NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCES
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

PROGRAM OF POSTGRADUATE STUDIES

PhD THESIS

Quality of experience characterization and provisioning in mobile cellular networks

Eirini V. Liotou

ATHENS

NOVEMBER 2017

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Χαρακτηρισμός και παροχή ποιότητας εμπειρίας σε κινητά κυψελωτά δίκτυα

Ειρήνη Β. Λιώτου

ΑΘΗΝΑ

ΝΟΕΜΒΡΙΟΣ 2017

PhD THESIS

Quality of experience characterization and provisioning in mobile cellular networks

Eirini V. Liotou

SUPERVISOR: Lazaros Merakos, Professor NKUA

THREE-MEMBER ADVISORY COMMITTEE:
Lazaros Merakos, Professor NKUA
Efstathios Hadjiefthymiades, Associate Professor NKUA
Athanasia Alonistioti, Assistant Professor NKUA

SEVEN-MEMBER EXAMINATION COMMITTEE

Lazaros Merakos,
Professor NKUA

Efstathios Hadjiefthymiades,
Associate Professor NKUA

Athanasia Alonistioti,
Assistant Professor NKUA

Georgios Polyzos,
Professor AUEB

Dimitrios Varoutas,
Associate Professor NKUA

Christos Xenakis,
Associate Professor UNIPI

Alexandros Kaloxylos,
Assistant Professor University of Peloponnese

Examination Date 16/11/2017

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Χαρακτηρισμός και παροχή ποιότητας εμπειρίας σε κινητά κυψελωτά δίκτυα

Ειρήνη Β. Λιώτου

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: Λάζαρος Μεράκος, Καθηγητής ΕΚΠΑ

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ ΠΑΡΑΚΟΛΟΥΘΗΣΗΣ:
Λάζαρος Μεράκος, Καθηγητής ΕΚΠΑ
Ευστάθιος Χατζηευθυμιάδης, Αναπληρωτής Καθηγητής ΕΚΠΑ
Αθανασία Αλωνιστιώτη, Επίκουρη Καθηγήτρια ΕΚΠΑ

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Λάζαρος Μεράκος,
Καθηγητής ΕΚΠΑ

Ευστάθιος Χατζηευθυμιάδης,
Αναπληρωτής Καθηγητής ΕΚΠΑ

Αθανασία Αλωνιστιώτη,
Επίκουρη Καθηγήτρια ΕΚΠΑ

Γεώργιος Πολύζος,
Καθηγητής ΟΠΑ

Δημήτριος Βαρουτάς,
Αναπληρωτής Καθηγητής ΕΚΠΑ

Χρήστος Ξενάκης,
Αναπληρωτής Καθηγητής ΠΑΠΕΙ

Αλέξανδρος Καλόξυλος,
Επίκουρος Καθηγητής Πανεπιστημίου
Πελοποννήσου

Ημερομηνία εξέτασης 16/11/2017

ABSTRACT

Traditionally, previous generations of mobile cellular networks have been designed with Quality of Service (QoS) criteria in mind, so that they manage to meet specific service requirements. Quality of Experience (QoE) has, however, recently emerged as a concept, disrupting the design of future network generations by giving clear emphasis on the actually achieved user experience. The emergence of the QoE concept has been a result of the inevitable strong transition that the Telecom industry is currently experiencing from system-centric networks to more user-centric solutions and objectives. Mobile network operators, service providers, application developers, as well as other stakeholders involved in the service provisioning chain have been attracted by the opportunities that the integration of the QoE concept could bring to their business; indeed, the provisioned QoE constitutes a determining factor of differentiation among different stakeholders, a tendency which is expected to become even more intense in the years to come.

Motivated by this boost towards user-centricity, the objective of the research conducted in this thesis is to explore the challenges and opportunities that arise in modern mobile cellular networks when QoE is considered. Such opportunities concern, first of all, the possibility to comprehend the QoE that a provider achieves when provisioning a service. This can be enabled by the implementation and integration of QoE assessment methods into the real-time operation of a network. Then, the next step is the exploitation of collected QoE-related intelligence in order to re-examine existing network-layer mechanisms (e.g., radio scheduling), or application-layer mechanisms (e.g., video streaming), as well as propose novel cross-layer approaches towards ameliorating the achieved QoE. Moreover, the opportunity emerges to propose novel algorithms that stem from the inherent idiosyncrasies of QoE, such as the non-linear impact of QoS-related parameters on QoE, as a way to further enhance the users' QoE. In this direction, throughout this thesis, QoE estimation models and metrics are explored and exploited in order to quantify QoE and thus, to improve existing mechanisms of mobile cellular networks.

The core of this thesis is the proposal of a QoE provisioning cycle that allows the control, monitoring (i.e., modeling) and management of QoE in a cellular network. Each one of these functions is further analyzed, while emphasis is given on the modeling and management operations. In terms of modeling, QoE assessment methods and QoE-related performance indicators are described and classified. Parametric quality estimation is identified as the most appealing type of QoE estimation in mobile cellular networks, thus, it is thoroughly described for widely used types of services, such as Voice over IP (VoIP) and video streaming.

In terms of QoE management, novel QoE-aware mechanisms that demonstrate QoE improvements for the users are proposed, namely: a) a QoE-driven Device-to-Device (D2D) communication management scheme that enhances end-user QoE, b) a “consistent” radio scheduling algorithm that improves the end-user QoE by mitigating throughput fluctuations, and c) a context-aware HTTP Adaptive Streaming (HAS) mechanism that successfully mitigates stallings (i.e., video freezing events) in the context of bandwidth-challenging scenarios. Moreover, a programmable QoE-SDN APP into the Software-Defined Networking (SDN) architecture is introduced, which enables network feedback exposure from mobile network operators to video service providers, revealing QoE benefits for the customers of video providers and bandwidth savings for the network operators.

Overall, this thesis promotes the uniting of the domain of QoE with the domain of mobile communications, as well as the collaboration of mutual-interest between mobile network operators (network layer) and service providers (application layer), presenting the high potential from such approaches for all involved stakeholders.

SUBJECT AREA: Mobile Communication Networks

KEYWORDS: Quality of Experience (QoE), HTTP Adaptive Streaming (HAS), radio resource scheduling, Software-Defined Networking (SDN), mobile cellular networks

ΠΕΡΙΛΗΨΗ

Παραδοσιακά, οι προηγούμενες γενεές κινητών κυψελωτών δικτύων έχουν σχεδιαστεί με κριτήρια Ποιότητας Υπηρεσίας, έτσι ώστε να πληρούν συγκεκριμένες απαιτήσεις διαφόρων υπηρεσιών. Η «Ποιότητα Εμπειρίας» έχει, ωστόσο, πρόσφατα εμφανιστεί ως έννοια, επηρεάζοντας το σχεδιασμό των μελλοντικών γενεών των δικτύων, δίνοντας σαφή έμφαση στην πραγματικά επιτευχθείσα εμπειρία του τελικού χρήστη. Η εμφάνιση της έννοιας της Ποιότητας Εμπειρίας οφείλεται στην αναπόφευκτη, ισχυρή μετάβαση που βιώνει η βιομηχανία των Τηλεπικοινωνιών από συστημο-κεντρικά δίκτυα σε πιο χρηστο-κεντρικές λύσεις και στόχους. Οι πάροχοι κινητών δικτύων, οι πάροχοι υπηρεσιών, οι προγραμματιστές εφαρμογών, αλλά και άλλα ενδιαφερόμενα μέλη που εμπλέκονται στην αλυσίδα παροχής υπηρεσιών προσελκύονται από τις ευκαιρίες που μπορεί να προσφέρει η ενσωμάτωση γνώσης Ποιότητας Εμπειρίας στο επιχειρηματικό τους μοντέλο. Πράγματι, η παρεχόμενη Ποιότητα Εμπειρίας αποτελεί έναν καθοριστικό παράγοντα διαφοροποίησης μεταξύ των διαφόρων παικτών, μία τάση που αναμένεται να γίνει ακόμη πιο έντονη τα επόμενα χρόνια.

Υποκινούμενη από αυτή την χρηστο-κεντρική τάση, η έρευνα που διεξάγεται σε αυτή τη διατριβή έχει ως στόχο την διερεύνηση των προκλήσεων και των ευκαιριών που προκύπτουν στα σύγχρονα κινητά κυψελωτά δίκτυα όταν λαμβάνεται υπόψιν η έννοια της Ποιότητας Εμπειρίας. Τέτοιες ευκαιρίες αφορούν, καταρχήν, τη δυνατότητα κατανόησης της Ποιότητας Εμπειρίας που επιτυγχάνει ένας πάροχος κατά την προσφορά μίας υπηρεσίας. Αυτό μπορεί να επιτευχθεί με την υλοποίηση και ενσωμάτωση μεθόδων αξιολόγησης Ποιότητας Εμπειρίας στην πραγματικού-χρόνου λειτουργία ενός δικτύου. Εν συνεχεία, ακολουθεί η εκμετάλλευση της συλλεγμένης ευφυΐας που σχετίζεται με την Ποιότητα Εμπειρίας, προκειμένου να επανεξεταστούν υφιστάμενοι μηχανισμοί επιπέδου δικτύου (π.χ., χρονο-προγραμματισμός ραδιοπόρων) ή μηχανισμοί επιπέδου εφαρμογής (π.χ., ροή βίντεο), αλλά και να προταθούν καινοτόμες διαστρωματικές προσεγγίσεις προς όφελος της Ποιότητας Εμπειρίας. Επιπλέον, υπάρχει η δυνατότητα πρότασης νέων αλγορίθμων που προκύπτουν από τα εγγενή χαρακτηριστικά της Ποιότητας Εμπειρίας, όπως η μη γραμμική επίδραση μετρικών Ποιότητας Υπηρεσίας στην Ποιότητα Εμπειρίας, με στόχο την περαιτέρω βελτίωσή της. Σε αυτή την κατεύθυνση, στην παρούσα διατριβή, διερευνώνται και αξιοποιούνται μοντέλα και μετρικές εκτίμησης Ποιότητας Εμπειρίας με στόχο την ποσοτικοποίησή της, έχοντας ως απώτερο στόχο την εισαγωγή βελτιώσεων στους υφιστάμενους μηχανισμούς κινητών κυψελωτών δικτύων.

Ο πυρήνας αυτής της διατριβής είναι η πρόταση μίας κυκλικής διεργασίας παροχής Ποιότητας Εμπειρίας που επιτρέπει τον έλεγχο, την παρακολούθηση (ήτοι, τη μοντελοποίηση) και τη διαχείριση της Ποιότητας Εμπειρίας σε ένα κυψελωτό δίκτυο. Κάθε μία από αυτές τις λειτουργίες αναλύεται περαιτέρω, ενώ έμφαση δίνεται στις λειτουργίες μοντελοποίησης και διαχείρισης. Όσον αφορά τη μοντελοποίηση, γίνεται περιγραφή και ταξινόμηση των μεθόδων εκτίμησης και των δεικτών επιδόσεων Ποιότητας Εμπειρίας. Η παραμετρική εκτίμηση της ποιότητας αναδεικνύεται ως η πιο ελκυστική κατηγορία μοντελοποίησης Ποιότητας Εμπειρίας σε κινητά κυψελωτά δίκτυα, οπότε και περιγράφεται διεξοδικά για ευρέως χρησιμοποιούμενους τύπους υπηρεσιών, όπως η συνομιλία (φωνή) μέσω Internet Protocol (IP) και η μετάδοση βίντεο.

Όσον αφορά τη διαχείριση Ποιότητας Εμπειρίας, προτείνονται νέοι μηχανισμοί που επιδεικνύουν βελτιώσεις στην εμπειρία των τελικών χρηστών, και συγκεκριμένα: α) ένα σχήμα ελέγχου των επικοινωνιών συσκευής-προς-συσκευή που λαμβάνει υπόψιν την εμπειρία των χρηστών, β) ένας «συνεπής» αλγόριθμος χρονο-προγραμματισμού ραδιοπόρων που βελτιώνει την Ποιότητα Εμπειρίας του χρήστη μετριάζοντας τις διακυμάνσεις της ρυθμαπόδοσης του δικτύου, και γ) ένας μηχανισμός προσαρμοστικής ροής βίντεο με γνώσεις «πλαισίου», ο οποίος επιτυγχάνει την εξάλειψη διακοπών του βίντεο σε συνθήκες χαμηλού εύρους ζώνης. Επιπλέον, προτείνεται μία εφαρμογή Ποιότητας Εμπειρίας βασισμένη στην αρχιτεκτονική Software-Defined Networking (SDN), ονόματι “QoE-SDN APP”, η οποία επιτρέπει την ανάδραση πληροφοριών δικτύου από παρόχους κινητής τηλεφωνίας σε παρόχους υπηρεσιών βίντεο, αναδεικνύοντας πλεονεκτήματα ως προς την Ποιότητα Εμπειρίας για τους πελάτες των παρόχων βίντεο αλλά και ως προς την εξοικονόμηση εύρους ζώνης για τους φορείς εκμετάλλευσης δικτύου.

Εν κατακλείδι, η παρούσα διατριβή προωθεί την ενοποίηση του ερευνητικού πεδίου της Ποιότητας Εμπειρίας με τον τομέα των κινητών επικοινωνιών, καθώς και τη συνεργασία αμοιβαίου ενδιαφέροντος μεταξύ των παρόχων δικτύου (επίπεδο δικτύου) με τους παρόχους υπηρεσιών (επίπεδο εφαρμογής), αναδεικνύοντας την δυναμική από τέτοιου είδους προσεγγίσεις για όλους τους εμπλεκόμενους φορείς.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Δίκτυα Κινητών Επικοινωνιών

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Ποιότητα εμπειρίας, προσαρμοστική ροή βίντεο μέσω HTTP, χρονο-προγραμματισμός ραδιοπόρων, Software-Defined δικτύωση, κινητά κυψελωτά δίκτυα

Dedicated to my husband Alex, and our baby boy on the way…

Αφιερώνεται στον σύζυγό μου Αλέξη, και στο αγοράκι μας που περιμένουμε…

ACKNOWLEDGMENTS

My enrollment to the PhD program of the Department of Informatics and Telecommunications in the National and Kapodistrian University of Athens, and the research conducted ever since would have never been a reality without the guidance and inspiration offered by my supervisor, Prof. Lazaros Merakos. I am heartedly grateful to him, since he has patiently and methodically helped me reach the point to write this thesis, always being kind and supportive, and always showing trust to me and to any initiatives that I have taken.

My sincerest thanks and gratitude also go to Dr. Nikos Passas, who has continuously mentored and monitored all the phases of my work, with care, conscience and optimism. I cannot overlook the fact that he has been very supportive throughout all the years of my PhD, constantly offering me opportunities to develop as a researcher, to undertake responsibilities related to project management, as well as to attend a plethora of research-, training- and outreach-related activities all over the world.

Moreover, I cannot thank enough Dr. Dimitris Tsolkas, an accomplished researcher and good friend, who has significantly encouraged me and helped me, especially during the first years of my PhD, when I needed his help the most. His careful and methodical assistance has led to the co-authoring of many publications. Furthermore, I would like to express my gratitude to Dr. Dionysis Xenakis, for all his heartfelt advice and our inspiring conversations during my PhD.

Moreover, I would like to thank Prof. Tobias Hoßfeld, for hosting me in the University of Duisburg-Essen in July 2015, and for offering me crucial guidance on QoE topics without any hesitation, as well as Dr. David Lister, who has been a very warm and inspiring Team Leader during my secondment in Vodafone R&D, UK, during January-May 2015. Also, I would like to sincerely thank Dr. Konstantinos Samdanis, Principal Researcher at Huawei, for enlightening me about topics related to SDN, and for being an admirable colleague to collaborate with.

Finally, I cannot miss to express my deepest thanks to my parents Vasilis and Varvara, to my wonderful and supportive husband Alexis, and to my wonderful nephew Spyros and niece Elisabeth, who have all helped me in their own unique ways to make the past four years of my PhD a wonderful, happy, and life-changing journey.

Eirini Liotou

18 October 2017

LIST OF PUBLICATIONS

Conference proceedings:

1. D. Tsolkas, **E. Liotou**, N. Passas, and L. Merakos, “A graph-coloring secondary resource allocation for D2D communications in LTE networks,” *17th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (IEEE CAMAD)*, Barcelona, Spain, September 2012.
2. D. Tsolkas, **E. Liotou**, N. Passas, and L. Merakos, “Enabling D2D communications in LTE networks,” *24th International Symposium on Personal, Indoor and Mobile Radio Communications (IEEE PIMRC)*, London, United Kingdom, September 2013.
3. **E. Liotou**, E. Papadomichelakis, N. Passas, and L. Merakos, “Quality of Experience-centric management in LTE-A mobile networks: The Device-to-Device communication paradigm,” *6th International Workshop on Quality of Multimedia Experience (IEEE QoMEX)*, Singapore, September 2014.
4. **E. Liotou**, D. Tsolkas, N. Passas, and L. Merakos, “Ant Colony Optimization for resource sharing among D2D communications,” *19th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (IEEE CAMAD)*, Athens, Greece, December 2014.
5. **E. Liotou**, G. Tseliou, K. Samdanis, D. Tsolkas, F. Adelantado, and C. Verikoukis, “An SDN QoE-Service for dynamically enhancing the performance of OTT applications,” *7th International Workshop on Quality of Multimedia Experience (IEEE QoMEX)*, Costa Navarino, Greece, May 2015.
6. **E. Liotou**, H. Elshaer, R. Schatz, R. Irmer, M. Dohler, N. Passas, and L. Merakos, “Shaping QoE in the 5G ecosystem,” *7th International Workshop on Quality of Multimedia Experience (IEEE QoMEX)*, Costa Navarino, Greece, May 2015.
7. D. Tsolkas, **E. Liotou**, N. Passas, and L. Merakos, “Addressing traffic demanding scenarios in cellular networks through QoE-based rate adaptation,” *26th International Symposium on Personal, Indoor and Mobile Radio Communications (IEEE PIMRC)*, Hong Kong, China, August 2015.
8. **E. Liotou**, T. Hoßfeld, C. Moldovan, F. Metzger, D. Tsolkas, and N. Passas, “Enriching HTTP adaptive streaming with context awareness: A tunnel case study,” *International Conference of Communications (IEEE ICC)*, Kuala Lumpur, Malaysia, May 2016.
9. **E. Liotou**, D. Tsolkas, and N. Passas, “A roadmap on QoE metrics and models,” *23rd International Conference of Telecommunications (IEEE ICT)*, Thessaloniki, Greece, May 2016.
10. **E. Liotou**, D. Tsolkas, K. Samdanis, N. Passas, and L. Merakos, “Towards Quality of Experience management in the next generation of mobile networks,” *25th European Conference on Networks and Communications (EuCNC)*, Athens, Greece, June 2016.
11. **E. Liotou**, R. Schatz, A. Sackl, P. Casas, D. Tsolkas, N. Passas, and L. Merakos, “The beauty of consistency in radio-scheduling decisions,” *59th Global Communications Conference (IEEE Globecom Wkshps) - International Workshop on*

Quality of Experience for Multimedia Communications (QoEMC), Washington, DC, USA, December 2016.

12. **E. Liotou**, A. Sfikopoulos, P. Kaltzias, and V. Tsolkas, "An evaluation of buffer- and rate-based HTTP adaptive streaming strategies," *22nd International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (IEEE CAMAD)*, Lund, Sweden, June 2017.
13. S. Tennina, I. Tunaru, G. Karopoulos, D. Xenakis, **E. Liotou**, and N. Passas, "Secure energy management in smart energy networks," *22nd International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (IEEE CAMAD)*, Lund, Sweden, June 2017.
14. **E. Liotou**, A. Marotta, L. Pomante, and K. Ramantas, "A middleware architecture for QoE provisioning in mobile networks," *22nd International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (IEEE CAMAD)*, Lund, Sweden, June 2017.

Book chapters / Lecture Notes in Computer Science (LNCS):

1. D. Tsolkas, **E. Liotou**, N. Passas, and L. Merakos, "LTE-A access, core, and protocol architecture for D2D communication," *Smart Device to Smart Device Communication*, Springer International Publishing, Editors: S. Mumtaz and J. Rodriguez, ISBN: 978-3-319-04963-2, pp. 23-40, April 2014.
2. D. Tsolkas, **E. Liotou**, N. Passas, and L. Merakos, "The need for QoE-driven interference management in femtocell-overlaid cellular networks," *Mobile and Ubiquitous Systems: Computing, Networking, and Services*, Springer International Publishing, Editors: I. Stojmenovic, Z. Cheng, and S. Guo, ISBN: 978-3-319-11569-6, vol. 131, pp. 588-601, September 2014 (*Mobiquitous, Tokyo, Japan, December 2013*).
3. F. Metzger, T. Hoßfeld, L. Skorin-Kapov, Y. Haddad, **E. Liotou**, P. Pocta, H. Melvin, V. Siris, A. Zgank, and M. Jarschel, "Context monitoring for improved system performance and QoE," *Autonomous Control for a Reliable Internet of Services: Methods, Models, Approaches, Techniques, Algorithms and Tools*, Springer International Publishing, Editors: I. Ganchev, R. van der Mei, and J. L. van den Berg, *to appear*.
4. R. Schatz, S. Schwarzmann, T. Zinner, O. Dobrijevic, **E. Liotou**, P. Pocta, S. Barakovic, J. Barakovic Husic, and L. Skorin-Kapov, "QoE Management for future networks," *Autonomous Control for a Reliable Internet of Services: Methods, Models, Approaches, Techniques, Algorithms and Tools*, Springer International Publishing, Editors: I. Ganchev, R. van der Mei, and J. L. van den Berg, *to appear*.
5. **E. Liotou**, T. Hoßfeld, C. Moldovan, F. Metzger, D. Tsolkas, and N. Passas, "The value of context-awareness in bandwidth-challenging HTTP Adaptive Streaming scenarios," *Autonomous Control for a Reliable Internet of Services: Methods, Models, Approaches, Techniques, Algorithms and Tools*, Springer International Publishing, Editors: I. Ganchev, R. van der Mei, and J. L. van den Berg, *to appear*.

