ACK/NACK	UE sends ACK/NACKS to the eNB corresponding to the Hybrid ARQ transmissions	P(ACK|NACK) = 0.01% P(NACK|ACK) = 1% P(ACK|DTX) = 1%
RLC retransmissions	They can occur if using RLC acknowledged mode	Typically less than 1% RLC retransmissions

(i.e. no delay requirements) to 0.05. The reliability, captured in the BER requirement, ranges from 10^{-2} to 10^{-4}.

For stringent latency requirements, the maximum achievable throughput approaches zero, being limited by system factors such as e.g. the subframe length. As the latency constraint is relaxed (higher D^t or higher ϵ), the achievable rate increases, in a region that is dominated by the tradeoff between load and latency. It is noticeable that with $\epsilon = 1$ the rate is insensitive to the other delay parameter, D^t, and the achievable rate reaches the Shannon capacity of the channel, which naturally is the upper bound for the considered channel model. It is also observed how the throughput increases as the reliability constraint is relaxed, due to the increase of modulation level in the adaptive modulation scheme. The behaviour is the same if coding is added, with a low value of error rate implying lower coding rate and consequently lower data rate.

IV. DISCUSSION

The analytical framework presented in Section III gives a good insight of the main mechanisms impacting the KPIs of interest. One observation is that in a multiuser environment the three KPIs can be guaranteed only for a fraction of the load in the system, and at the expenses of larger latencies for the rest of best effort users in the network. Moreover, the main procedures relevant for the study can be incorporated into the system model to get a good approximation to the final values. However, simplifications and idealizations limit the scope of purely analytical studies when dealing with very complex systems, as it is the case of LTE. In this sense, system-level simulations can model not only all the relevant elements but also different sources of imperfection.

Besides the random elements described in Section II, several sources of imperfection at the PHY/MAC level are also important for the latency and reliability performance, as summarized in Table I. For example, AMC makes use of the received CSI feedback, which is subject to various imperfections such as measurement imperfections, quantization, reporting delay, and reception errors [21]. All of them can be represented by a random process, which essentially means that there is certain probability that the AMC selected by the base station deviates from the ideally desired selection. In addition, the HARQ performance is influenced by the randomness associated with occasionally rare mis-detection of ACK/NACK at the base station from the terminals. The composite variability of all of the sources of imperfection identified in the table can indeed lead to long tails of the transmission delay, although at low probabilities.

V. OUTLOOK TOWARDS 5G

The next generation of mobile radio access technologies (5G), expected to become available for commercial launch around 2020, is right now in its early exploration phase with several unknowns regarding the requirements and the technologies to be used. We have indicated in Table II the various sources of variability with impact on the link performance in 4G LTE (see also Section II) and how to further improve them for 5G. Hence, pointing to the candidate techniques that could help in enabling the support of ultra low latency and high reliable communications in the future [22].

In the frequency and spatial domain, the increased diversity given by larger bandwidths and higher order MIMO (massive MIMO) will help in improving the tradeoff between latency and reliability. As for the interference, the goal is to achieve a more stable interference footprint. The use of advanced interference management comprising both network-based coordination and receiver-based techniques is expected to be an integral part of 5G, providing both general capacity benefits and improved reliability by reducing the interference vulnerability [23]. A promising enhancement in the network side is the use of multi-cell baseband pooling as suggested in [24], which offers opportunities for centralized multi-cell scheduling, reducing some of the uncertainties that would otherwise be present if conducting independent scheduling and resource allocation per cell. Further enhancements aiming at reducing the latency include the use of a shorter subframe duration for a reduced HARQ transmission delay.

As indicated in Table II, the end-to-end performance is also impacted by the core network. Architecture enhancements to reduce the impact from the core network are also an active research topic; e.g. by allowing local data connectivity for base stations to reduce end-to-end delays. One example is allowing the data path of communication between two terminals without the core network participating, as follows: From terminal A to base station A, directly to base station B that is serving terminal B, and to terminal B. Similarly, native support for direct device-to-device (D2D) communication is anticipated to be an integrated part of future 5G systems [25].

VI. CONCLUSIONS

In a wireless system it is challenging to fulfil simultaneously stringent reliability, latency and throughput. In this paper we

TABLE II. SUMMARY OF THE SOURCES OF VARIABILITY INFLUENCING ON MEETING THE REQUIREMENTS FOR ULTRA RELIABILITY, INCLUDING ASSOCIATED TECHNIQUES LTE AND OUTLOOK ON CANDIDATE TECHNIQUES FOR FUTURE 5G.

Sources of variabiality	4G LTE	Candidate techniques for 5G
Base station transmitter queueing delays	QoS-aware scheduling to prioritize transmission of critical messages	QoS-aware scheduling to prioritize transmission of critical messages. Inter-cell load balancing or distributed queues
Radio channel variability	Closed loop code-book MIMO, wideband transmission	Massive MIMO, bandwidth beyond current LTE
Interference variability	Advaced receivers such as MMSE-IRC, simple ICIC techniques, Network Assisted Interference Cancellation and Suppression (NAICS)	Optimized system designed for advanced interference mitigation receivers, advanced ICIC techniques offering reduced and stable interference footprint
Channel state information	Comprises CQI, RI, PMI, and also multi-process CSI as defined for CoMP	More accurate CSI to better guide the base station to select the optimum transmission
Link adaptation	Relies on CSI, first transmission target BLER is typically 10%-30% although there are options for also using lower BLER for time-critical transmissions	Based on more accurate CSI, increased flexibility for differentiated BLER depending on reliability requirements
HARQ delays	8 ms delay between HARQ transmissions on the same SAW channel	Reduced HARQ transmission delay; e.g. by using shorter TTI size and stricter processing requirements
Core network (CN) delays	Data coming from the serving gateway via the S1 interface to the eNB, simple D2D solution under development for LTE	Localized routing of data with reduced CN delay, optimized D2D for applications where this is feasible
Handover	Asynchronous HO based on RACH access by the UE to access the target eNB	Synchronous HO (e.g. with synchronized base stations) removes the need for RACH access, increasing reliability and latency