Peer-reviewed journals:

1. **E. Liotou**, D. Tsolkas, N. Passas, and L. Merakos, “Quality of Experience management in mobile cellular networks: Key issues and design challenges,” *IEEE Communications Magazine, Network & Service Management Series*, vol. 53, no. 7, pp. 145-153, July 2015.
2. D. C. Mocanu, J. Pokhrel, J. P. Garella, J. Seppänen, **E. Liotou**, and M. Narwaria, “No-reference video quality measurement: Added value of machine learning*,” Journal of Electronic Imaging*, vol. 24, no. 6, December 2015.
3. F. Metzger, **E. Liotou**, C. Moldovan, and T. Hoßfeld, “TCP video streaming and mobile networks: Not a love story, but better with context,” *Elsevier Computer Networks, Special Issue on “Traffic and Performance in the Big Data Era*,” vol. 109, pp. 246-256, November 2016.
4. D. Tsolkas, **E. Liotou**, N. Passas, and L. Merakos, “A survey on parametric QoE estimation for popular services,” *Elsevier Network and Computer Applications*, vol. 77, pp. 1-17, January 2017.
5. **E. Liotou**, K. Samdanis, E. Pateromichelakis, N. Passas, and L. Merakos, “QoE-SDN APP: A rate-guided QoE-aware SDN-APP for HTTP adaptive video streaming,” *IEEE Journal on Selected Areas in Communications*, *Series on Network Softwarization & Enablers*, *under review*.

Other publications:

1. **E. Liotou**, N. Passas, and L. Merakos, “Towards QoE provisioning in next generation cellular networks,” *IEEE Communications Society, Multimedia Communications Technical Committee E-Letter, Special Issue on “QoE Management for Next Generation Multimedia Services”*, vol. 10, no. 3, May 2015 (*invited article*).
2. **E. Liotou**, N. Passas, and L. Merakos, “The emergence of experience packages in the 5G era,” *IEEE 5G Tech Focus online journal*, September 2017 (*editor-reviewed*).

ΣΥΝΟΨΗ

Κατά τη διάρκεια των τελευταίων ετών, έχει παρατηρηθεί μία εκθετική αύξηση της δικτυακής κίνησης που προκαλείται από κινητούς χρήστες, ένα φαινόμενο που οφείλεται σε πολλαπλούς παράγοντες. Από τη μία πλευρά, η εμφάνιση των έξυπνων τηλεφώνων και tablets μαζί με την τεράστια ανάπτυξη εφαρμογών λογισμικού έχουν αλλάξει το τοπίο στον τομέα των τηλεπικοινωνιών. Παράλληλα, τα τέλη ακόμη και για εντατική χρήση δεδομένων είναι πλέον ανεκτά, δεδομένου ότι οι φορείς εκμετάλλευσης προσφέρουν πολύ ελκυστικά προφίλ συνδρομής για να προσελκύσουν πελάτες. Από την άλλη πλευρά, τα σύγχρονα δίκτυα, όπως η τεχνολογία Long Term Evolution - Advanced (LTE-A), μπορούν να προσφέρουν πολύ υψηλό εύρος ζώνης στους τελικούς χρήστες και να υποστηρίξουν μεγάλο αριθμό υπηρεσιών, προωθώντας περαιτέρω αύξηση στη ζήτηση κατανάλωσης δεδομένων.

Όλες αυτές οι συνθήκες μετατρέπουν τους χρήστες κινητής τηλεφωνίας σε όλο και πιο απαιτητικούς όσον αφορά την ποιότητα που επιδιώκουν να επιτύχουν, καθώς και σε αρκετά επικριτικούς όταν αυτή η ποιότητα δεν ανταποκρίνεται στις προσδοκίες τους. Αναγνωρίζοντας αυτό το γεγονός, τα τελευταία χρόνια έχει υπάρξει μία δυναμική που ωθεί το επίκεντρο του ενδιαφέροντος από το «δίκτυο» στο «χρήστη». Ως αποτέλεσμα, οι πάροχοι δικτύων καθώς και οι πάροχοι υπηρεσιών έχουν αρχίσει να λαμβάνουν μέτρα προς αυτήν την κατεύθυνση, τα οποία ενισχύονται περαιτέρω από τον έντονο ανταγωνισμό στην αγορά σε αυτή την περιοχή. Προκειμένου να περιγραφούν αυτές οι «χρηστο-κεντρικές» τάσεις, έχουν θεσπιστεί νέοι όροι στη βιβλιογραφία, με πιο κυρίαρχο τον όρο της «Ποιότητας Εμπειρίας» (Quality of Experience – QoE), που περιγράφει τη συνολική αποδοχή μίας εφαρμογής ή μίας υπηρεσίας από ένα χρήστη. Αυτό σημαίνει ότι παλαιότεροι όροι όπως αυτός της Ποιότητας Υπηρεσίας (Quality of Service – QoS), που χρησιμοποιείται παραδοσιακά εδώ και χρόνια, θεωρείται πλέον μόνο μερικός ή ελλιπής. Ο λόγος είναι, ότι η Ποιότητα Υπηρεσίας είναι σε θέση να καταγράψει μόνο τα τεχνικά χαρακτηριστικά μίας υπηρεσίας, αλλά δεν δίνει βέβαιη ένδειξη σχετικά με την ικανοποίηση του χρήστη κατά την αλληλεπίδρασή του με την υπηρεσία. Μάλιστα, η σχέση μεταξύ αυτών των δύο μετρικών (Ποιότητας Υπηρεσίας και Εμπειρίας) είναι μη γραμμική, ενώ πιο συγκεκριμένα έχει αποδειχθεί με υποκειμενικά πειράματα ότι υπάρχει μία εκθετική σχέση μεταξύ τους.

Η έννοια της Ποιότητας Εμπειρίας έρχεται να γεμίσει αυτό το κενό, καθώς αποτελεί υπερσύνολο της Ποιότητας Υπηρεσίας, καθώς και υποκειμενικών και λοιπών παραγόντων «πλαισίου», δηλαδή παραγόντων του ευρύτερου περιβάλλοντος που επηρεάζουν συνειδητά ή ασυνείδητα την εμπειρία του χρήστη. Λόγω αυτής της εγγενούς υποκειμενικότητας, η Ποιότητα Εμπειρίας είναι ένας αρκετά γενικός όρος, που είναι δύσκολο να ποσοτικοποιηθεί. Ωστόσο, η προσεκτική εκτέλεση υποκειμενικών πειραμάτων με ανθρώπινους αξιολογητές έχει οδηγήσει σε αντικειμενικά μοντέλα που είναι σε θέση να μετρήσουν αυτόματα την Ποιότητα Εμπειρίας που συνδέεται με μία συγκεκριμένη σύνδεση και υπηρεσία, «προσομοιώνοντας» τη γνώμη του ίδιου του χρήστη. Κάθε μοντέλο μέτρησης Ποιότητας Εμπειρίας που έχει προταθεί ή προτυποποιηθεί αναφέρεται σε πολύ συγκεκριμένο πεδίο εφαρμογής και σενάριο και προ-απαιτεί την τήρηση υποθέσεων, ώστε να θεωρηθεί έγκυρο. Κατά συνέπεια, η αποκαλούμενη «μοντελοποίηση» της Ποιότητας Εμπειρίας είναι μία πολύ σημαντική ερευνητική πρόκληση.

Η επίγνωση Ποιότητας Εμπειρίας είναι πολύ σημαντική, καθώς μπορεί να αξιοποιηθεί άμεσα από τους παρόχους δικτύων και υπηρεσιών. Πρώτα απ' όλα, αποτελεί τον πιο ελκυστικό και απόλυτο τρόπο αξιολόγησης της απόδοσης των προσφερόμενων υπηρεσιών. Δεύτερον, προβλήματα δικτύου, όπως σημεία συμφόρησης, μπορούν να εντοπιστούν από κατώφλια Ποιότητας Εμπειρίας πυροδοτώντας διορθωτικές ενέργειες

στο δίκτυο (προληπτικά ή εκ των υστέρων). Τέλος, προκύπτει η δυνατότητα ενσωμάτωσης της ίδιας της γνώσης Ποιότητας Εμπειρίας στους μηχανισμούς του δικτύου και συγκεκριμένα στις διαδικασίες λήψης αποφάσεων, ώστε αυτό να λειτουργεί με πιο αποδοτικό και αποτελεσματικό τρόπο. Για παράδειγμα, η Ποιότητα Εμπειρίας μπορεί να αποτελέσει ένα νέο κριτήριο ενεργοποίησης ήδη υπαρχόντων μηχανισμών δικτύου (π.χ., κριτήριο μετάβασης σε λειτουργία συσκευής-προς-συσκευή, μετρική χρονο-προγραμματισμού ραδιοπόρων, κτλ.), αντικαθιστώντας προϋπάρχοντα κριτήρια και μετρικές, όπως είναι οι μετρήσεις ισχύος σήματος. Τέλος, η κατανόηση και αναγνώριση των παραγόντων-κλειδιών που επηρεάζουν την εμπειρία ενός χρήστη με τον πιο ουσιαστικό τρόπο δίνουν τη δυνατότητα πρότασης καινοτόμων αλγορίθμων, που ειδάλλως δε θα μπορούσαν να προκύψουν.

Η μοντελοποίηση και διαχείριση Ποιότητας Εμπειρίας σε κινητά κυψελωτά δίκτυα, και μάλιστα, σε πραγματικό χρόνο, αποτελούν θεμελιώδη υποσυστήματα ενός ευρύτερου πλαισίου για την ολοκληρωμένη παροχή Ποιότητας Εμπειρίας στους τελικούς χρήστες. Ένα τέτοιο πλαίσιο περιλαμβάνει και ευρύτερες προκλήσεις, όπως η συλλογή κατάλληλων δεδομένων εισόδου που θα οδηγήσουν σε επίγνωση Ποιότητας Εμπειρίας, η ρεαλιστική υλοποίηση ενός τέτοιου πλαισίου σε πραγματικά δίκτυα, και η ενδεχόμενη αλληλεπίδραση μεταξύ παρόχων δικτύων και παρόχων υπηρεσιών, με στόχο την ολιστική παροχή βέλτιστης Ποιότητας Εμπειρίας στους τελικούς χρήστες.

Η παρούσα διδακτορική διατριβή εστιάζει στην διερεύνηση των προκλήσεων αλλά και ευκαιριών που προκύπτουν στα σύγχρονα κινητά κυψελωτά δίκτυα ως προς την παροχή Ποιότητας Εμπειρίας στους τελικούς χρήστες. **Συγκεκριμένα, στοχεύει στον χαρακτηρισμό και στην εκμετάλλευση μοντέλων μέτρησης και μετρικών αξιολόγησης Ποιότητας Εμπειρίας, προκειμένου να βελτιωθούν υπάρχοντες μηχανισμοί κυψελωτών δικτύων προτυποποιημένων από την 3GPP (3rd Generation Partnership Project), αλλά και δικτύων στον ορίζοντα του 5G, όπως ο μηχανισμός ραδιο-προγραμματισμού πόρων, η εκκίνηση απευθείας επικοινωνίας συσκευής-προς-συσκευή, και η προσαρμοστική ροή βίντεο.**

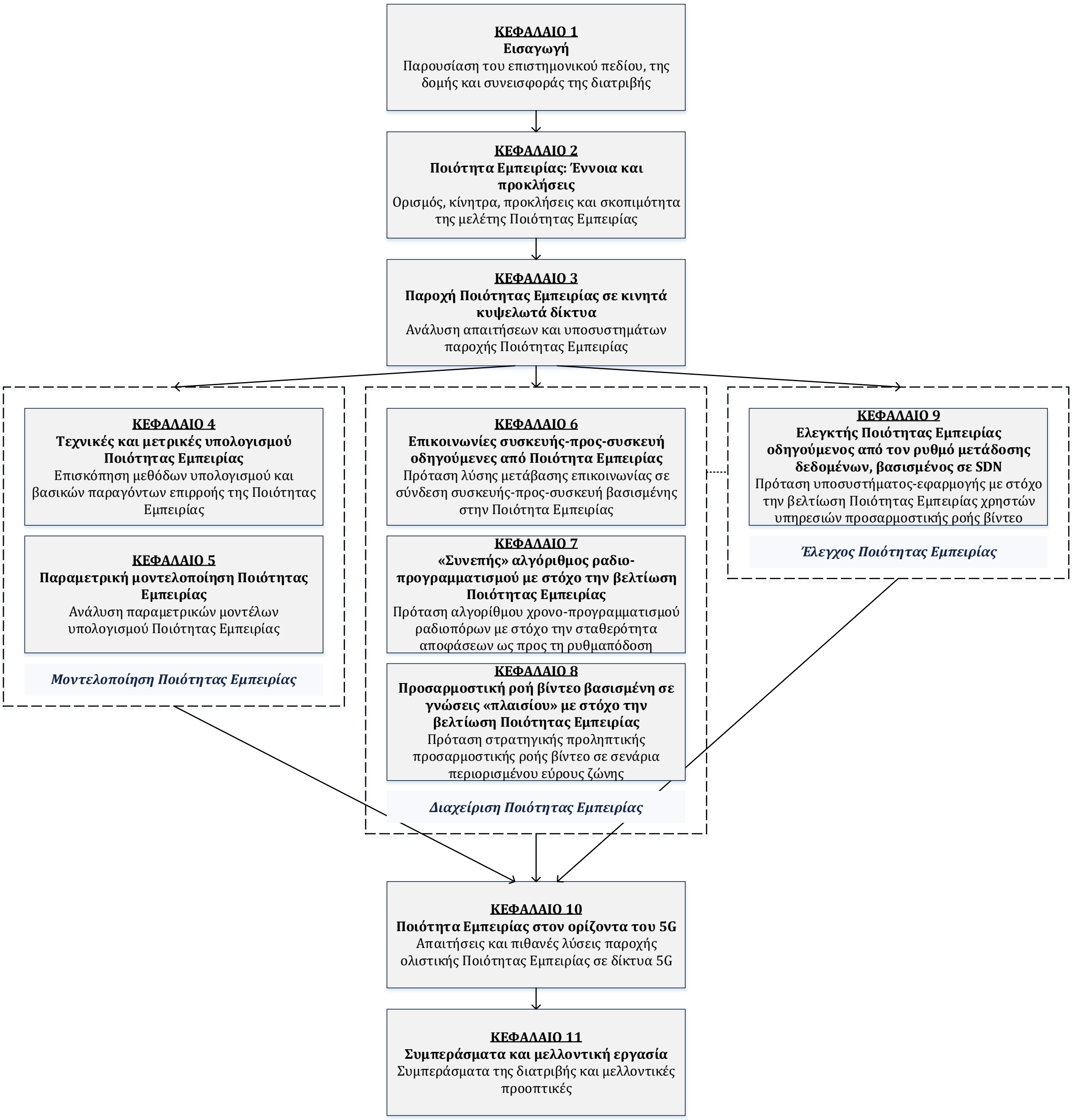
Το περιεχόμενο της διατριβής χωρίζεται σε δέκα κεφάλαια, και ακολουθεί τη δομή που φαίνεται στο Σχήμα I.

Συγκεκριμένα:

Στο **1^{ο} κεφάλαιο** περιγράφεται το θέμα της διδακτορικής διατριβής στα πλαίσια του ευρύτερου επιστημονικού πεδίου όπου ανήκει. Επιπλέον, επεξηγούνται τα κίνητρα, η σκοπιμότητα και η συνεισφορά της διατριβής, καθώς και η δομή που ακολουθεί.

Στο **2^{ο} κεφάλαιο** επεξηγείται ο όρος της Ποιότητας Εμπειρίας, η ανάγκη μετάβασης σε κριτήρια Ποιότητας Εμπειρίας για την αξιολόγηση της απόδοσης σύγχρονων κινητών δικτύων επικοινωνιών, καθώς και οι τεχνικές αλλά και γενικότερες προκλήσεις που προκύπτουν. Επιπλέον, επεξηγείται η συσχέτιση μεταξύ Ποιότητας Εμπειρίας και Υπηρεσίας, ενώ αναδεικνύεται η σημασία αλλά και η σκοπιμότητα διαχείρισης δικτύων και υπηρεσιών με κριτήρια Ποιότητας Εμπειρίας.

Στο **3^{ο} κεφάλαιο** προτείνεται ένα πλαίσιο παροχής Ποιότητας Εμπειρίας σε χρήστες κινητών κυψελωτών δικτύων, το οποίο αποτελείται από τρεις βασικές δομές-υποσυστήματα: α) τη δομή *ελέγχου* μετρικών Ποιότητας Εμπειρίας, β) τη δομή *μοντελοποίησης* (και παρακολούθησης) Ποιότητας Εμπειρίας, και γ) τη δομή *διαχείρισης* Ποιότητας Εμπειρίας στο δίκτυο. Το πλαίσιο αυτό περιγράφεται αναλυτικά ως προς αυτές τις βασικές δομές και τις μεταξύ τους αλληλεπιδράσεις, καθώς και τις προκλήσεις υλοποίησής τους. Αυτές οι δομές ενεργοποιούνται κυκλικά, έτσι ώστε οι τελικές αποφάσεις διαχείρισης και ελέγχου Ποιότητας Εμπειρίας να είναι ανά πάσα στιγμή αποτέλεσμα επικαιροποιημένης και «πραγματικού χρόνου» γνώσης για την κατάσταση του δικτύου και για την εμπειρία των χρηστών. Μία μελέτη αξιολόγησης στο τέλος του κεφαλαίου αποδεικνύει το “proof-of-concept” και τα πιθανά οφέλη από την εφαρμογή ενός τέτοιου συστήματος διαχείρισης ποιότητας πάνω από τις τρέχουσες ή ακόμη και μελλοντικές γενεές κινητών κυψελωτών δικτύων.


ΚΕΦΑΛΑΙΟ 1
Εισαγωγή
Παρουσίαση του επιστημονικού πεδίου, της δομής και συνεισφοράς της διατριβής
ΚΕΦΑΛΑΙΟ 2
Ποιότητα Εμπειρίας: Έννοια και προκλήσεις
Ορισμός, κίνητρα, προκλήσεις και σκοπιμότητα της μελέτης Ποιότητας Εμπειρίας
ΚΕΦΑΛΑΙΟ 3
Παροχή Ποιότητας Εμπειρίας σε κινητά κυψελωτά δίκτυα
Ανάλυση απαιτήσεων και υποσυστημάτων παροχής Ποιότητας Εμπειρίας
ΚΕΦΑΛΑΙΟ 4
Τεχνικές και μετρικές υπολογισμού Ποιότητας Εμπειρίας
Επισκόπηση μεθόδων υπολογισμού και βασικών παραγόντων επιρροής της Ποιότητας Εμπειρίας
ΚΕΦΑΛΑΙΟ 5
Παραμετρική μοντελοποίηση Ποιότητας Εμπειρίας
Ανάλυση παραμετρικών μοντέλων υπολογισμού Ποιότητας Εμπειρίας
Μοντελοποίηση Ποιότητας Εμπειρίας
ΚΕΦΑΛΑΙΟ 6
Επικοινωνίες συσκευής-προς-συσκευή οδηγούμενες από Ποιότητα Εμπειρίας
Πρόταση λύσης μετάβασης επικοινωνίας σε σύνδεση συσκευής-προς-συσκευή βασισμένης στην Ποιότητα Εμπειρίας
ΚΕΦΑΛΑΙΟ 7
«Συνεπής» αλγόριθμος ραδιο-προγραμματισμού με στόχο την βελτίωση Ποιότητας Εμπειρίας
Πρόταση αλγορίθμου χρονο-προγραμματισμού ραδιοπόρων με στόχο την σταθερότητα αποφάσεων ως προς τη ρυθμαπόδοση
ΚΕΦΑΛΑΙΟ 8
Προσαρμοστική ροή βίντεο βασισμένη σε γνώσεις «πλαισίου» με στόχο την βελτίωση Ποιότητας Εμπειρίας
Πρόταση στρατηγικής προληπτικής προσαρμοστικής ροής βίντεο σε σενάρια περιορισμένου εύρους ζώνης
Διαχείριση Ποιότητας Εμπειρίας
ΚΕΦΑΛΑΙΟ 9
Ελεγκτής Ποιότητας Εμπειρίας οδηγούμενος από τον ρυθμό μετάδοσης δεδομένων, βασισμένος σε SDN
Πρόταση υποσυστήματος-εφαρμογής με στόχο την βελτίωση Ποιότητας Εμπειρίας χρηστών υπηρεσιών προσαρμοστικής ροής βίντεο
Έλεγχος Ποιότητας Εμπειρίας
ΚΕΦΑΛΑΙΟ 10
Ποιότητα Εμπειρίας στον ορίζοντα του 5G
Απαιτήσεις και πιθανές λύσεις παροχής ολιστικής Ποιότητας Εμπειρίας σε δίκτυα 5G
ΚΕΦΑΛΑΙΟ 11
Συμπεράσματα και μελλοντική εργασία
Συμπεράσματα της διατριβής και μελλοντικές προοπτικές

Σχήμα I: Δομή διδακτορικής διατριβής.

Έχοντας ορίσει το πλαίσιο παροχής Ποιότητας Εμπειρίας, το οποίο αποτελείται από τις τρεις προαναφερθείσες δομές, πρώτα εστιάζουμε στο θέμα της *μοντελοποίησης*, που αποτελεί αντικείμενο των κεφαλαίων 4 και 5. Συγκεκριμένα:

Στο **4^{ο} κεφάλαιο** εξετάζεται το θέμα της μοντελοποίησης Ποιότητας Εμπειρίας. Συγκεκριμένα, γίνεται ταξινόμηση και συγκριτική μελέτη των διαφόρων μοντέλων αξιολόγησης Ποιότητας Εμπειρίας, καθώς και καταγραφή των βασικών παραγόντων επιρροής της τελικής εμπειρίας ενός χρήστη.

Στο **5^o κεφάλαιο** εντοπίζονται και περιγράφονται παραμετρικές φόρμουλες υπολογισμού Ποιότητας Εμπειρίας για τα πιο δημοφιλή είδη υπηρεσιών (π.χ., Voice over IP (VoIP), βίντεο πραγματικού χρόνου, video-on-demand, περιήγηση στο Διαδίκτυο, Skype, Internet Protocol Television (IPTV) και υπηρεσίες λήψης δεδομένων), καταλήγοντας στους βασικούς δείκτες απόδοσης και παραμετροποίησης ανά τύπο υπηρεσίας. Αυτή η μελέτη έχει ως κύριο στόχο να καλύψει το κενό στη βιβλιογραφία που προκύπτει από την έλλειψη ενός κατάλληλου εγχειριδίου σχετικά με την αντικειμενική εκτίμηση Ποιότητας Εμπειρίας και του συνεχώς αυξανόμενου ενδιαφέροντος προς αυτή την κατεύθυνση. Από τη μελέτη αυτή, αναδεικνύεται ότι οι δείκτες απόδοσης είναι στενά εξαρτώμενοι από τον τύπο υπηρεσίας, και ότι, ακόμη και για την ίδια υπηρεσία, διαφορετικοί παράγοντες συμβάλλουν με διαφορετικό βάρος στην αντίληψη Ποιότητας Εμπειρίας. Αυτό το εύρημα μπορεί να επιτρέψει μία πιο ουσιαστική παροχή πόρων σε διαφορετικές εφαρμογές, σε σύγκριση με αγνωστικά συστήματα ως προς την Ποιότητα Εμπειρίας.

Όσον αφορά τη *διαχείριση* Ποιότητας Εμπειρίας, στα επόμενα κεφάλαια προτείνονται νέοι δικτυακοί μηχανισμοί που μπορούν να βελτιώσουν την αντίληψη των χρηστών ως προς την ποιότητα της εφαρμογής που χρησιμοποιούν. Η περιγραφή αυτών των μηχανισμών αποτελεί το μεγαλύτερο μέρος της διατριβής (κεφάλαια 6, 7, 8 και μέρος του 9). Πιο λεπτομερώς:

Στο **6^o κεφάλαιο** περιγράφεται ένας μηχανισμός μετάβασης μίας σύνδεσης από κυψελωτή λειτουργία σε λειτουργία συσκευής-προς-συσκευή (Device-to-Device – D2D). Οι επικοινωνίες συσκευής-προς-συσκευή αποτελούν αναπόσπαστο μέρος των μελλοντικών κινητών κυψελωτών δικτύων, λόγω των σημαντικών ωφελειών που προσφέρουν τόσο για τους παρόχους δικτύων όσο και για τους τελικούς χρήστες. Υπό αυτή την οπτική γωνία, και συνειδητοποιώντας ότι το κύριο πλεονέκτημα των επικοινωνιών συσκευής-προς-συσκευή είναι η ενδεχόμενη βελτίωση της εμπειρίας των χρηστών, προτείνεται ένα πλαίσιο βασισμένο στην Ποιότητα Εμπειρίας για τη διαχείριση αυτού του τύπου επικοινωνιών. Τα αποτελέσματα προσομοίωσης σε δίκτυο LTE δείχνουν ότι αυτό το πλαίσιο είναι ικανό να μετρήσει και να ενισχύσει τη συνολική εμπειρία των χρηστών κινητής και, κατά συνέπεια, να επιτρέψει αναλογικά οικονομικά οφέλη για τους παρόχους δικτύων.

Στο **7^o κεφάλαιο** περιγράφεται ένας προτεινόμενος μηχανισμός ραδιο-προγραμματισμού με επίγνωση μετρικών Ποιότητας Εμπειρίας. Παρόλο που το πρόβλημα του ραδιο-προγραμματισμού έχει μελετηθεί εκτενώς τις τελευταίες δεκαετίες, πρόσφατα συμπεράσματα από τον τομέα της Ποιότητας Εμπειρίας έρχονται να δώσουν μία νέα προοπτική στις παραδοσιακές προσεγγίσεις. Η συγκεκριμένη μελέτη εκμεταλλεύεται τέτοιου είδους πρόσφατα υποκειμενικά ευρήματα σχετικά με την επίδραση των διακυμάνσεων της ρυθμαπόδοσης δικτύου στην Ποιότητα Εμπειρίας διαδραστικών εφαρμογών, και επανεξετάζει γνωστούς αλγορίθμους ραδιο-προγραμματισμού. Ποσοτικοποιώντας τις επιπτώσεις των παραδοσιακών αλγορίθμων στην αντίληψη Ποιότητας Εμπειρίας του χρήστη, εξάγονται νέα συμπεράσματα, όπως η σημασία και ο αντίκτυπος της «συνέπειας» της κατανομής των πόρων στην Ποιότητα Εμπειρίας των χρηστών. Ως βασικό αποτέλεσμα, οι δίκαιοι αλγόριθμοι φαίνεται να είναι εγγενώς πιο συνεπείς από «άπληστους» αλγορίθμους, παρέχοντας λιγότερες διακυμάνσεις ρυθμαπόδοσης και, ως εκ τούτου, καλύτερη Ποιότητα Εμπειρίας. Με βάση αυτό το συμπέρασμα, προτείνεται μία νέα προσέγγιση ραδιο-προγραμματισμού, η οποία βελτιώνει την Ποιότητα Εμπειρίας των χρηστών, μετριάζοντας τις διακυμάνσεις της ρυθμαπόδοσης.

Στο **8^o κεφάλαιο** περιγράφεται ένας προληπτικός μηχανισμός προσαρμοστικής ροής βίντεο με επίγνωση πληροφοριών «πλαισίου». Η παροχή ροής βίντεο από "Over-The-Top (OTT)" παρόχους υπηρεσιών μέσω ενός κυψελωτού δικτύου είναι ένα πολύ συνηθισμένο σενάριο σήμερα. Ωστόσο, ενώ η ροή βίντεο λειτουργεί αρκετά καλά σε ένα στατικό σενάριο, προκύπτουν διάφορα ζητήματα για κινητούς χρήστες. Για παράδειγμα, η κίνηση εν μέσω σύντομων περιοχών χωρίς δικτυακή κάλυψη, όπως ένα τούνελ, έχει συχνά ως αποτέλεσμα την υποβάθμιση της ποιότητας ή τη διακοπή ενός βίντεο (stalling). Προκειμένου να αντιμετωπιστεί αυτό το πρόβλημα, η παρούσα μελέτη παρέχει μία αναλυτική προσέγγιση του προβλήματος υποβάθμισης της ποιότητας του βίντεο όπως αυτή βιώνεται από κινητούς χρήστες, και προτείνει μία στρατηγική προσαρμοστικής παροχής ροής βίντεο μέσω Hypertext Transfer Protocol (HTTP) (HTTP Adaptive Streaming – HAS) για την πρόληψη διακοπών και, κατ' επέκταση, την ελαχιστοποίηση των αρνητικών επιπτώσεων στην Ποιότητα Εμπειρίας. Επίσης, παρέχει μία λύση που μπορεί να αποτρέψει εντελώς τις διακοπές του βίντεο, όταν κατάλληλες πληροφορίες γενικότερου πλαισίου (όπως πληροφορίες θέσης από δορυφορική πλοήγηση) είναι διαθέσιμες. Τα αποτελέσματα της αξιολόγησης ενθαρρύνουν την περαιτέρω έρευνα σχετικά με το πώς γνώσεις για το γενικότερο πλαίσιο ενός σεναρίου (context awareness) μπορούν να αξιοποιηθούν για την περαιτέρω ενίσχυση της παροχής υπηρεσιών βίντεο από τους OTT παρόχους.

Η τρίτη δομή του πλαισίου παροχής Ποιότητας Εμπειρίας αφορά τις διαδικασίες *ελέγχου* που απαιτούνται για την επίτευξη της προσδοκώμενης ποιότητας. Αυτοί οι μηχανισμοί περιλαμβάνουν, μεταξύ άλλων, την αλληλεπίδραση με το δίκτυο υποδομής για συλλογή πληροφοριών, αλλά και την επικοινωνία με παρόχους υπηρεσιών για την καλύτερη κατανόηση και αποτελεσματικότερη διαχείριση της πραγματικής εμπειρίας των τελικών χρηστών. Συγκεκριμένα:

Στο **9^{ο} κεφάλαιο**, το προτεινόμενο πλαίσιο παροχής Ποιότητας Εμπειρίας που προτάθηκε στο κεφάλαιο 3 επανεξετάζεται με βάση μία Software-Defined Networking (SDN) αρχιτεκτονική, αποκαλύπτοντας με αυτό τον τρόπο νέες προκλήσεις και ευκαιρίες. Συγκεκριμένα, προτείνεται μία αρχιτεκτονική παροχής Ποιότητας Εμπειρίας, η οποία προωθεί τη συνεργατική ανταλλαγή πληροφορίας μεταξύ παρόχων υπηρεσιών βίντεο και παρόχων δικτύου κινητής τηλεφωνίας με στόχο την επίτευξη υψηλότερων επιπέδων Ποιότητας Εμπειρίας χρηστών βίντεο. Κλειδί στην προτεινόμενη αρχιτεκτονική είναι η εφαρμογή "QoE-SDN APP", που βρίσκεται στο επίπεδο ελέγχου της αρχιτεκτονικής SDN, και αναλαμβάνει το ρόλο διαμεσολαβητή μεταξύ των δύο ενδιαφερόμενων μερών, τροφοδοτώντας τον πάροχο υπηρεσιών βίντεο με πληροφορίες που είναι διαθέσιμες μόνο στον πάροχο δικτύου, όπως η ρυθμαπόδοση. Οι δυνατότητες του προτεινόμενου συνεργατικού μοντέλου αναδεικνύονται προτείνοντας και αξιολογώντας τρεις νέες περιπτώσεις χρήσης που προκύπτουν από την εν λόγω αρχιτεκτονική, στα πλαίσια της προσαρμοστικής ροής βίντεο. Σε αυτά τα σενάρια, γνώση σχετικά με τη ρυθμαπόδοση των χρηστών παρέχεται σε έναν πάροχο βίντεο, προκειμένου αυτός να είναι σε πιο ισχυρή θέση να επαναπροσδιορίσει την κωδικοποίηση, αποθήκευση (caching), αλλά και την ανά-χρήστη επιλογή κατάλληλων κωδικοποιήσεων βίντεο (video segment selection).

Στο **10^{ο} κεφάλαιο** γίνεται μία μελέτη ως προς την ενσωμάτωση απαιτήσεων Ποιότητας Εμπειρίας στο οικοσύστημα δικτύων πέμπτης γενιάς (5G). Για το σκοπό αυτό, εντοπίζονται και αναλύονται ουσιαστικά χαρακτηριστικά που μπορούν να διαμορφώσουν χρηστο-κεντρικά δίκτυα. Τέλος, προτείνεται η υιοθέτηση «πακέτων εμπειρίας» (experience packages), που οδηγούν σε μία πιο προσωποποιημένη παροχή υπηρεσιών στους χρήστες, λαμβάνοντας υπόψιν όχι μόνο τεχνικές παραμέτρους, αλλά και το προφίλ του χρήστη, καθώς και το γενικότερο πλαίσιο (context) της επικοινωνίας.

Τέλος, στο **11^{ο} κεφάλαιο** παρουσιάζονται τα συμπεράσματα της διδακτορικής διατριβής, καθώς και ανοιχτές δυνατότητες για περαιτέρω μελλοντική έρευνα.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

This thesis describes original research work conducted by the author, Mrs. Eirini Liotou, during April 2013 to November 2017 under the supervision of Prof. Lazaros Merakos and the mentorship of Dr. Nikos Passas. It has been realized in the Green, Adaptive and Intelligent Networking Group (GAIN), a research group within the Communication Networks Laboratory (CNL) of the Department of Informatics and Telecommunications in the National and Kapodistrian University of Athens (NKUA). The results of this work have been published in international conferences, peer-reviewed journals and book chapters. Mrs. Liotou's published papers have received 154 citations to date, according to Google Scholar.

Along with her research activity, Mrs. Liotou has participated in the realization of European research projects (e.g., CROSSFIRE, SMART-NRG, CASPER, COST-QUALINET, COST-ACROSS, etc.), and has made contributions to the writing of proposals for new research projects within the framework of the European research program HORIZON2020, but also to national calls. Moreover, she has co-supervised undergraduate and postgraduate theses in the Department of Informatics and Telecommunications, NKUA (11 completed theses and 3 currently under development).

In the context of scientific contributions to the broader research community, Mrs. Liotou has been a reviewer at scientific conferences (IEEE ICC, IEEE GLOBECOM, IEEE CAMAD, IEEE QoMEX, IEEE PIMRC, EuCNC, etc.) and journals (IEEE Communications Magazine, IEEE Journal of Selected Topics in Signal Processing, IEEE Transactions on Vehicular Technology, Elsevier Computer Communications, ACM Transactions on Multimedia Computing Communications and Applications, etc.) as well as Technical Program Committee Member in IEEE QoMEX 2015, QEEMS 2017, INFOCOM-CNTCV 2017, and ACROSS 2017. She is an IEEE Graduate Student member, as well as a member of the IEEE Young Professionals, IEEE Communications Society, IEEE Women in Engineering, IEEE 5G Initiative, IEEE 5G Community, and ETSI STQ. In parallel, she has given invited talks at several workshops and events, such as the Qualinet final workshop (8-10/10/2014, Delft), Dagstuhl seminar on "Quality of Experience: From Assessment to Application" (4-7/1/2015, Schloss Dagstuhl), CROSSFIRE open day (25/3/2015, King's College London), and ETSI workshop on "Telecommunication Quality beyond 2015" (21/10/2015, A1 Telekom Austria).

During her PhD, she visited the M2M & 5G Technologies Group of VODAFONE R&D Technology, Newbury, UK, in the context of studying Quality of Experience towards the horizon of 5G networks (January-May 2015). She also visited the Chair of Modeling of Adaptive Systems at the University of Duisburg-Essen, Germany, in the context of designing advanced adaptive video streaming mechanisms (July 2015). Finally, she has participated in two summer schools (3rd Qualinet summer school and 1st ACROSS summer school), while she has attended six scientific conferences to present her research work.

Eirini Liotou

18 October 2017

1. INTRODUCTION

1.1 Thesis motivation and scope

Over the last few years, there has been a tremendous increase in the network traffic generated by mobile users, a phenomenon which can be attributed to multiple factors. On the one hand, the emergence of smart phones and tablets along with the huge, recently emerged app market have changed the landscape in the telecommunications sector. In parallel, the charges even for intensive data usage are tolerable, as network operators offer very attractive subscription packets to attract customers. On the other hand, modern networks, such as the Long Term Evolution - Advanced (LTE-A) and emerging 5G networks, can offer very high bandwidth to their users, supporting a plethora of diverse, resource-hungry services, and further boosting the demand for data consumption. All these conditions make mobile users more and more demanding in terms of the quality they expect to achieve.

Recognizing this fact, there has lately been a momentum that pushes the epicenter of interest from the "network" to the "user". While network and service providers are trying to create or follow this "user-centric" trend, new terms have been coined that allow its more comprehensive description. The term "Quality of Experience" (QoE) is irrefutably the most dominant one, as it describes "*the overall acceptability of an application or service, as perceived subjectively by the end-user*". This means that older terms such as Quality of Service (QoS), traditionally used for years, are now considered only partial or incomplete. The reason behind that is that QoS can only record the technical characteristics of a service without giving a clear indication of the user's satisfaction when interacting with this service. In fact, the relationship between these two metrics (QoS and QoE) has been found to be non-linear.

The definition of QoE makes clear that it is a very broad and generic concept, and as such, it incorporates the complete end-to-end system effects (terminal, network, services, etc.) together with the human impressions of these effects. QoE actually incorporates all conscious and unconscious aspects that affect the overall satisfaction of a user, including the overall context of the communication scenario (e.g., communication task, surrounding environment, pricing, etc.). As vague as the concept of QoE may sound, reliable estimation methods have been developed with the assistance of subjective experiments with human evaluators. These experiments lead to reliable QoE assessment methods, which manage to automatically evaluate and rate the QoE of a user with respect to a specific application or service. This procedure is called "QoE modeling", and it is the most important first step towards QoE provisioning.

The awareness of an overall QoE score is very important for all involved stakeholders in the service communication chain. Once QoE is measured, this may be exploited in many aspects by network operators and service providers. First of all, the extraction of a QoE score of a service with respect to a user is the most attractive and absolute way to evaluate the performance of the offered services. Second, network problems such as bottlenecks or local failures may be identified by predefined QoE thresholds, and proactive or reactive actions may be triggered to correct them. A third important motive for QoE awareness is the possibility to incorporate QoE intelligence in the network mechanisms, and specifically in the network decision processes. This may lead to "QoE-driven" or otherwise called "QoE-aware" algorithms that can help the network function in a more efficient and effective way. For instance, QoE may become the criterion or trigger mechanism of standard network algorithms (e.g., radio resource scheduling, mobility management, power control, etc.) replacing current QoS-based criteria, such as plain signal strength measurements. What is more, understanding and identifying the key factors that truly affect the user's experience creates the possibility to propose innovative algorithms that focus on targeted QoE performance indicators. Finally, QoE-awareness may drive a more resource-efficient network operation, by helping recognize moments and cases of operation when providing extra resources to the users would not improve their perceived QoE. In other words, "over-engineering" could be avoided.

QoE modeling and management in mobile cellular networks are fundamental components, part of a wider framework that enforces the end-to-end QoE provisioning. This framework also includes wider challenges such as the collection of appropriate input data that will lead to the awareness of QoE (i.e., QoE monitoring), the realistic implementation of such a framework in real networks, and the possible interaction between network providers and service providers, aiming at the holistic delivery of optimal QoE to the end-users, among others.