have studied the fundamental tradeoffs among these three KPIs in a 4G network. We have first described the main sources of variability having a relevant impact on them. An analytical framework to calculate the tradeoffs among the three KPIs is outlined. Different aspects of the PHY/MAC procedures can be studied by means of joint use of the effective bandwidth and the effective capacity theory. Guidelines to extend the theory to cases in which the channel and / or the traffic model are not tractable from a mathematical point of view are also provided, as well as the advisability of system-level simulations in such a complex system with several random elements and related sources of non-idealities. Finally, we identified different candidate techniques to improve the trade-off among the KPIs in future 5G systems.

REFERENCES

[1] G. Wu et al., "M2M: From Mobile to Embedded Internet," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 36-43, April 2011.

[2] A. Osseiran et al., "Scenarios for 5G Mobile and Wireless Communications: The Vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26-35, May 2014.

[3] G. Gomez, et al., "QoS Modeling for End-to-End Performance Evaluation over Networks with Wireless Access," *EURASIP Journal on Wireless Communications and Networking*, 2010.

[4] C. S. Chang, J. A. Thomas, "Effective Bandwidth in High-Speed Digital Networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, Aug. 1995.

[5] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Transactions on Wireless Communications*, vol., 2, no. 4, pp. 630-643, July 2003.

[6] B. Soret, M.C. Aguayo-Torres, J.T. Emtrambasaguas, "Capacity with explicit delay guarantees for generic sources over correlated Rayleigh channel," *IEEE Transactions on Wireless Communications*, vol. 9, no. 6, pp. 1901-1911, 2010.

[7] B. Soret, M. C. Aguayo-Torres, and J. T. Entrambasaguas, "Multiuser capacity for heterogeneous QoS constraints in uncorrelated Rayleigh channels,", *Proc. Information Theory Workshop ITW'09*, October 2009.

[8] B. Soret, M.C. Aguayo-Torres, J.T. Emtrambasaguas, "Analysis of Delay Constrained communications over OFDM systems," *Proc. IEEE Global Communication Conference*, November 2009.

[9] M. C. Gursoy, "MIMO Wireless Communications Under Statistical Queueing Constraints," *IEEE Transactions on Information Theory*, vol. 57, no. 9, pp. 5897-5917, Sep. 2011.

[10] L. Liu, et al., "Resource Allocation and Quality of Service Evaluation for Wireless Communication Systems Using Fluid Models," *IEEE Transactions on Information Theory*, vol. 53, no. 5, pp. 1767-1777, May 2007.

[11] Y. Li, M. Cenk Gursoy, and S. Velipasalar, "On the Throughput of Hybrid-ARQ under QoS Constraint," *arXiv:1312.0882v1 [cs.IT]*, December 2013.

[12] S. Sesia, I. Toufik, et M. Baker (Ed), "LTE-The UMTS Long Term Evolution: From Theory to Practice,", *John Wiley & Sons Ltd*, 2011.

[13] K. Park, and W. Willinger (Ed), "Self-Similar Network Traffic and Performance Evaluation," *John Wiley & Sons Ltd*, 2000.

[14] K. I. Pedersen et al., "An Overview of Downlink Radio Resource Management for UTRAN Long-Term Evolution," *IEEE Commun. Mag.*, vol. 47, no. 7, pp. 86-93, July 2009.

[15] 3GPP, TR 36.866 "Study on Network-Assisted Interference Cancellation and Suppression (NAICS) for LTE," v. 12.0.0, March 2014.

[16] K. I. Pedersen et al., "Enhanced inter-cell interference coordination in co-channel multilayer LTE Advanced networks," *IEEE Wireless Communications Magazine*, vol. 20, no. 3, pp. 120-127, June 2013.

[17] J. Lee et al., "Coordinated multipoint transmission and reception in LTE-Advanced systems," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 44-50, November 2012.

[18] F. P. Kelly, "Notes on the effective bandwidth, " *F. P. Kelly, S. Zachary, and I. Zeidins (editors) Stochastic Networks: Theory and Applications*, vol. 4, pp. 141-168, 1996.

[19] N. G. Dufield et al., "Entropy of ATM traffic streams: a tool for estimating QoS parameteres," *IEEE J. Sel. Areas Commun.*, vol. 13, pp. 981-990, August 1995.

[20] N. X. Liu, and J. S. Baras, "Measurement and Simulation Based Effective Bandwidth Estimation," *Proc. IEEE Global Communication Conference*, November 2004.

[21] K. I. Pedersen et al., "Performance Analysis of Simple Channel Feedback Schemes for a Practical OFDMA System, " *IEEE Transactions on Vehicular Technology*, vol 58, no. 9, pp. 5309-5314, Nov. 2009.

[22] P. Mogensen, et.al., "5G Small Cell Optimized Radio Design", *Proc. IEEE Global Communication Conference*, December 2013.

[23] W. Nam et al., "Advanced Interference Management for 5G Cellular Networks," *IEEE Commun. Mag.*, pp. 52-60, Vol 52, No. 5, May 2014.

[24] C. Lin et al., "Towards Green and Soft: A 5G Perspective", *IEEE Commun. Mag.*, pp. 66-73, February 2014.

[25] M. Nader Tehrani et al., "Device-to-Device Communication in 5G Cellular Networks: Challenges, Solutions, and Future Directions," *IEEE Commun. Mag.*, pp. 86-92, Vol 52, No. 5, May 2014.