This PhD thesis focuses on exploring the challenges and opportunities that arise in modern mobile cellular networks in terms of QoE provisioning to end-users. Specifically, **this thesis aims to characterize and exploit QoE models and metrics in order to improve existing mechanisms in mobile cellular networks standardized by 3GPP (3rd Generation Partnership Project), but also towards the 5G horizon, such as the radio resource allocation, Device-to-Device communication setup, and adaptive video streaming mechanisms.**

1.2 Thesis contributions

In this thesis, the reader will delve into details regarding the topic of QoE management in mobile cellular communication networks. The main contributions of the research conducted in this thesis are the following:

1. Proposal of a conceptual framework for achieving end-to-end QoE provisioning in mobile cellular networks. This framework is analyzed in terms of its design, its constituents and their interactions, as well as key implementation challenges, while its proof-of-concept in an LTE network is assessed. Related publication:

    - *E. Liotou, D. Tsolkas, N. Passas, and L. Merakos, "Quality of Experience management in mobile cellular networks: Key issues and design challenges," IEEE Communications Magazine, Network & Service Management Series, vol. 53, no. 7, pp. 145-153, July 2015.*

2. The identification and analysis of parametric QoE formulas and Key Performance Indicators (KPIs) that can be used for real-time QoE assessment of popular service types in communication networks (i.e., VoIP, online video, video streaming, web browsing, Skype, IPTV and file download services). Related publications:

    - *E. Liotou, D. Tsolkas, and N. Passas, "A roadmap on QoE metrics and models," 23rd International Conference of Telecommunications (IEEE ICT), Thessaloniki, Greece, May 2016.*

    - *D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, "A survey on parametric QoE estimation for popular services," Elsevier Network and Computer Applications, vol. 77, pp. 1-17, January 2017.*

3. A network management framework that exploits QoE awareness for controlling the operational mode of mobile users in LTE-A networks with Device-to-Device (D2D) support. Simulation studies have revealed the twofold benefits of this mechanism, i.e., both for the users (increase in QoE) and the operators (increase in offered throughput). Related publication:

    - *E. Liotou, E. Papadomichelakis, N. Passas, and L. Merakos, "Quality of Experience-centric management in LTE-A mobile networks: The Device-to-Device communication paradigm," 6th International Workshop on Quality of Multimedia Experience (IEEE QoMEX), Singapore, September 2014.*

4. Proposal of a new radio scheduling logic, which takes into account the impact of throughput fluctuations on the QoE of interactive applications. By quantifying how traditional radio scheduling decisions influence the user-perceived QoE, a novel “consistent” resource allocation process is proposed, which further improves users’ QoE by moderating these fluctuations. Related publication:

   - *E. Liotou, R. Schatz, A. Sackl, P. Casas, D. Tsolkas, N. Passas, and L. Merakos, “The beauty of consistency in radio-scheduling decisions,” 59th Global Communications Conference (IEEE Globecom Wkshps) - International Workshop on Quality of Experience for Multimedia Communications (QoEMC), Washington, DC, USA, December 2016.*

5. Analytical investigation of the video quality degradation problem as it is experienced by mobile users in vehicles, and proposal of a proactive context-aware HTTP Adaptive Streaming (HAS) strategy, which helps prevent stallings in light of bandwidth-challenging situations. Related publications:

   - *E. Liotou, T. Hoßfeld, C. Moldovan, F. Metzger, D. Tsolkas, and N. Passas, “Enriching HTTP adaptive streaming with context awareness: A tunnel case study,” International Conference of Communications (IEEE ICC), Kuala Lumpur, Malaysia, May 2016.*
   - *F. Metzger, E. Liotou, C. Moldovan, and T. Hoßfeld, “TCP video streaming and mobile networks: Not a love story, but better with context,” Elsevier Computer Networks, Special Issue on “Traffic and Performance in the Big Data Era,” vol. 109, pp. 246-256, November 2016.*
   - *E. Liotou, T. Hoßfeld, C. Moldovan, F. Metzger, D. Tsolkas, and N. Passas, “The value of context-awareness in bandwidth-challenging HTTP Adaptive Streaming scenarios,” Autonomous Control for a Reliable Internet of Services: Methods, Models, Approaches, Techniques, Algorithms and Tools, Springer International Publishing, Editors: I. Ganchev, R. van der Mei, and J. L. van den Berg, to appear.*

6. Proposal of a Software-Defined Networking (SDN)-based architecture that promotes and enables a technologically feasible realization of a collaboration paradigm between service providers and mobile network operators. The potential of this architecture is highlighted through the proposal and evaluation of three use cases that are unlocked by this architecture, in the context of HAS. In this paradigm, feedback about the network throughput is provided to a video service provider so that he can be in a stronger position to redefine encoding, caching, and per-user video segment selection. Related publications:

   - *E. Liotou, G. Tseliou, K. Samdanis, D. Tsolkas, F. Adelantado, and C. Verikoukis, “An SDN QoE-Service for dynamically enhancing the performance of OTT applications,” 7th International Workshop on Quality of Multimedia Experience (IEEE QoMEX), Costa Navarino, Greece, May 2015.*
   - *E. Liotou, D. Tsolkas, K. Samdanis, N. Passas, and L. Merakos, “Towards Quality of Experience management in the next generation of mobile networks,” 25th European Conference on Networks and Communications (EuCNC), Athens, Greece, June 2016.*
   - *E. Liotou, K. Samdanis, E. Pateromichelakis, N. Passas, and L. Merakos, “QoE-SDN APP: A rate-guided QoE-aware SDN-APP for HTTP adaptive video streaming,” IEEE Journal on Selected Areas in Communications, Series on Network Softwarization & Enablers, under review.*

7. Identification of the essential attributes that can shape QoE-centric networks towards the 5G era, and introduction of the “experience package” concept. Experience packages can lead to a more personalized service provisioning to users, considering not only technical parameters, but also the user profile and the context of the communication. Related publications:

   - *E. Liotou, N. Passas, and L. Merakos, “Towards QoE provisioning in next generation cellular networks,” IEEE Communications Society, Multimedia Communications Technical Committee E-Letter, Special Issue on “QoE Management for Next Generation Multimedia Services”, vol. 10, no. 3, May 2015.*

- *E. Liotou, H. Elshaer, R. Schatz, R. Irmer, M. Dohler, N. Passas, and L. Merakos, “Shaping QoE in the 5G ecosystem,” 7th International Workshop on Quality of Multimedia Experience (IEEE QoMEX), Costa Navarino, Greece, May 2015.*
- *E. Liotou, N. Passas, and L. Merakos, “The emergence of experience packages in the 5G era,” IEEE 5G Tech Focus online journal, September 2017.*

1.3 Thesis structure

This thesis consists of 11 chapters and follows the conceptual structure that is depicted in Figure 1.

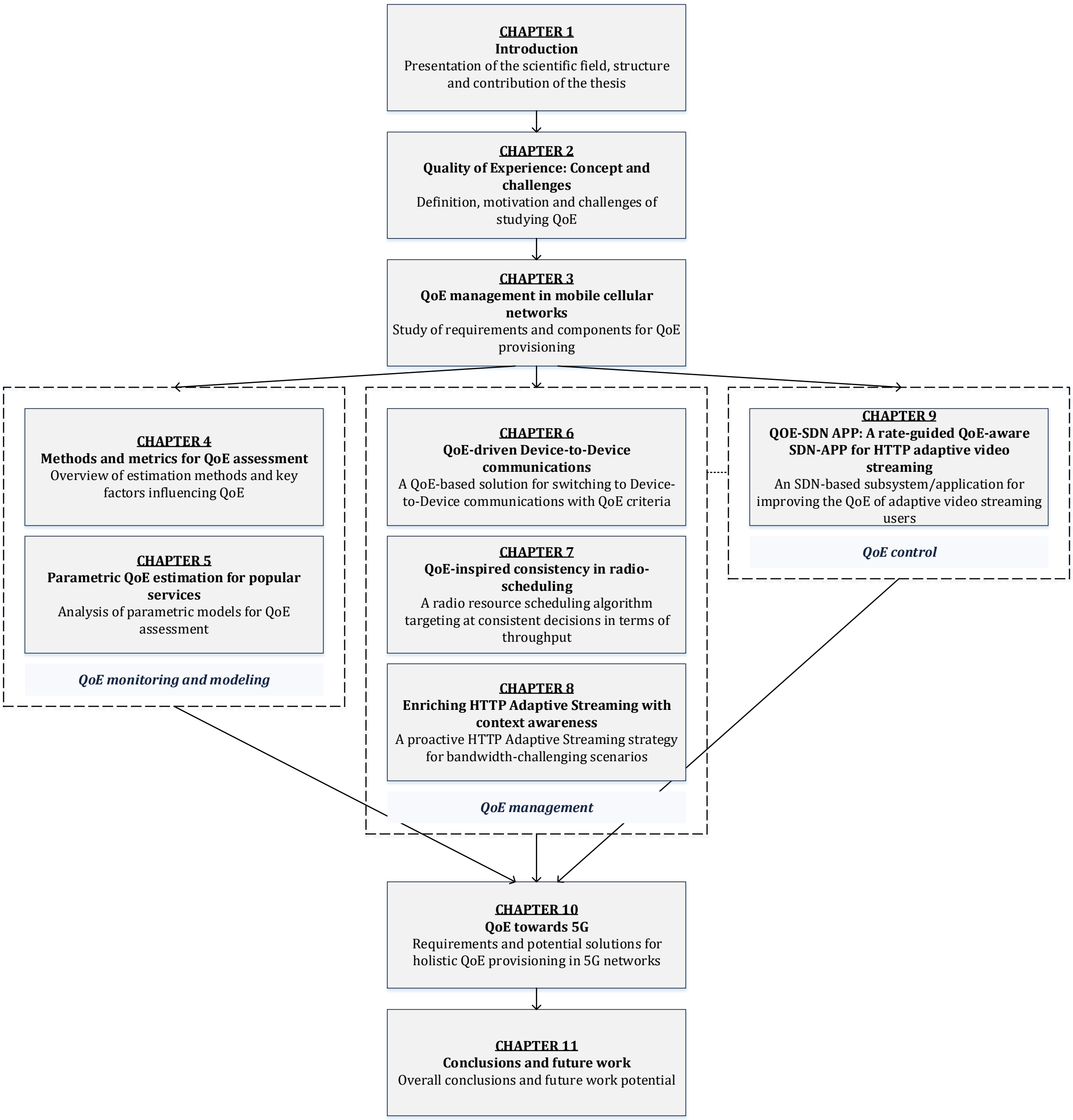

CHAPTER 1
Introduction
Presentation of the scientific field, structure and contribution of the thesis
CHAPTER 2
Quality of Experience: Concept and challenges
Definition, motivation and challenges of studying QoE
CHAPTER 3
QoE management in mobile cellular networks
Study of requirements and components for QoE provisioning
CHAPTER 4
Methods and metrics for QoE assessment
Overview of estimation methods and key factors influencing QoE
CHAPTER 5
Parametric QoE estimation for popular services
Analysis of parametric models for QoE assessment
QoE monitoring and modeling
CHAPTER 6
QoE-driven Device-to-Device communications
A QoE-based solution for switching to Device-to-Device communications with QoE criteria
CHAPTER 7
QoE-inspired consistency in radio-scheduling
A radio resource scheduling algorithm targeting at consistent decisions in terms of throughput
CHAPTER 8
Enriching HTTP Adaptive Streaming with context awareness
A proactive HTTP Adaptive Streaming strategy for bandwidth-challenging scenarios
QoE management
CHAPTER 9
QOE-SDN APP: A rate-guided QoE-aware SDN-APP for HTTP adaptive video streaming
An SDN-based subsystem/application for improving the QoE of adaptive video streaming users
QoE control
CHAPTER 10
QoE towards 5G
Requirements and potential solutions for holistic QoE provisioning in 5G networks
CHAPTER 11
Conclusions and future work
Overall conclusions and future work potential

Figure 1: Thesis structure.

Following the current Chapter 1 that gives an overview of the scope and contributions of this thesis, Chapter 2 introduces the reader to the concept of QoE and sets the background for the rest of this thesis. Then, Chapter 3 describes the basic framework and functionalities for the purposes of QoE provisioning in mobile cellular networks. These functionalities are related to QoE monitoring and modeling - further analyzed in

Chapters 4 and 5, b) QoE management - further analyzed in Chapters 6, 7 and 8, and c) QoE control - further analyzed in Chapter 9. More specifically, Chapter 4 describes basic methods, tools and metrics for the assessment of QoE, while Chapter 5 elaborates on a subset of these methods, called parametric methods, which allow the real-time QoE monitoring in a communication network. With respect to QoE management, Chapter 6 describes an algorithm for switching from cellular mode to D2D communication mode, based on QoE criteria. Moreover, Chapter 7 describes a QoE-inspired radio scheduler that stems from subjective studies' findings in the context of QoE, while Chapter 8 uses context-awareness to improve the QoE of adaptive video streaming users. Chapter 9 describes an SDN-based architecture for end-to-end QoE improvement of video services, which includes all QoE functionalities (monitoring-management-control), while its core lies in the QoE control function. Finally, Chapter 10 discusses some insights towards QoE provisioning in the 5G era, while Chapter 11 concludes this thesis and presents ideas and opportunities for future work.

2. QUALITY OF EXPERIENCE: CONCEPT AND CHALLENGES

2.1 Definitions

The notion of Quality of Experience (QoE) has appeared at around the beginning of this century. It is probably impossible to trace back exactly when or who coined this term; however, many references to QoE appear at around that time. For instance, in 2000 we can find a reference to QoE by Patricia Seybold consulting group [1], as a quality benchmark that measures how well an e-business delivers the expected branded experience to its customers. Then, in 2001, we can find a reference of QoE in [2], where Aad van Moorsel from Hewlett-Packard Laboratories supports that the user experience becomes increasingly important in the "Internet age". Then, a 2004 white paper from Nokia [3] clearly defines QoE, stating that: "*The ultimate measure of a network and the services it offers is how subscribers perceive the performance. QoE is the term used to describe this perception and how usable the subscribers think the services are*.". Moreover, this insightful white paper discusses QoE implications on business, as well as groups Key Performance Indicators (KPIs) into two main categories, i.e., reliability and comfort, where reliability is defined as "*the availability, accessibility and maintainability of the content, the service network and/or the user device application software*", while comfort refers to "*the quality of the content, the bearer service and/or the software features of the user device and application*".

Later, formal definitions of QoE also appeared by various standardization bodies and other groups. The formal definition of QoE is provided by the International Telecommunication Union - Telecommunication Standardization Sector (ITU-T) Rec. P.10 (Amendment 2, 2008) [4], as "*the overall acceptability of an application or service, as perceived subjectively by the end-user*". Based on this approach, two issues need to be noted, namely: a) "*QoE includes the complete end-to-end system effects (client, terminal, network, services infrastructure, etc.)*", and b) "*the overall acceptability may be influenced by user expectations and context*". The purpose of this new concept is to provide means to track the degree of user satisfaction of a network's performance in a qualitative or quantitative manner and to try to improve it in order to meet or exceed the users' expectations. As an overall, QoE addresses the issue of a service's acceptability, attractiveness and sale-ability.

The European Telecommunications Standards Institute (ETSI) provides another formal definition of QoE, as: "*A measure of user performance based on both objective and subjective psychological measures of using an ICT service or product*". Moreover, it notes that QoE "*takes into account technical parameters (e.g. QoS) and usage context variables (e.g. communication task) and measures both the process and outcomes of communication (e.g. user effectiveness, efficiency, satisfaction and enjoyment)*" [5].

Finally, "QUALINET", the European Network of Excellence on QoE in Multimedia Systems and Services, provides an insight to the QoE notion and its underlying principles [6]. It first defines the term *quality* as "*the outcome of an individual's comparison and judgment process*" and the term *experience* as "*an individual's stream of perception and interpretation of one or multiple events*". Subsequently, QUALINET defines QoE as: "*The degree of delight or annoyance of the user of an application or service*", also adding that "*it results from the fulfilment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state*".

A user's impression on QoE may start to be formed much before the actual usage of a service (or product) and may continue even after usage. Specifically, according to [7], the experience of a user spans across four different and subsequent time events.

Initially, the potential user of a service forms an *anticipated* experience, which refers to the effect on QoE before really using a service, i.e., based on one's own anticipations, other people's opinions, advertisements, brand, etc. Later, during the actual user interaction with the service, the *momentary* experience is formed, causing either positive or negative feelings, whereas the *episodic* experience is based on the user's reflection of this interaction, after its completion. Finally, after a person has used a service multiple times over a larger period of time, this person has created a *cumulative* view about this experience.

Many terms related to quality are available in the literature and most of them are presented below. All of these terms may be assumed to be incorporated into the much broader and generic concept of QoE.

- **Quality of Service (QoS):** As defined by the ITU-T Rec. E.800 [8], QoS is "*the totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service*". Alternatively, according to ITU Development Sector (ITU-D) Study Group 2 [9], QoS is "*a collective of service performances that determine the degree of satisfaction of a user of a service*", or according to the Internet Engineering Task Force (IETF) [10], it is "*a set of service requirements to be met by the network while transporting a flow*".

  QoS may be further divided into four main viewpoints, depicted in Figure 2: a) the QoS requirements of the user/customer, b) the QoS offered/planned by the service provider, c) the QoS delivered/achieved by the service provider, and d) the QoS actually experienced/perceived by the user. The latter viewpoint (d) actually corresponds to the QoE concept itself. Nevertheless, despite these definitions that closely relate QoS to the user's satisfaction, QoS has been traditionally handled as a pure technical term, providing system-centric rather than user-centric quality guarantees.

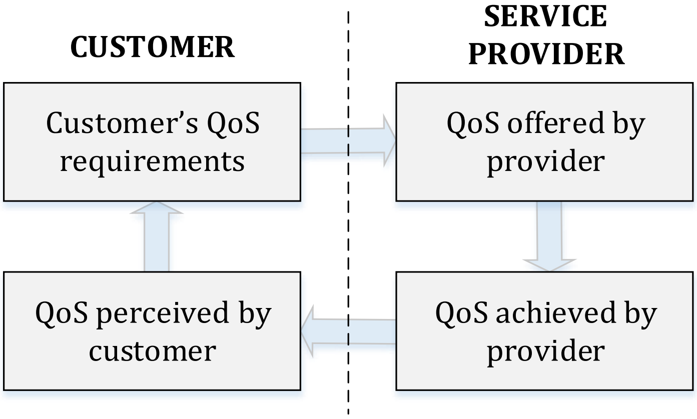


Figure 2: The four viewpoints of QoS [8].

- **Grade of Service (GoS):** The GoS term incorporates the quality a user can expect to experience when initiating a service and mainly relates to the network's availability, together with the call setup blocking probability and session establishment delays.
- **Quality of Resilience (QoR):** This term describes the network's reliability and survivability against disastrous situations such as local failures or malicious attacks. Hence, it embraces security and privacy issues.
- **Quality of Perception (QoP):** QoP represents the user's side of the more technical and traditional QoS. According to [11], "*QoP encompasses not only a user's*

satisfaction with the quality of multimedia presentations, but also his/her ability to analyze, synthesize and assimilate the informational content of multimedia displays".

- **Quality of Design (QoD)** and **Quality of Conformance (QoC):** The former term (QoD) refers to the fitness for use of a product or service, i.e., the level at which the operator or producer intends to fulfill the customer requirements. This is indicated by the completeness and correctness of the service's specifications and is closely related to the QoS offered by the service provider. The latter term (QoC), refers to the quality actually produced and delivered to the customers, complying fully or partially with the originally planned QoD [12]. So, we may identify a connection between the QoC and the QoS achieved by the provider.
- **Quality of Business (QoB or QoBiz):** This term appears in [2] and is a metric expressed in terms of money, such as the average amount of money received per executed transaction. QoB is mainly influenced by cost and revenue considerations.
- **Quality of User Experience (QoUE)** and **Quality of Customer Experience (QoCE):** Both are synonyms to QoE, with a focus on the specific different role of the person using a service, i.e., "user" or "customer".
- **QoX:** This is just another way to abbreviate "Quality of eXperience", that may be found in literature.

2.2 QoE dependencies

From the previous section, it is inferred that QoE is a multi-factor concept, depending on a plethora of multiple and diverse parameters. According to [13], the main properties of QoE are User-, Application-, Terminal-, and Time-dependency.

- **User dependency** means that users may perceive QoE in different ways even when receiving the same service, they may show different preferences regarding their sessions, or they may prioritize different factors as important. Moreover, due to their variations in emotions, expectations or experiences, they may evaluate services that offer the same QoS much differently.
- **Application dependency** describes the different impact of different applications on QoE. This is a main property of QoE. Different applications have different technical requirements, influence factors and constraints. For instance, VoIP applications are delay-sensitive, whereas video applications are bandwidth-sensitive. This implies that QoE should be evaluated in a completely different way per application and that different QoE management objectives should be devised per application type.
- **Terminal dependency** describes the impact of diverse devices on QoE in terms of their technical characteristics, capabilities and limitations. For instance, characteristics such as resolution, colour or screen size seem to play a key role in the perceived QoE of the user. However, potential device limitations may be sometimes falsely attributed to network or service deficiencies. Moreover, powerful devices may increase user expectations in terms of achieved QoE.
- **Time dependency**, finally, stems from the fact that many of the QoE influence factors are time-variant and thus, difficult or impossible to control. These factors may range from fluctuating user subjectivity to unstable wireless channel conditions.

The authors in [13] conclude that, due to the above dependencies, QoE needs to be managed on a per-user, per-application, and per-terminal basis in a real-time way.

As shown in Figure 3 in a relatively abstract way, the QoE finally perceived by the user is the result of many parameters of different layers. It is the result of a) QoS-related KPIs related to the network infrastructure and network mechanisms, b) application-related parameters that depend on the type of application considered, c) user-related Key Quality Indicators (KQIs) such as reliability and integrity, and finally d) factors related to user personality, usage context, device, etc. According to [14], each one of the lower layer parameters can be mapped to higher layer ones, so that eventually the final QoE is the weighted sum of multiple KQIs.

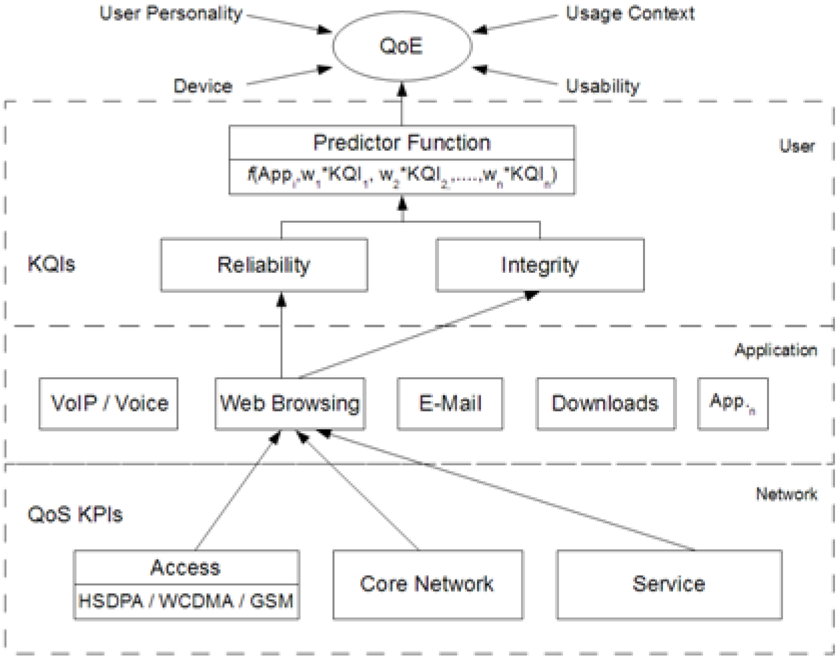


Figure 3: Overall QoE formation [14].

There also exist diverse approaches in the literature, which try to explain how a QoE opinion is formed, namely which dimensions influence the user perception. Many works differentiate the “Content” factor as significant, and break QoE into System, Human, Context and Content parts. Another approach is the ARCU multi-dimensional model, which is composed of the Application, Resource, Context and User space components and is proposed in [15]. Moreover, [16] describes QoE through four main attributes: the communication situation, service prescription, technical parameters and user experience. In Chapter 3, we thoroughly present the approach proposed by QUALINET in [6].

2.3 The importance of QoE

The acquisition of a QoE score of an application or service is of crucial importance, not only to the user but also to various stakeholders in the service provisioning chain. For instance, service providers, network operators, equipment manufacturers, marketing teams and customer support agents with the privilege of knowing the QoE of their offered product/service, may have direct financial advantages.

The importance of QoE awareness is justified, if we have a look at statistics regarding customer churn [17]:

- 82% of customer defections are due to frustration and the provider’s inability to deal with this effectively.
- For 1 person who calls with a problem, 29 others never will.
- 1 frustrated customer will tell 13 others.
- A 90% of the customers abandons a service without even complaining.

Especially for the case of communication networks, QoE may provide a better insight to the network operators regarding the quality of their offered services. More specifically, QoE intelligence is invaluable to telecom operators, since it can (Figure 4): a) Enable Customer Experience Management (CEM) through QoE-oriented data analytics (e.g., automate service configuration, facilitate self-care and self-diagnosis through QoE analytics, reduce or prevent customer churn and offer troubleshooting), b) Drive business operations, enable strategic business decisions and build more meaningful Service Level Agreements (SLAs) or Experience Level Agreements (ELAs), c) Decrease churn, by comprehending users' and applications' requirements and controlling the network accordingly (namely, avoid under-engineering, proactively predict and prevent network problems, or reactively improve QoE), and finally, d) Increase network efficiency through identifying and exploiting the non-linear relationships between QoS and QoE (namely, avoid over-engineering, e.g., reduce energy consumption or save spectrum resources without sacrificing the QoE).

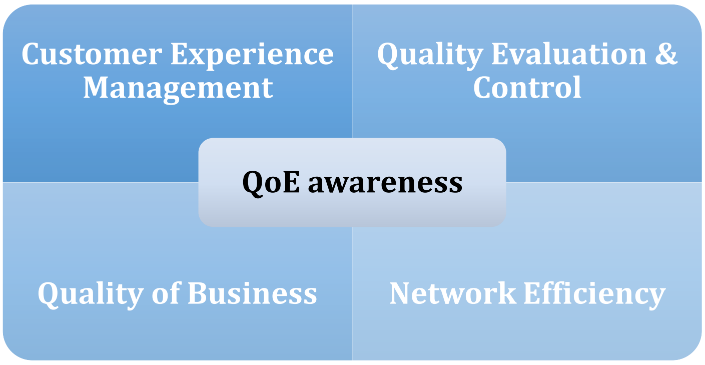


Figure 4: The importance of QoE awareness for network operators.

In more detail, we identify the usefulness and importance of QoE awareness during the whole lifecycle of a network, from its original design and planning, to its testing, maintenance and improvement, in the following aspects:

- **Network design and planning:** QoE-awareness can help design resource- and energy-efficient networks from scratch, by avoiding "under-engineering" cases of providing fewer resources than required. Similarly, "over-engineering" may be avoided, through the release of occupied network resources that are redundant in terms of the finally perceived quality. Hence, a more resource-efficient network operation would be possible, by helping recognize moments and cases of operation when the provisioning of extra resources to the users would not improve the QoE perceived, and by parameterizing the network accordingly. Hence, infrastructure and capacity planning or network reconfiguration may be performed using continuous QoE assessment scores as a feedback, to be carefully considered for re-parameterizing and re-dimensioning the network before this is actually deployed.
- **Quality evaluation and control:** QoE is the most appealing and ultimate way to evaluate the performance of any offered service, mechanism, or algorithm. By keeping track of the actually offered QoE, the provider becomes able to control and optimize the quality of the offered services to the user. Guaranteed QoE scores, both advertised by the providers and equivalently experienced by the users, is what makes one network provider more competent over another.
- **Troubleshooting:** Network "health" problems such as bottlenecks and local failures may be identified via QoE-based alarms (e.g., based on user-centric KQIs), thus causing corrective mechanisms to be immediately launched inside the network. Such problems may be either predicted, a priori, and then proactively resolved or

they may be identified, a posteriori, and reactively corrected, as long as feasible solutions are available.

- **Decision-making:** QoE may be incorporated in any network decision-making mechanisms, such as mobility management, radio resource scheduling, power control, rate adaptation, etc. New “QoE-driven” / “QoE-aware” inspired algorithms could help the network operate in a more effective way or in a fairer way, by using the user’s perceived quality as the ultimate criterion of decision making. In parallel, the economic impact of these mechanisms can be evaluated and considered by network operators during the decision-making process, optimizing the system from a combined user- and network-centric perspective.
- **CEM:** Through QoE awareness, service providers or network operators may gain access to user-related data, such as profile information, type of usage, communication statistics, user mobility patterns, etc. This acquired information may not only assist in QoE-based network/application control, but also in the better management of the customers’ overall experience, e.g., in terms of charging and pricing, SLAs, subscription profiles, customer support, customer behavior forecasting, etc.
- **Handset and service performance benchmarking:** This refers to the possibility of evaluating and classifying hardware (e.g., mobile phones) and software (e.g., applications or services) based on their performance and quality experienced by the users.
- **Business planning:** Finally, QoE intelligence helps stakeholders drive their business operations, prioritize investments, build SLAs, and enable informed strategic business decisions.

Apart from mobile network operators, who have an interest in collecting QoE awareness, as explained before, other stakeholders who may find interest are shown in Table 1.

Table 1: Stakeholders with QoE intelligence interests.

QoE stakeholders	
Network operators	Service providers
Network designers	Customer support
Marketing teams	Sales support
Equipment manufacturers	User Experience (UX) designers
Infrastructure planners	Application developers
Product strategists	SLA negotiators

Depending on each stakeholder’s interests and incentives, the target of acquiring QoE intelligence may differ. For instance, some network operators may focus on how QoE can help decrease customer churn, others may explore ways in which QoE intelligence enables a more efficient network resource usage, while others may be more interested in maximizing the average QoE of all subscribers, or in achieving QoE fairness among them [18]. Based on each stakeholder’s interests, QoE will be interpreted, monitored, and managed in a different way, depending also on the parameters that this stakeholder can control. As an example, equipment manufacturers may focus on how hardware decisions affect the user experience, network operators will control network-layer parameters, while service providers will work on improving the impact of application-layer parameters on QoE. Nevertheless, if different stakeholders view beyond their local optimum, mediations or collaborations of mutual interests may emerge, as will be further argued in this thesis.

2.4 The relationship between QoS and QoE

As discussed previously, QoS is not considered sufficient for the thorough characterization of a product or service as opposed to the most appealing QoE notion. The reasons to differentiate between QoS and QoE and to adopt QoE as the most suitable criterion for quality evaluation are twofold.

First of all, QoS handles purely technical aspects regarding a service and does not incorporate any kind of human-related quality-affecting factors. This means that the same QoS level might not guarantee the same QoE level for two different users. Apart from the system's technical characteristics, other factors such as the context of use, the user-specific characteristics such as users' experiences and expectations, the delivered content and the pricing of a service make a significant impact on the finally perceived QoE as well.

The second reason for this differentiation is that, QoS does not reflect the impact that the technical factors have on the user's quality perception, since there is no straightforward connection defined. This implies that, for instance, the constant amelioration of one technical parameter does not linearly and infinitely improve the user's QoE. Based on this gap between QoS and QoE, some formulas have emerged that attempt to map QoS parameters to the overall QoE value. Two different approaches have dominated in the literature: the perception-centric and the stimulus-centric one.

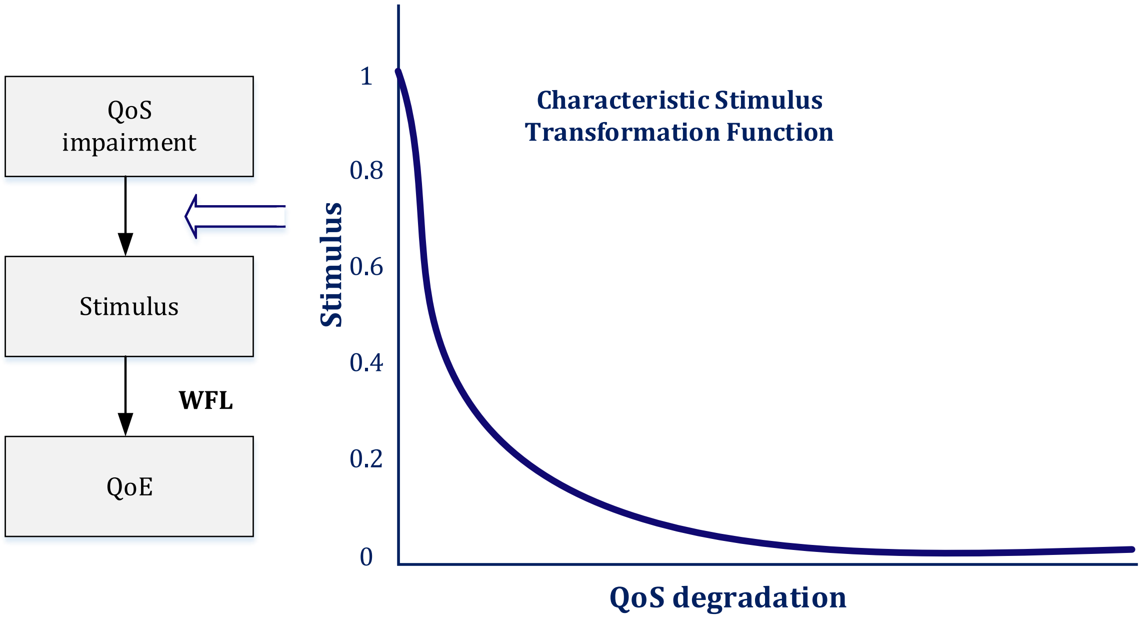


Figure 5: QoS-stimulus-QoE-perception chain.

The stimulus-centric approach is based on the "WQL hypothesis" inspired by the so-called "Weber-Fechner Law (WFL)", which describes the effect of a physical stimulus on the human perception according to the principles of Psychophysics [19]. This law claims that the relationship between stimulus and perception is *logarithmic*, which drives the conclusion that in order for a stimulus' change to be reliably detected by an observer, this has to differ from its original value by a constant fraction. From this law, the notion of "just noticeable differences" emerges, which describes the smallest detectable difference between two sequential levels of a particular stimulus.

Regarding the perception-centric QoS-QoE mapping, the so-called "IQX hypothesis" (i.e., Exponential Interdependency of QoE and QoS) has been proposed in [20]. According to this famous approach, the relationship between the QoE and one QoS degrading parameter is negative *exponential* and the change of QoE actually depends on the current level of QoE. The IQX hypothesis can be mapped to the WQL hypothesis, if a transformation function is considered that maps the QoS degrading parameter to stimulus values, as presented in Figure 5. Then, the stimulus may be mapped to perception (i.e., QoE) using the WQL hypothesis.

Figure 6 exhibits the IQX hypothesis. We observe three different regions of QoE evolution, split by the thresholds x_1 and x_2:

- Region 1 (constant optimal QoE): This region implies that minimal disturbance in QoS does not translate in QoE reduction at all. For instance, small delays and delay variations may be eliminated by a jitter buffer, without the user noticing the additional delay.
- Region 2 (sinking QoE): When the disturbance exceeds a certain threshold x_1, it is no longer transparent to the user. Consequently, the QoE starts to sink. It is interesting that the negative gradient of QoE diminishes as QoE values get lower. Intuitively, this means that a user can be very sensitive to a certain QoS disturbance while experiencing a high-quality service, but the exact same disturbance can go unnoticed when QoE is already low.
- Region 3 (unacceptable QoE): As soon as the disturbance reaches another threshold, x_2, the value of QoE becomes indifferent, implying that the user has possibly given up using the service, or the service has stopped working due to technical constraints such as timeouts.

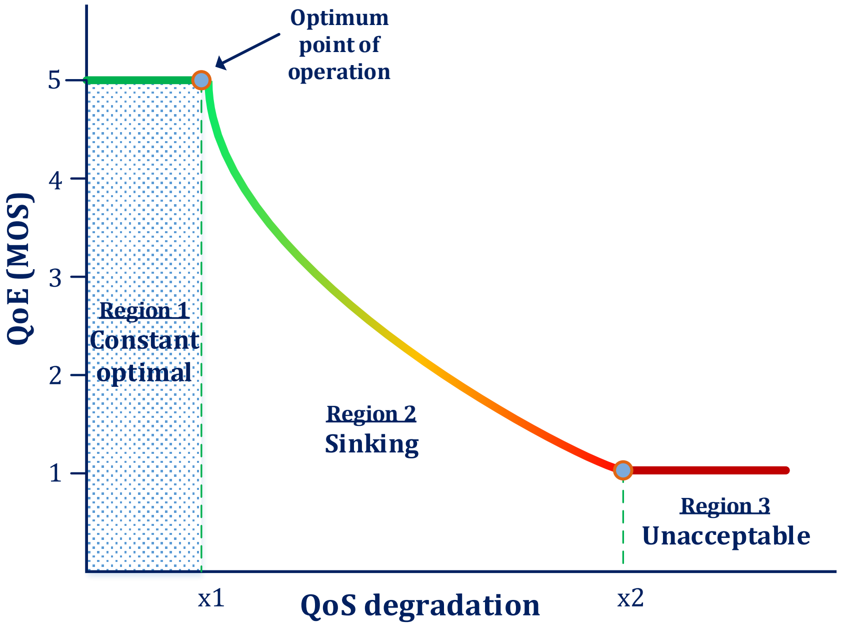


Figure 6: The IQX hypothesis.

Some of the QoS parameters that have been successfully mapped to QoE using the above laws are the: a) packet loss ratio, b) type-p reordered ratio (i.e., the percentage of packets in the received stream that are reordered, which quantifies the jitter), c) weighted session time for web browsing / page load time, d) total setup time of a wireless connection, e) delivery bandwidth, f) image quality perception as a function of blur, and g) download time perception as a function of response time.

The QoS to QoE mapping may also be of power-law type in some instances (following the Steven's power law [21]), such as for the session volume as a function of bandwidth, and the video perception as a function of jitter.

A summary of the above laws governing the relationship between QoS and QoE is presented in Table 2.

Table 2: QoS-to-QoE laws.

Law	Trend	Relation	Form
Steven's Power Law	Stimulus-centric	$QoE = \alpha QoS^{\beta}$	Power
WFL	Stimulus-centric	$QoE = \alpha \ln(QoS)$	Logarithmic
IQX	Perception-centric	$QoE = ae^{-\beta QoS} + \gamma$	Exponential

In the next subsections, we describe the key challenges associated to QoE, namely technical, economic/business and legal issues that need to be addressed before QoE becomes the de facto way of quality provisioning.

2.5 Key challenges in the QoE domain

The research area of QoE is multi-dimensional, spanning across many scientific domains, even different from the IT and Telecommunications sectors, such as the area of Psychophysics, Psychology, Sociology, Decision theory, Microeconomics, Business, etc. Some of the most important research issues that are associated with the understanding and the provisioning of QoE in a network have been identified as the following:

- A QoE management procedure needs to be standardized in networks with QoE-awareness. This procedure, as will be thoroughly explained in next chapters, should implement some kind of QoE estimation (a.k.a. QoE modeling), QoE monitoring and, ultimately, QoE control. The awareness of QoE is an important asset of network operators just by itself. If, however, it is further exploited, it becomes a powerful tool for optimizing the network and delivering service management in a QoE-centric way.
- The mapping of QoS to QoE is another important area that may be found useful for a fast adaptation of QoE into the networks, i.e., by exploiting the current QoS mechanisms and transforming them to QoE-aware mechanisms. In this research field, current approaches (e.g., IQX) map a single QoS parameter to QoE values, but there is still research needed so that multiple QoS parameters can be mapped at once in a single QoE value.
- Another area of particular interest is the exploitation of QoE provisioning for resource and energy savings. Based on human perception principles described by Psychophysics' laws such as the WFL, Steven's power law and IQX hypothesis presented before, this is possible, and could provide valuable insights for QoE-based resource management techniques. Hence, the impact of human perception and cognition may be exploited for designing smarter network mechanisms that optimize both for QoE and network resources' utilization (e.g., [22]).
- Another main challenge that needs to be addressed in mobile cellular networks is the end-to-end provisioning of quality, irrespective of the multi-vendor, multi-operator, multi-network environments where the packets traverse, in parallel with the diverse transport technologies and differentiated quality assurance requirements that the providers face. Efficient signaling protocols may be proposed that overcome this issue, or, novel solutions based on softwarization and virtualization may be designed.

- What is more, a service may be generated by a third party, e.g., a service provider, while the mobile network infrastructure is used just as a communication pipe for this service. Currently, the underlying infrastructure is a black box as far as service providers are concerned; however, it might be profitable to raise this isolation between the two stakeholders, and propose new technical and business schemes, where they join forces towards a higher user experience.
- Respecting the users' privacy is another crucial challenge in networks with QoE support. QoE awareness requires some kind of behavioral monitoring and user-specific data collection; however, trust and security should be somehow guaranteed.
- The design of new business models, SLAs and subscription profiles are also required, that take into account the special QoE-based characteristics.

2.5.1 Technical issues

Below, further technical challenges and constraints in the QoE provisioning process in mobile cellular networks are presented. The Long Term Evolution (LTE) / LTE-Advanced (LTE-A) network is used as a reference in this discussion. (Note: this network type is considered throughout this thesis as well, including the simulations conducted.)

Heterogeneity

One of the most important characteristics of LTE-A networks is their inherent heterogeneity. With this term we refer to the dynamic and constantly increasing emergence of geographically distributed and overlapping smaller cells (e.g., femtocells and picocells), a characteristic that plays a drastic role in the offered QoE. On the one hand, this heterogeneity better supports the ever-increasing user traffic requirements and pushes towards an increase of the user QoE, since users are served by closer base stations (higher throughput, less energy, less delays). On the other hand, this phenomenon inevitably imposes higher interference and severer competition over the, anyway, scarce spectrum resources, thus pushing towards a decrease of user QoE. Consequently, there is a delicate balance to be considered in modern mobile networks regarding QoE control, not only during the network's planning phase but also while the network is operational.

In Figure 7, a typical LTE-A network is presented (access and core), where heterogeneity is evident. By taking advantage of its impact, the opportunity to incorporate the network heterogeneity into the QoE provisioning chain emerges. More specifically, it may be exploited both as a source of input for QoE modeling and as a technique for quality control. For instance, the corrective action of initiating a vertical handover from the macro-cell to a small cell may significantly improve the perceived quality. (The LTE notation is adopted in Figure 7, i.e., evolved eNB - eNB is the LTE base station, Home eNB – HeNB is the femto base station, and User Equipment – UE is the mobile user).

QoE monitoring approach

An important challenge in QoE monitoring is the collection of QoE-related input information from the appropriate network nodes and devices. The dilemma in this problem is whether input will be collected centrally by the various distributed network nodes (network-centric approach) or using agents installed locally at the user devices (agent-based approach). Agent-based approaches have the advantage of being able to capture also more subjective QoE influence factors, such as the context of use. Moreover, if these agents are not silent probes but also require some user feedback, they are able to capture the unique human factor's characteristics. Hence, they are able to provide a clearer understanding of the perceived quality. Moreover, agent-based solutions have the benefit of capturing any potential problems due to the access (wireless) part of the end-to-end communication path as well as problems occurring inside the handsets themselves.

On the other hand, these approaches that rely on collecting data at the network edges have the disadvantage of not capturing the problems occurring inside the core network, and thus, do not provide diagnostic information. Apart from that, a major disadvantage is the dependability on the manufacturers' willingness to implement such solutions inside the mobile handsets, as well as compatibility issues. Furthermore, it is expected that such solutions are not scalable, and they significantly overload the network with QoE-specific signaling and, therefore, also drain the devices' battery faster. Finally, implementing monitoring solutions inside the user terminals and transferring the monitored information through the network raises privacy and security issues that need to be considered in terms of the users' acceptability of such solutions [23]. Due to the co-existence of equally important advantages and disadvantages of the agent-based solutions, it appears to be a good option to combine both approaches in a carefully distributed way.

User versus provider

We may identify two contradicting forces in the QoE provisioning process: On the one side, there are the network operators and service providers, who want to maximize their revenue, and, on the other side, there are the users, who desire the maximum experienced quality, and in fact, at the lowest possible cost. Nevertheless, in order to increase their revenue, providers have to sometimes reduce the offered quality of their services, through for instance the installing of less infrastructure devices or less powerful nodes, or due to buying and disposing less spectrum resources to their subscribers, etc. However, since a reduced quality will inevitably produce customer churn, in the end, their revenues will be significantly decreased or even the company reputation will be affected. Consequently, it becomes crucial that the golden section between these two contradictory forces is found, i.e., that the operators offer the maximum possible quality at the least possible charge for the users, while achieving the maximum possible revenues. Hence, it is essential that new, QoE-based business plans and charging schemes emerge.

Network diversity

Providing high QoE to a mobile subscriber does not necessarily depend only on the technical efforts (namely hardware equipment, software functionalities, network management, etc.) of the network operator to which this subscriber belongs. The finally perceived QoE of the user will be formed during the complete end-to-end path, starting from the source of data (which might even not be a node in the possession of the network operator) and ending at the user terminal. This means, that there will be cases when this communication passes through different operators or vendors, through different mobile technologies, through different networks, or even through different countries or continents. This raises two issues. First, in order to ensure proper QoE at the user, collaborations and agreements among different parties (e.g., operators) are required, which will sufficiently define the obligations of each party. Moreover, agreements or collaborations between network operators and service providers are becoming essential, in order to provide QoE in the optimal way by joining forces of both stakeholders. These agreements require some kind of signaling, especially at interconnection interfaces, and moreover some diagnosis tools to be able to identify problem roots along such diverse communication paths. Second, security and privacy issues are raised, because user-sensitive information that is used for better QoE management is traversed through different stakeholders.

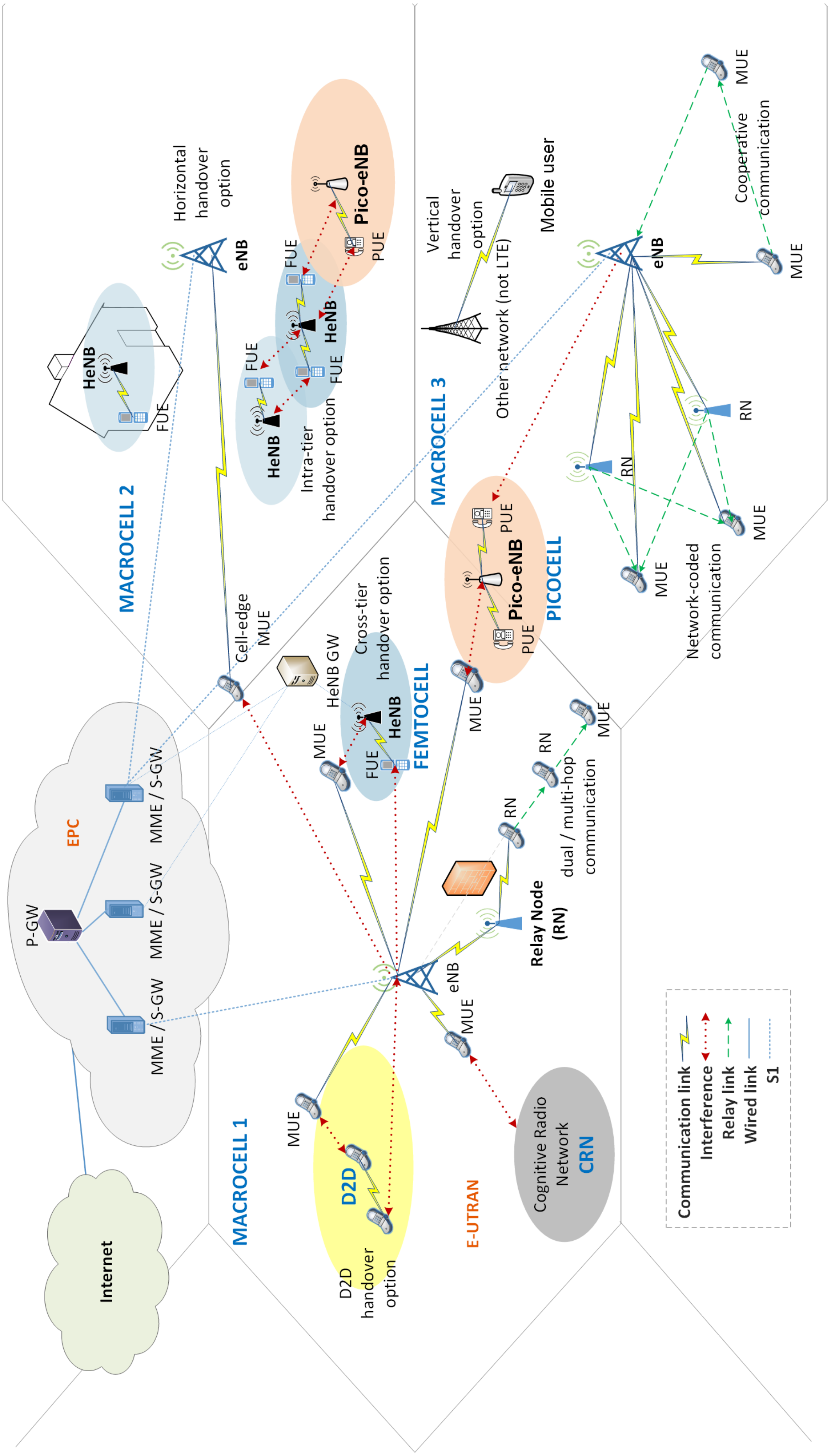


Figure 7: Heterogeneity in LTE-A.

Scalability and complexity

The end-to-end QoE support requires feedback mechanisms in two directions: First, the collected QoE-input information by either agents or network probes needs to be transferred to a central QoE modeling and management entity, and second, any control decisions of this entity need to be disseminated back to the network affected nodes. Consequently, as the number of end-devices and core network nodes increase, the QoE monitoring procedure suffers from scalability problems. Moreover, since QoE modeling and management decisions need to be performed per user's request (i.e., per user's flow) to account for the unique session characteristics, and since the number of users in the network may be large, complexity issues are raised regarding the network's optimization decisions. This is further deteriorated due to the large number of input factors that should be taken into consideration by a reliable QoE estimation model.

Encryption

Contradicting interests emerge between operators and service/content providers (such as Facebook, Google, etc.). The latter design their new technologies and services with security (i.e., encryption/content labeling) in mind. Encryption, however, might become an "enemy" for QoE-awareness and in turn, for flexible QoE delivery. Unencrypted data, on the contrary, can be a powerful tool for the operators as the source of information to ensure, enhance or adjust QoE, or to provide service differentiation. To achieve such a differentiation, the operator needs to know the application type, its current state, etc. As an example, by having access to the buffer state information of a video playout, the operator can prioritize the limited resources available at a specific time and location in order to maximize a certain utility function, e.g., maximize the number of satisfied users.

Energy consumption

Finally, the required energy consumption for supporting QoE in a network seems to be a very crucial issue, due to the involvement of new network entities, the increase in the processing tasks of users and nodes (especially the monitoring of QoE influence factors from e.g., the surrounding environment), the extra signaling imposed, etc. Hence, energy efficient solutions targeting at minimizing the consumed power required for QoE-awareness collection and quality provisioning should be considered in future research.

2.5.2 Economic and business issues

Apart from the technical factors influencing QoE, also pricing/charging greatly impacts the user's opinion. The issue of QoE charging is studied in [24], where it is described in terms of a fix-point problem. In addition to the delivered service quality, it considers user context and expectations, as well as economic feedback from the subscribers.

The basic models studied are given in Figure 8 and Figure 9. The first one refers to QoS-based charging. In this model, several QoS parameters are measured or estimated and then used as input to the charging mechanism, which determines the corresponding price based on predefined tariff functions (Figure 8). This produces a feedback, since the chosen tariff influences the customer demand, which in turn shapes the network load, and finally, the delivered service quality.

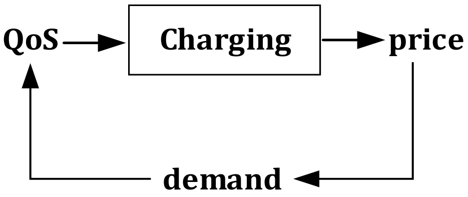


Figure 8: Charging for QoS model [19].

This model can be described as follows (where p is the price, d is the demand and q is the QoS):

$$price\ function: \quad p = p(q) \tag{2-1}$$

$$demand\ function: \quad d = d(p) \tag{2-2}$$

$$QoS\ function: \quad q = q(d) \tag{2-3}$$

The second model is an extended feedback model for QoE charging. Similar to the first one, the provided QoS along with the price affect the charging mechanism (Figure 9). The difference here is the fact that they serve as determinants for the QoE evaluation, which is considered as the essential input for the charging mechanism. That is, there is an additional feedback, which is the influence that the price has on the perceived QoE. For example, a user that pays for a service perceives a worse QoE than one that doesn't pay for it. A comparison of the studied model with user trials on QoE for Video on Demand (VoD) has shown that the model can be considered as a representative for a broad set of relevant scenarios.

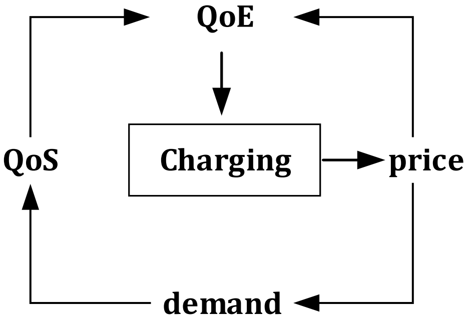


Figure 9: Charging for QoE model [19].

The main logic behind this model may be briefly described using the following formulas:

$$demand\ function: \quad d = d(p) \tag{2-4}$$

$$QoS\ function: \quad q = q(d) \tag{2-5}$$

$$price\ function: \quad p = p(x) \tag{2-6}$$

$$QoE\ function: \quad x = x(q, p) \tag{2-7}$$

which may be further expressed as:

$$QoE\ function: \quad x = x_Q(q) * x_E(p) \tag{2-8}$$

where x is the QoE, x_Q is the "quality function" and x_E is the "expectation function".

Finally, one important tool in the problem of understanding and quantifying the QoE that is worth mentioning comes from the microeconomic utility theory. This theory helps describe the preferences of a user through a "utility function" [25]. This function is denoted as $u_i(x)$, for user i and refers to the consumption of the resource "x". In this sense, if $u_i(x) \leq u_i(y)$, this implies that the user prefers y over x. Three typical examples of utility functions are presented in Figure 10: the linear, elastic and non-elastic functions. The linear utility function describes a scenario where constantly increasing a resource or a metric, such as the capacity, linearly and infinitely increases the utility as well. Being more realistic, the elastic traffic describes a concave increase in the user's utility while increasing his capacity, increasing faster in the beginning and slower while capacity is already large. Finally, the non-elastic traffic refers to an "ON-OFF" scenario, where the user has a perceivable and fixed utility only after a certain threshold.

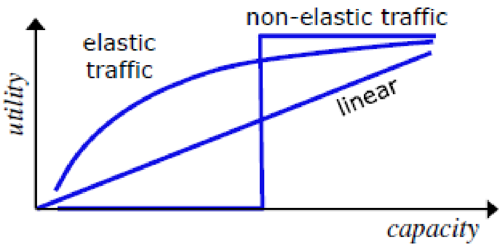


Figure 10: Examples of utility functions [25].

If multiple users in a cell are taken into consideration, say N, then the overall social welfare is defined by the weighted sum of logarithmic utility functions of the form $u_i(x_i) = w_i \log x_i$, as follows [25]:

$$U(x_1, x_2, \ldots, x_N) = \sum_i w_i \log x_i \tag{2-9}$$

Apart from finding a proper scheme for charging for QoE, the question of how QoE profits will be distributed among the involved parties is still open. This is valid for instance for the case where multiple network operators are involved in the service provisioning chain. Another example is for the way of distributing the QoE profits to both service providers (e.g., a VoD service provider) and network operators or Internet Service Providers (ISPs) [26].

Marketing is another issue that should be taken into consideration, namely how QoE will be advertised to the market as an extra service that users will normally have to buy, and motivate them to do so. Competition among different providers may be enforced through advertisements that claim that one provider offers higher QoE to its customers, similarly to how advertising is performed today based on QoS criteria such as download speeds.

2.5.3 Legal issues

In [26], several legal challenges linked with QoE support in the networks are described. “Quality” may be considered as a public good, which should be available to everyone as long as this is feasible. In other words, it may be considered incorrect to deliberately prevent users from getting a high QoE when they need it, because they haven’t paid for it, even though this would be technically possible. This would cause discrimination among users. Hence, this is one of the legal issues that need to be investigated, referred to as the challenge of net(work) neutrality. From the operator’s perspective, net neutrality regulations may not leave enough space for innovation and investment in the networks in terms of QoE. Furthermore, even though the recently voted net neutrality regulatory framework [27] has been welcomed by most service/content providers as a way to allow flawless access to their services, it is not a black or white issue. For instance, the dynamic allocation of “fast lanes” may no longer be allowed by the network providers to pass, say, Netflix content to premium users or to do any other type of service differentiation.

Another legal issue that needs further research is the problem of “double selling”. This refers to the decision about whether the QoE will be sold as an add-on service to existing network connections or as an indispensable element of the offered services.

Moreover, SLAs need to be revisited. SLAs are a type of contract between the provider of a service and the client and describe the service type and quality that the customer should expect to receive. If the requirements described in the SLA are not respected by the provider, i.e., violated, then legal issues arise. Presently, such requirements are described using QoS terms, such as the maximum tolerable delay and packet loss, etc. However, as explained before, QoS values above a threshold do not directly imply a proportionately satisfactory QoE. Consequently, new types of SLAs or ELAs may be considered that define the required quality using QoE terminology. The great challenge of this new approach is to find a way to clearly define the various QoE classes and to be able to measure this, so that customers do not arbitrarily complain about their perceived quality. Besides, it may be difficult for the customers to distinguish from e.g., “very poor” and “poor” quality; hence a common “vocabulary” and understanding between users and providers needs to be thoroughly defined and described.

Similarly, SLAs that are signed between a network operator and various service/content providers need to be revisited. In this case, it is further required to devise indisputable methods of measuring the QoE at the various interconnection points, of checking it against the ELAs’ QoE requirements, and of finding which side of an interconnection is legally responsible in case of QoE deficiencies.

Last but not least, privacy and fidelity issues arise when providing QoE support into the network. QoE-related information has to potentially pass through different provider domains, different countries or even continents, through both the wireless and wired medium. Offering an end-to-end QoE would require, though, the transfer of such sensitive information throughout this whole path, raising issues about whether information about e.g., user profiles and demographics, user statistics, usage patterns, etc. are confidentially transferred. Moreover, it needs to be guaranteed that collected information about the users will not be used for any other reason rather than QoE provisioning and customer support in general, and moreover, that this information will not be provided to third parties. Such privacy considerations may make the users skeptical towards accepting an add-on QoE service, not to mention paying for it.

A similar aspect that may raise legal privacy issues is the potential requirement for various providers to cooperate, especially at the points of interconnections. Hence, information about each other’s network status and configuration may need to be shared. This is another legal challenge that needs to be settled before QoE provisioning becomes an integral part of the communication networks in the future.

Having provided the general background regarding QoE, in the next chapter we present the requirements towards collecting QoE intelligence and, in sequence, managing a mobile cellular network in a QoE-aware manner.

3. QoE MANAGEMENT IN MOBILE CELLULAR NETWORKS

Telecom operators are facing the need for a radical shift from technical quality requirements to customer experience guarantees. This trend has emerged due to the constantly increasing number of mobile devices and applications and the explosion of the overall traffic demand, forming a new era: that of “the rise of the consumer”. QoE is the most dominant term coined in order to quantify, manage and improve the experienced user quality. However, QoE has been more of an afterthought for network providers, and, thus, numerous research questions need to be answered prior to a shift from conventional network-centric paradigms to more user-centric approaches. To this end, it is crucial to provide insights on the issue of network-level QoE management, identifying the open issues and prerequisites towards acquiring QoE awareness and enabling QoE support in mobile cellular networks.

In this chapter, a conceptual framework for achieving end-to-end QoE provisioning is proposed, and described in detail in terms of its design, its constituents and their interactions, as well as the key implementation challenges. An evaluation study serves as a proof of concept for this framework, as well as demonstrates the potential benefits of implementing such a quality management scheme on top of current or future generations of mobile cellular networks.

3.1 Introduction

As also discussed in Chapter 2, QoE is defined by ITU-T as “*the overall acceptability of an application or service, as perceived subjectively by the end-user*”. Otherwise put, it describes the degree of the end-user’s “*delight or annoyance”* when using a product or service [6]. Inherently, QoE is a very broad and generic concept, and, as such, it incorporates any conscious or unconscious aspects that affect the overall user satisfaction.

This generic notion of QoE has opened up research to a variety of systems and application domains. In this chapter, we narrow down the scope to the telecommunications domain, where QoE intelligence is of crucial importance, not only to the end-consumers but also to any stakeholders involved in the service provisioning chain. In telecommunication networks, despite the catholic presence of inherently deployed QoS mechanisms, QoE has been an “afterthought”. No generation of telecommunication networks has been originally designed with QoE principles so far. Nevertheless, the system-centric view of QoS provisioning is no longer sufficient, and it needs to be replaced or complemented with more user-centric approaches [26]. Therefore, the shift from QoS- to QoE-centric networks remains an emerging, open challenge.

Towards this direction, new architectures have been proposed regarding the collection and exploitation of QoE-related information. For instance, a block diagram for the QoE management of Next Generation Networks (NGNs) is proposed in [13], where adaptations to the NGN-specific Network Attachment Control and Resource and Admission Control Functions are described. Furthermore, a novel architecture for QoE support in LTE systems requiring new, proprietary interfaces is described in [23]. Other works focus on specific services, such as the CEM system for IPTV described in [28]. Similarly to the aforementioned examples, the majority of current works proposes solutions tailored to concrete systems or services. In parallel, standardization activities mainly handle the issue of QoE estimation, a.k.a. “QoE modeling” [29], leaving the end-to-end QoE provisioning realization out of discussion. Motivated by this observation, the current study proposes the required steps for enabling QoE-based management in the environment of *mobile cellular networks*. Our contribution lies in identifying the design challenges and requirements towards QoE provisioning, namely a) gaining QoE awareness, and b) using this awareness to enable effective QoE-centric decisions on top of mobile cellular networks (e.g., GSM, UMTS, LTE/LTE-A). In this way, a better understanding of the challenging topic of QoE in mobile cellular networks is gained.

The remainder of this chapter is organized as follows. We first provide a comprehensive composition of the QoE notion from a mobile cellular network's perspective by describing, in an end-to-end manner, the most important quality influence factors. Following this, we present a conceptual framework towards QoE support, described in terms of functionalities, interactions and design challenges. Afterwards, realization issues for the tight integration of the proposed QoE provisioning framework in mobile cellular networks are identified, and evaluation results are presented, using the LTE network as a case study.

3.2 Breaking down QoE provisioning in a mobile cellular network

QoE is a broad concept, embracing influence factors from different domains and disciplines. We adopt the approach of [6] and categorize those factors into three major pillars, namely *System* (here, *Network*), *Human* and *Context*, which compound together, formulate the overall user QoE. Moving one step further, we group the most dominant factors per pillar, and illustrate how QoE opinions are progressively formed during a communication session (i.e., how these pillars are connected) (Figure 11).

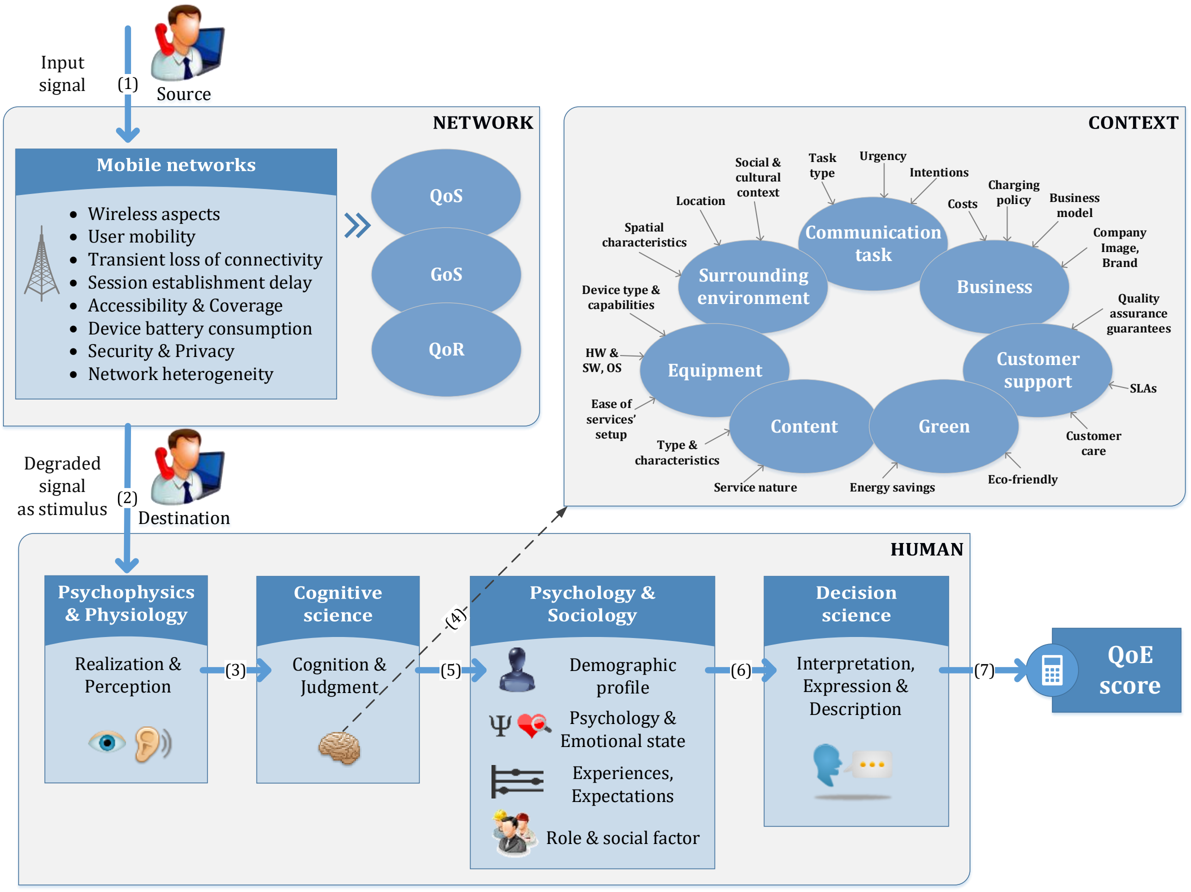


Figure 11: The three QoE influence pillars.

The *Network* pillar consists of any end-to-end quality affecting parameters, as these are described by the QoS, GoS and QoR terms [26]. It embraces technical characteristics of the traversed network, equipment specifications, application characteristics, etc. This pillar is strongly connected with network-specific factors, which are particularly important and decisive for the operator (see the “Network” box in Figure 11). In the case of mobile cellular networks, the most challenging and less investigated factor is their inherent heterogeneity, referring to the dynamic emergence of geographically distributed and overlapping smaller cells. As also discussed in Section 2.5.1, on the one hand, this heterogeneity helps support higher traffic requirements, pushing towards an increase of the user QoE, while, on the other hand, it imposes higher interference and severer competition over the bandwidth, pushing simultaneously towards a QoE decrease.

Moving on to the *Human* pillar, we describe it as the superset of four subcategories, where each one comprises a unique scientific area that influences the overall user’s quality impression. Initially, the area of Psychophysics quantifies the relationship between a physical stimulus (e.g., sound/image) and the resulting perception to the human sensory system. Then, the Cognitive Science studies the human mind and how this works in terms of interpretation, reasoning, judgment, information processing, etc. Psychology and Sociology help understand the human character and behavior both as a unity and part of the society, which uniquely affect the user’s understanding of quality. Finally, Decision Theory studies the rationality and optimality in decision making.

Finally, the *Context* includes any kind of background information that consciously or unconsciously affects the user’s judgment. For instance, QoE is influenced by the spatiotemporal environment where the service is provisioned (open-air crowded place vs. quiet office); the equipment under use (mobile phone vs. tablet); the service and content type (audio/video/text/graphics); the content characteristics (head-and-shoulder video vs. football game); the communication task (public safety vs. leisure browsing); and other contextual information related to business or financial aspects (e.g., charging policy, marketing, brand effect).

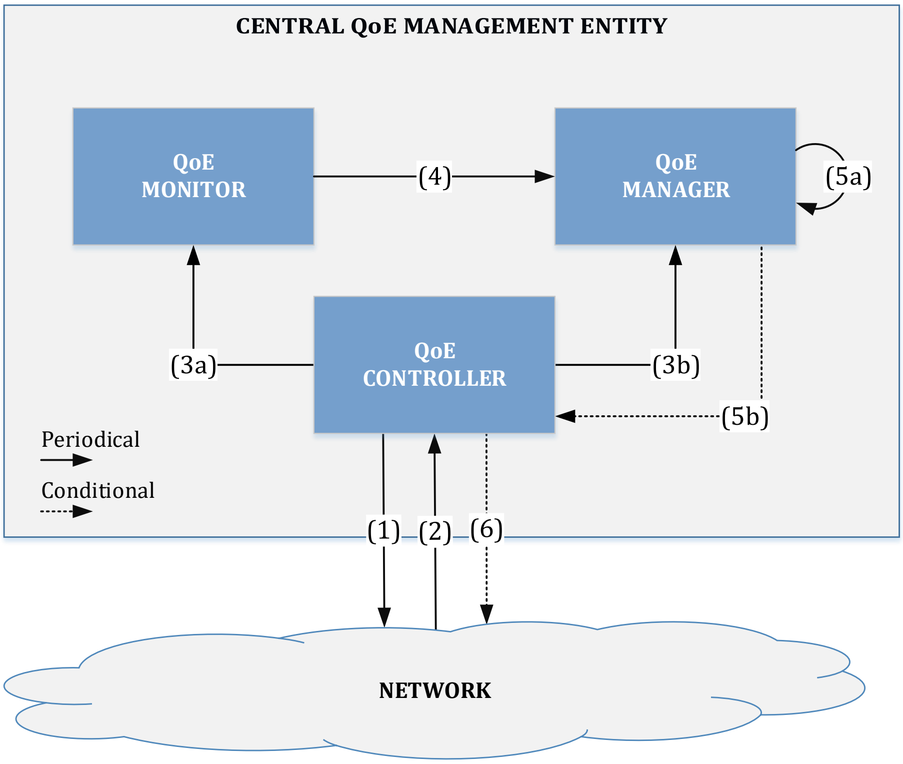
Depicted in Figure 11 is the progressive formulation of a user’s QoE during a communication session, presented in chronological order (steps (1)-(7)). One source-generated signal is entering the network (1), and its distorted version reaches its destination (2), where it is perceived by the target user as a visual/audio stimulus (3). This stimulus is internally represented into the human brain, processed as information content and in terms of quality (5). This quality judgment is significantly affected by numerous external factors, which all together constitute the context of this communication scenario ((4)-dashed). Following this, the quality impression is further influenced by unique characteristics of the human subject (e.g., demographic profile, current psychology, expectations) (6). Finally, the formed quality perception is expressed as a QoE score in a given scale (7), such as the 5-point Mean Opinion Score (MOS).

3.3 A conceptual framework towards QoE management in mobile cellular networks

In this section, we propose a framework that enables QoE management in mobile cellular networks. To this end, we identify its required building blocks, their inner functionalities and in-between interactions.

The structure of the quality provisioning chain and the required interactions ((1)-(6)) are presented in Figure 12. In the core of the proposed framework is a *central QoE management entity,* which is implemented at an administrative location of the operator’s network, on top of the mobile cellular network depicted in Figure 12 by the “Network” cloud. This entity is able to collect QoE-related input and apply QoE-driven network management decisions. It consists of three main building blocks, namely the *QoE-Controller*, *QoE-Monitor* and *QoE-Manager*.

The **QoE-Controller** plays the role of an interface between the central entity and the underlying network, synchronizing communication exchange in both directions. It is in charge of configuring the data acquisition process, by requesting and collecting feedback from appropriate data sources (e.g., some QoS indicators), as will be further analyzed below (interactions (1) and (2) in Figure 12, respectively). The QoE-Controller also decides and imposes the periodicity of this process (through (1)), namely it controls how often QoE input should be generated/gathered, and consequently how often QoE will be assessed. Having collected the required data, this component provides input of interest both to the QoE-Monitor and the QoE-Manager ((3a) and (3b), respectively). More specifically, it provides QoE-input data on a per flow basis to the former, and information regarding the current network state to the latter (e.g., network topology, resources' availability, etc.). Finally, the QoE-Controller applies any QoE-aware control decisions back to the network, during the final step of the QoE management loop (6).



(1) Instructions controlling the QoE-input data generation are sent to the network
(2) Input data from all data sources are collected by the QoE-Controller
(3a) Processed QoE-data per flow are sent to the QoE-Monitor
(3b) Information regarding the current network state is sent to the QoE-Manager
(4) Estimated QoE scores are reported to the QoE-Manager per flow
(5a) Customer Experience Management procedures are performed
(5b) Corrective actions are triggered, if required
(6) The QoE-Controller actualizes these corrective actions

Figure 12: The proposed QoE management framework.

Second, the **QoE-Monitor** is responsible for estimating the QoE per flow, i.e., per user's session, and for reporting this to the QoE-Manager (4). Using network-derived input available through the QoE-Controller, the QoE-Monitor initially performs traffic classification to deduce the type of traffic of the considered flow. This procedure is feasible using statistical analysis, e.g., [30]. Inside the QoE-Monitor, already built-in QoE assessment functions, referred to as "QoE models" (i.e., formulas for quantifying a service's QoE) are available, different per traffic/service type (e.g., video/voice/data).

Depending on the identified traffic type, the proper QoE estimation model is selected by the QoE-Monitor, followed by an estimation of the QoE. It needs to be noted, that all available QoE models are integrated offline into the QoE-Monitor by the operators, namely during the design phase of the central QoE management entity and prior to its real-time operation, which makes the original selection of QoE models very crucial.

The last component of the proposed framework is the **QoE-Manager**, responsible for conducting any type of CEM (5a) or QoE-aware network management (5b). It uses a) input from the QoE-Controller regarding the current network state, b) estimated QoE scores through the QoE-Monitor, and c) operator-specific information, such as network policies or SLAs/ELAs, as a way to decide and dictate the necessary measures that need to be imposed to the network for solving quality problems at hand. Decisions are taken per flow or catholically, respecting user policies (e.g., subscription profile, charging information, etc.) and current network constraints (e.g., availability in resources). Any QoE-triggered decisions are clearly system-specific, in the sense that their actualization depends on the underlying network. The adaptation/control actions that realize these decisions are applied to the network through the QoE-Controller (6).

Next, we analyze key design issues per building block, starting by the QoE-Monitor, which performs the key process of estimating the QoE per flow.

3.3.1 The QoE-Monitor

The main challenge in the implementation of the QoE-Monitor is the thoughtful selection of QoE estimation models, different per traffic/service type, to be integrated offline (a priori) into this block. QoE models imitate the *Human* processes that occur inside a specific *Context* each time, given the *Network* characteristics at hand. Formally defined, a QoE model is "*a procedure that aims to model the relationship between different measurable QoE influence factors and quantifiable QoE dimensions for a given service scenario*" [31]. Consequently, the main purpose of this block is to reliably estimate QoE, as if this assessment was done by humans.

A plethora of QoE models can be found in standardization bodies' recommendations and in the literature. For instance, ITU standardization activities for IPTV QoE assessment can be found in [32], while a detailed taxonomy of objective speech quality models can be found in [33]. For VoIP services, the "E-model" is commonly used, mainly due to its valuable characteristic of providing distinct formulas for quantifying the impact of packet delays and loss rates on QoE (the "Delay impairment factor" and "Equipment impairment factor", respectively). For web browsing services, QoE is strongly affected by the web pages' response/loading time, while for file download services by the effective data rate. The experienced quality in real-time video applications (e.g., IPTV) is mainly influenced by the packet loss rate and burstiness, frame-rate, bitrate and content type. Finally, the QoE for lossless video streaming services (e.g., YouTube) is significantly affected by the number and duration of stallings, as well as the video start-up delays.

QoE models are mainly classified based on their evaluation method [34]:

a) *Media-layer models* make use of transmitted and/or received signals. Based on the need or not for the original source signal to be used as input, they are further characterized as Full-Reference, Reduced-Reference or No-Reference.

b) *Packet-layer models* extract information from packet headers, while bitstream models use both headers and payload information.

c) *Parametric models* use specific network planning parameters and metrics, as well as terminal design parameters.

d) *QoS-to-QoE mapping models* are based on the non-linear dependencies between QoS parameters and QoE values.

More details about the QoE-Monitor, mainly in terms of QoE modeling, and specifically a more elaborate classification of QoE models and a description of QoE parametric models, are presented in Chapters 4 and 5, respectively.

3.3.2 The QoE-Controller

The QoE-Controller realizes the interface between the central QoE management entity and the underlying network, by enabling a bi-directional communication exchange. Specific design decisions need to be taken when designing this building block.

Regarding the communication direction from the network to the QoE-Controller (illustrated as (2) in Figure 12), the strategic selection of appropriate nodes used for the acquisition of QoE-related input is a challenging issue. Input can be collected by various distributed nodes located at the Core and Access Network (macro-/small-cell base stations, routers/servers/gateways) capturing service degradations, as well as by agents installed locally at end-devices, capturing more subjective QoE influence factors, such as context and human characteristics (Figure 13). Some guidelines on QoE/QoS data collection in 4G networks are given in [35]. In this work, the authors propose the integration of active probes within multiple network elements between the service provider's gateway and the access network, for measuring network QoS indicators (e.g., throughput, delay, jitter), transport KPIs (e.g., round-trip times) and application/service KQIs (e.g., video frame rate, blurriness).

The appropriate type of collected QoE-related input is another important issue. This input refers to any kind of raw network data, real-time measurements, statistical/historical information, or information at the operator's possession, obtainable through: a) active (intrusive) or passive (non-intrusive) probes on distributed network elements, b) embedded agents/sensors on user-devices that explicitly/silently collect usage data and statistics (e.g., monitor video playout buffers to predict stalling events), c) user-devices' applications that request user feedback, or d) any subscriber-related databases owned by the operator (Figure 13).

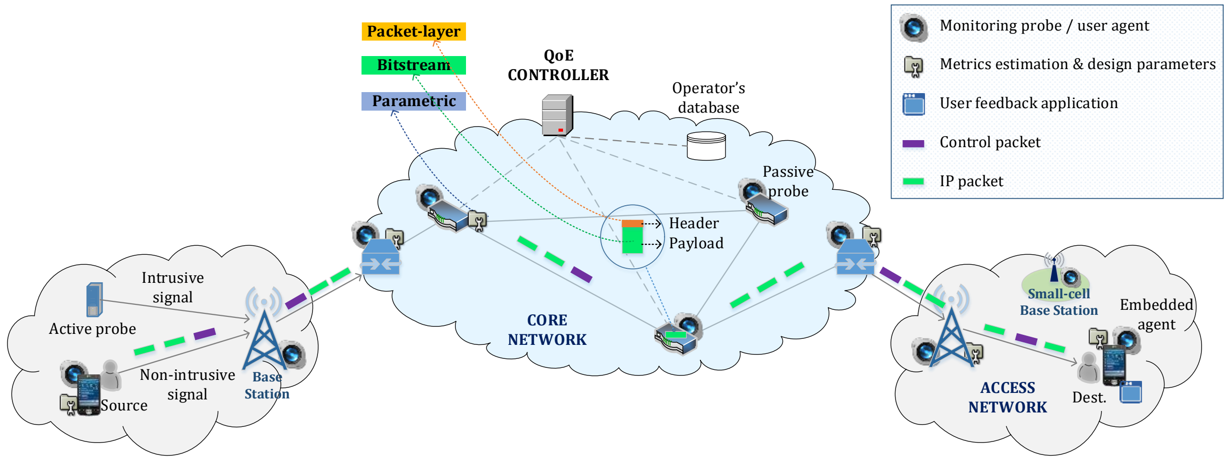


Figure 13: Illustration of the interactions between the QoE-Controller and the various QoE data sources.

The data acquisition process needs to be aligned with the QoE estimation models embedded inside the QoE-Monitor. Different input parameters are required per model, and therefore, the two phenomenally different procedures of the QoE-Controller and

QoE-Monitor have to be tuned offline. Therefore, the operator's first task is to select the appropriate QoE models, and then to fine-tune the data acquisition process accordingly (Figure 13). The collection of input may be based on packet-level information acquired through Deep Packet Inspection (DPI) techniques (applies to packet-layer/bitstream models) or by estimating communication-related metrics (parametric models). In the case that packet-layer models are used, the characteristics and configuration of endpoints should be known in advance, or be acquired using Real-Time Control Protocol-Extensive Reports (RTCP-XR). In addition, the data acquisition procedure needs to be tuned a priori with respect to the pool of decisions/actions embedded inside the QoE-Manager.

Regarding the communication direction from the QoE-Controller to the network (illustrated as (1) in Figure 12), we envision that the QoE-Controller is able to dynamically configure/administrate the data generation and the data collection periodicity, e.g., by switching ON/OFF some probes, based on the current network state. This periodicity needs to balance between the inevitable extra signaling overhead imposed in the network and the timeliness of the acquired data or, equivalently, the accuracy of QoE estimations.

A closer look into the QoE-Controller is given in Chapter 9, where the implementation of this component (as well as of the whole QoE management cycle) is put into the frames of the Software-Defined Networking (SDN) technology.

3.3.3 The QoE-Manager

Currently, the only opportunity for network providers to assess the offered QoE of their products or services is during the design phase, namely, prior to real-time operation. This may be accomplished by purchasing special equipment from third-party vendors, capable of performing measurements of voice/audio-visual quality through emulating the human perception. Operators may use such quality-measurement suites as a way of testing the performance of new services/devices, and thus, accelerate the time-to-market. This, however, is the only course of action currently feasible; on the contrary, the proposed framework opens up possibilities for real-time quality monitoring and smart network-centric QoE management based on the operator's actual customer portfolio and realistic communication conditions.

The first possibility enabled by the proposed QoE-Manager is to record and monitor real-time quality estimations per session. Acquired QoE intelligence can assist operators in comprehending and better managing their customers' overall, long-term experience, increasing thereafter their loyalty level. Operators may also benefit by offering personalized services based on customer profile analytics. Moreover, the opportunity emerges for creating new, QoE-based business models, to the benefit of both the users (e.g., receive differentiated quality upon demand) and the network providers (e.g., impose correlate charges).

Another possibility is to improve the QoE of a current flow, or to maximize the sum/average QoE of the served users catholically, e.g., by expressing the total QoE as a utility function. A quality improvement may be requested either proactively or reactively. The former approach requires the prediction of network problems via QoE-based alarms, while the latter means reacting to problems already present. Potentially, any network control measures (e.g., admission control, flow prioritization, cross-layer scheduling) may be implemented, respecting network policies and constraints. The QoE-Manager can also keep track of the effectiveness of these decisions, and hence, be able to self-adapt and optimize the methods used for solving the identified quality problems.

Finally, through the QoE-Manager, the opportunity emerges to exploit QoE awareness as a way to potentially save on network resources without compromising the overall customer experience. This may become possible either by identifying moments and cases of operation when providing extra resources to a user would not improve the QoE perceived, e.g., [22], or by exploiting the non-linear relationships between QoS and QoE, such as the ones quantified by the IQX hypothesis and the WFL law [19]. As described in Chapter 2, the former relationship claims a negative exponential dependency between the perceived QoE values and degrading QoS parameters, while the latter describes the logarithmic impact of physical stimuli on the human perception; therefore, such relationships provide the potential for devising novel QoE management algorithms that help avoid over-engineering phenomena in terms of QoE impact.

Various novel QoE management techniques, in the context of mobile cellular networks are presented in Chapters 6, 7, 8, and 9.

3.4 Enabling end-to-end QoE support in mobile cellular networks

3.4.1 Realization issues and challenges

Mobile cellular networks with QoE management aspirations may adopt and customize the proposed framework. The network-specific decisions that need to be taken are:

1. The physical location of the QoE management framework inside the operator's infrastructure: Challenges include determining whether this framework will be implemented as a stand-alone entity or not, centrally or in a distributed fashion, as well as developing new interfaces to support communication with other network nodes and the users.
2. The identification of the required QoE data sources, the configuration of the data collection periodicity, as well as the signaling between the network and the QoE-Controller: The main concern is the minimization of the extra signaling overhead imposed in the network, compromising between scalability and estimation accuracy issues. Also, the consumed power required for the QoE-data collection should be considered, mainly to avoid drainage of the handheld devices' battery.
3. The selection of appropriate QoE models for the QoE-Monitor: Research is needed on finding ways to limit the imposed signaling required by these models, and to reduce the complexity of the QoE estimation process. Moreover, new models will need to be devised in the future, mainly to capture the long-term QoE and customer churn, based on multiple, sequential episodes with the same service. Finally, traffic/service classification performed in the QoE-Monitor is a very challenging issue, especially in the content-encrypted domain (e.g., Hyper Text Transfer Protocol Secure (HTTPS)).
4. The type of decisions taken by the QoE-Manager and their actualization through the QoE-Controller: Since these decisions need to be performed on a per flow basis, and since the number of users in the network may be large, scalability and complexity issues are raised here as well.

Except for these technical challenges, the operator needs to account for some business and legal aspects too. First, ensuring end-to-end QoE may depend on multiple network, service or content providers, especially at infrastructure inter-connection points; therefore, collaborations and SLAs among different stakeholders are required. Second, security and privacy issues are raised, since potentially user-sensitive information has to be traversed through the network, for QoE management purposes. Net neutrality issues also emerge, especially if packet differentiation is selected for QoE provisioning. Finally, the operator needs to come up with proper business cases and monetary incentives, before being convinced to implement and commercialize such a QoE management scheme.

3.4.2 Evaluation results: The LTE case study

In this section, we use LTE as a case study to demonstrate the feasibility, performance and potential benefits of the proposed QoE management scheme, using simulation. To this end, we have expanded the open-source LTE-Sim [36] to support this framework.

We first estimate the amount of extra signaling imposed for QoE monitoring during the real-time operation of this framework, as well as the resulting accuracy of the QoE estimations. Overhead occurs due to the communication exchange between the QoE-Controller and the network, whereas communication among the three main building blocks of the framework takes place internally inside the central entity. The QoE-Controller is responsible for configuring the periodicity of the QoE-related data collection, referred to as the “QoE reporting period”.

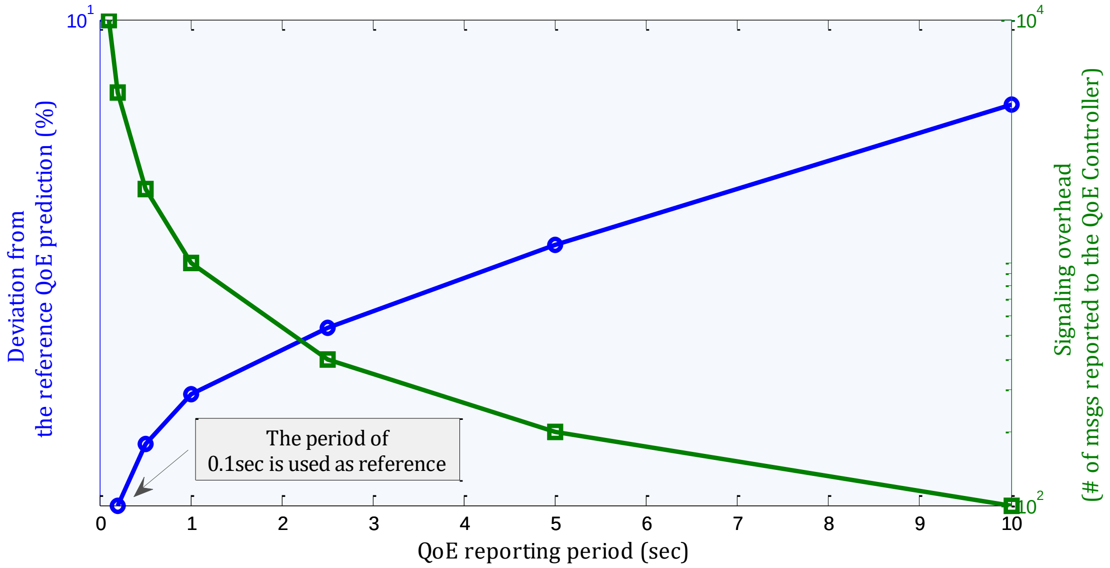
For this study, we simulate a heterogeneous network, consisting of one macro-cell served by an eNB, small-cells served by HeNBs located inside 5x5 3GPP-based building blocks, and finally uniformly distributed UEs. We count the number of messages collected by the QoE-Controller during configurable QoE reporting intervals (in this case, one message per UE per interval) roughly quantifying in this way the imposed overhead. With the input parameters of Table 3, we estimate how accurate the predicted QoE scores are per reporting period, using as reference the case where QoE-input is collected per 0.1 seconds. We report the obtained results in Figure 14a, reaching to the conclusion that there exists a trade-off between the amount of signaling overhead and the achieved accuracy in the QoE predictions. The results are closely dependent on the actual QoE estimation model used (here, we use the ITU-T G.107, “E-model”), while different signaling requirements are expected by different models.

Table 3: Basic simulation parameters for QoE-driven admission control.

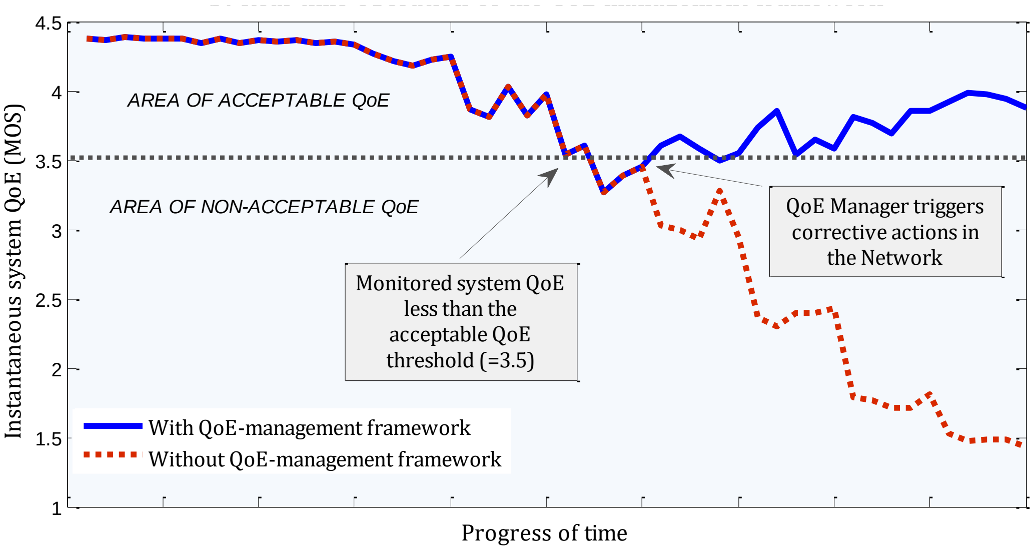
Parameter	Value
Macro-cell radius	1 km
eNB TX power	43 dBm
HeNBs TX power	23 dBm
Number of UEs	Scalable
Distribution of UEs	Uniform inside the attached cell
Traffic load per user	1 VoIP call
VoIP codec	G.729a
Duplex mode	Frequency Division Duplex (FDD)
Channel bandwidth	10 MHz (split between macro- and small-cell)
Scheduling algorithm	Proportional fair
Flow duration	10 sec
QoE reporting period	0.1 - 10 sec
Maximum acceptable delay	0.1 sec
Packet loss robustness factor	Zero
QoE estimation model	ITU-T G.107 (E-model)

Nevertheless, this overhead may be counterbalanced considering the new opportunities enabled for QoE-driven network management. As a characteristic example, we describe how the proposed framework may be customized and applied towards implementing a real-time QoE-aware Admission Controller. We study the case of a heavily congested outdoors small-cell, representing for instance scenarios where this small-cell is used to serve a stadium during a concert or football game. We evaluate the proposed QoE management framework and compare it with the conventional case, where, in the absence of QoE awareness, users are admitted based on their positions or on received signal strengths from surrounding base stations. The proposed framework is customized as follows:

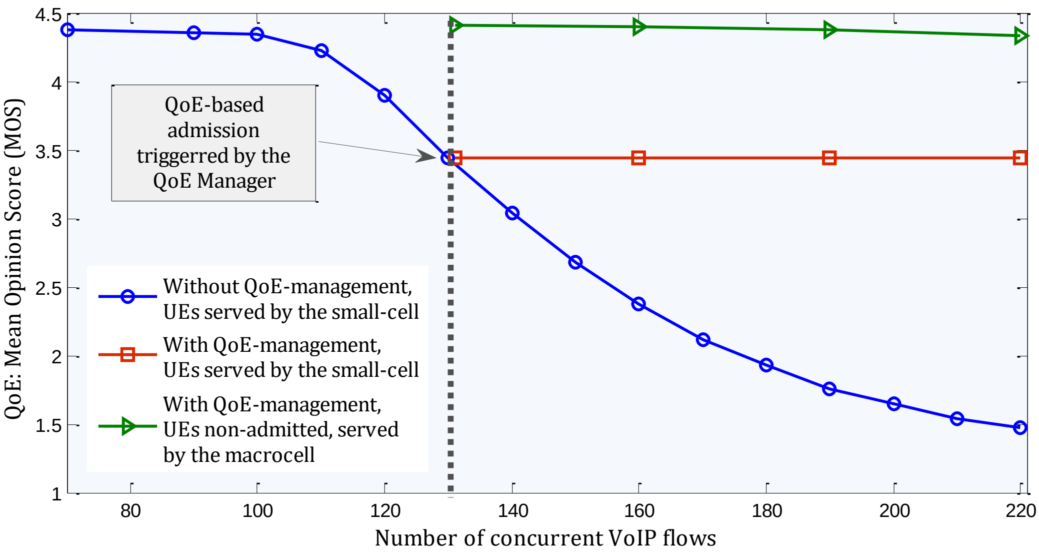
- **QoE-Monitor:** We study the case of UEs producing VoIP traffic and select the E-model implementation for the purposes of QoE estimation. Thus, the QoE-Monitor provides the QoE-Manager with real-time estimations of the QoE experienced per VoIP flow.
- **QoE-Controller:** The data collection procedure is tuned, a priori, with the QoE modeling function. Consequently, the E-model dictates the periodic collection of:
  a) The average delay associated with the transmitted packets, extracted through examining the timing information available inside the received packets.
  b) The packet loss rate, estimated as the number of erroneously received packets over the aggregate number of transmitted packets, measured by the number of negative acknowledgments produced throughout the QoE reporting period.
  c) The packet loss robustness factor (the average number of consecutively lost packets over this number for the case of random loss), acquired using statistical information by intermediate network nodes.
  d) The codec type of the UEs, required to select the appropriate E-model coefficients.
- **QoE-Manager:** The QoE-Manager is informed by the QoE-Monitor about the estimated QoE per VoIP flow, and consequently, is aware of the average QoE of the served UEs. If this QoE score reaches a minimum acceptable threshold (here, MOS=3.5), the QoE-Manager will restrict the admission of new flows inside the small-cell. Instead, those will be served by the macro-cell. In this way, a QoE-driven admission control mechanism is implemented.



(a) Trade-off between the network overhead and achieved accuracy in the QoE prediction.



(b) Real-time operation of the QoE management framework.



(c) QoE-driven admission control.

Figure 14: QoE management framework evaluation results.

To evaluate this framework, we generate a constantly increasing number of VoIP flows inside the small-cell, namely within the range of the HeNB, using the simulation parameters of Table 3, and we record the instantaneous average QoE in the system, while time progresses (Figure 14b). We observe a point when this QoE drops below the predefined MOS threshold, due to the increasing number of competing requests for spectrum resources. This event triggers the QoE-Manager to restrict the admission of new flows inside the small-cell, causing any new-comers to be admitted by the macro-cell instead. If the macro-cell is not severely congested, as is the case here, the average system QoE will be lifted above the threshold (blue plot in Figure 14b), which is not the case if this QoE admission mechanism is not present (red plot).

In Figure 14c, we look at the same experiment in a more microscopic level, namely we evaluate the achieved QoE level for users admitted either by the small-cell or the macro-cell. Again, we observe at some point a QoE drop below the threshold (specifically, for 130 concurrent VoIP flows inside the small-cell). At this point, the QoE-Manager does not allow any new flows to be admitted by the HeNB, and, so, the average QoE inside the small-cell remains constant onwards (red plot in Figure 14c). In parallel, the new flows, which are forced to be served by the eNB, also receive good

QoE (green plot), subject however to the current load of the macro-cell (note a small QoE decrease from 1 to 90 admitted flows). Consequently, we conclude that the application of this QoE management framework surpasses conventional admission control schemes, which would force all new flows to associate to the HeNB based on QoE-unaware criteria (blue plot).

3.5 Conclusions

Mobile cellular technologies, such as 4G and 5G, are moving from network-centric to user-centric approaches, by incorporating some kind of QoE logic and intelligence. Towards this direction, this study has focused on the integration of QoE acquisition and QoE management inside these networks. A framework for end-to-end QoE management is proposed, its viability is investigated, and key challenges for its realization are identified and discussed. Therefore, this work contributes to the need of providing more structured and focused insight on the issue of QoE management in mobile cellular networks, assisting operators with QoE aspirations to adopt this framework and customize it according to specific requirements and needs.

4. METHODS AND METRICS FOR QoE ASSESSMENT

In this chapter, we study QoE evaluation and estimation approaches (i.e., QoE models) towards a user-centric network management. Fundamental background on QoE quantification has been gathered with the following objectives: a) to describe the main QoE estimation methodologies, b) to classify these existing methods based on diverse criteria, c) to compare these methods based on their advantages, disadvantages and implementation challenges, d) to clarify the major QoE influence factors, and e) to reveal the most important objective QoE estimation requirements for mobile cellular networks. As a conclusion from this chapter, the importance of parametric QoE estimation is highlighted.

4.1 QoE estimation taxonomy

There are various different approaches for quantifying the QoE level of a provided service. A primary classification of the available approaches is based on whether QoE is evaluated directly by humans or automatically through technical factors. In the first case, specific assessment processes are used, referred to as subjective models/tests, while in the second case mathematical formulas or algorithms are exploited, referred to as objective models. The main classification of QoE models is presented graphically in Figure 15, and is further discussed in the next subsections.

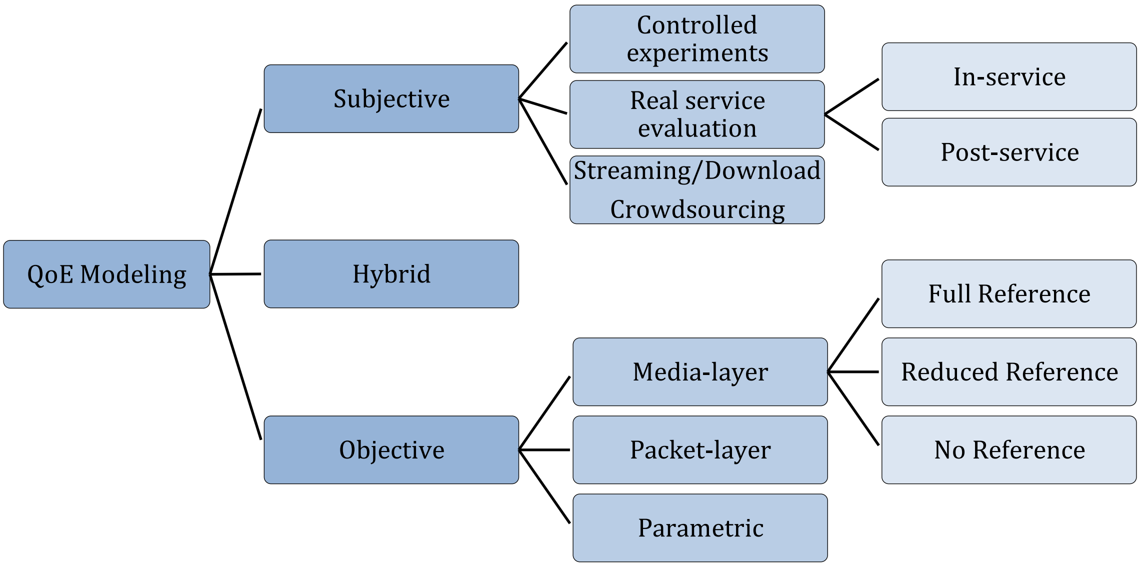


Figure 15: Classification of QoE modeling approaches.

4.1.1 Subjective QoE estimation

Subjective tests are usually based on controlled real-life experiments with human participants who directly evaluate their experience of an application or service. These users may be involved in the experiment in a passive way (just viewing/listening) or in an active/interactive way (participating in a conversation) and they judge the quality regarding some stimulus' presentation. For instance, the participants may be called to evaluate the listening or conversational quality of a phone service, the quality of a video, etc. These tests need to be thoroughly designed in advance and the user group needs to be properly selected based on guidelines and recommendations by standardization bodies. Perhaps the most important recommendation towards that direction is the ITU-T P.800 [37]. Various techniques may be used for subjective evaluation. For instance, users may score the quality using an absolute rating scale or they may compare sequential images/videos/sounds stating which one is better. The results are based on user opinions, past experiences, expectations, user perception, judgement and description capabilities, etc. and primarily quantify the effectiveness, efficiency and overall satisfaction of using a service.

These kinds of subjective tests are considered as the most reliable ones, since they incorporate any conscious and unconscious aspects of human quality evaluation, aspects that can otherwise not be captured. Indeed, only perceptual quality tests can validly and reliably express the internal state of the human factor. Nevertheless, such subjective techniques are considered reliable, if and only if they are designed carefully and users are unbiased and objective.

One drawback of the above method is that the results of such experiments are valuable only for the laboratory testing of some service, and not for real-time QoE support. One way to overcome this issue is to conduct "real-service" QoE evaluation, where users rate their experience on the run (in-service) or after a service has ended (post-service). Such an example is the "OneClick" paradigm, which may be used for real-time QoE monitoring and feedback, and consequently for QoE control. This framework only requires a subject to click a dedicated key whenever he/she feels dissatisfied with the quality of the application in use [38]. Furthermore, an example of post-service test is that of Skype, where users rate their experience once a session is terminated, using the MOS scale.

Subjective experiments in controlled laboratory environments need thorough design that strictly follows guidelines provided by standardization bodies. These guidelines describe all aspects such as room conditions (e.g., isolated room, without any noise), audio headset or generally the dedicated equipment used for hearing/viewing/talking, test methodologies, guidelines for the selection of the panel, etc. Regarding the latter, there are guidelines regarding the number of participants, their age, their background (experts or non-experts), their past involvement in similar experiments, the randomness in their selection, etc.

However, lately there is also a trend to evaluate the quality of an application in a more relaxed way, i.e., at one's own and familiar environment, using one's own equipment and so forth. In this kind of experiments, a service is evaluated using "streaming" or "download" approaches. These methods are considered as more realistic and are open to a much broader public as compared to laboratory experiments, thus allowing for better management. Indeed, a large number of participants may reveal very reliable and realistic QoE scores. Approaches that follow this paradigm are called "Crowdsourcing" techniques [39], because they outsource the task of quality evaluation to arbitrary anonymous online users. One such example is the Google Microworkers platform as well as the Amazon Mechanical Trunk, where an Internet user may conduct QoE experiments designed by other parties (such as researchers), who require a general public for an evaluation task.

Finally, an important issue in subjective test methodologies is the discrimination between "instantaneous" and "overall" quality evaluations. The former method implies a continuous evaluation of the perceived quality by the user during one experiment (see ITU-T P.880), whereas the latter simply requires that the user gives one cumulative score for his/her own experience at the end of each experiment. The first method gives a better insight to the system designers, since they can correlate the instantaneous quality with momentary technical parameters in the network; however, the latter better describes the overall user experience.

4.1.2 Objective QoE estimation

Subjective tests are costly, time-consuming and not reproducible on demand. Moreover, they are usually not real-time and hence cannot be used for in-service quality monitoring. These constraints have raised the need for the development of objective models that try to measure or predict the quality perceived by users, without their intervention. The objective models may be classified using various criteria [34], [40] (as also briefly mentioned in Section 3.3.1):

- **Reference signal utilization:** Regarding whether the source signal or part of it is required or not as input in the QoE estimation process, we distinguish the *Full Reference (FR) or reference-based or double-ended* models, the *Reduced Reference (RR)* models and the *No Reference (NR) or single-ended* models, where "reference" refers to the original signal.

  FR models do not require any a-priori information or assumptions about the underlying network, since they presuppose the exploitation of the source signal, and are highly accurate and robust, at the cost of not providing any insight about the system under test. NR and RR models, on the other hand, do not require the original source signal, but they do require prior knowledge about specific technical characteristics of the system. Despite their complexity, these models are more realistic as an implementation option in mobile cellular real-time networks. [41] conducts a survey on the evolution of video quality assessment methods using this classification.

- **Evaluation method:** Regarding the kind of input information that is used for QoE measurement, we distinguish the: *Media-layer (signal-based)*, *Packet-layer / Bitstream*, and *Parametric* models (see Figure 16). Media-layer models make use of transmitted and/or received signals and may be FR, RR or NR. Packet-layer models extract information from packet headers, while bitstream models may use both packet headers and payload data. Parametric (or parametric planning) models use specific network planning parameters and metrics, such as delay, packet loss, jitter, etc., as well as terminal quality parameters. Hybrid models, finally, combine characteristics of any of the above methods.

  Packet-layer and bitstream models are also referred in the literature as "protocol-information-based" models, because they base their estimations on parameters collected at run time from network processes and control protocols. Various surveys in the literature review media-layer models (e.g., [42] thoroughly discusses media-layer models for video quality assessment), while others focus on packet-layer/bitstream models (e.g., [43]). Finally, [44] conducts a study of the correlation models mapping QoS to QoE for multimedia services, providing in this way generic formulas that parametric models usually follow. In more detail:

  a. Media-layer: The major representative of this category is described in ITU-T P.862 [45]. It compares the original reference signal with the degraded output signal as it results from passing through a communication system. It is a perceptual and cognitive model where a Perceptual Evaluation of Speech Quality (PESQ) score is mapped to an objective MOS listening quality score. The model is applicable when it is implemented in specific environments where the input signal is reachable.

  b. Packet-layer / Bitstream: Models of this kind extract information from the packets travelling in the network. The most representative one is the ITU-T P.564 [46]. This is a no-reference type model that exploits packet header / payload information to acquire a QoE score. In Tables 8.1 and 8.2 of [46], the reader can find the detailed list of permitted input information that is used by this model for speech quality computation. However, the most important type of data used are the time-stamps and sequence numbers of the packets that travel in the network. The model is applicable for (passive) quality assessment and live QoE monitoring and assessment.

c. Parametric planning: Parametric QoE estimation models are currently the most appealing candidates for quantifying QoE levels in an indirect and user-transparent way in mobile networks. Thus, they will be further discussed in Chapter 5.

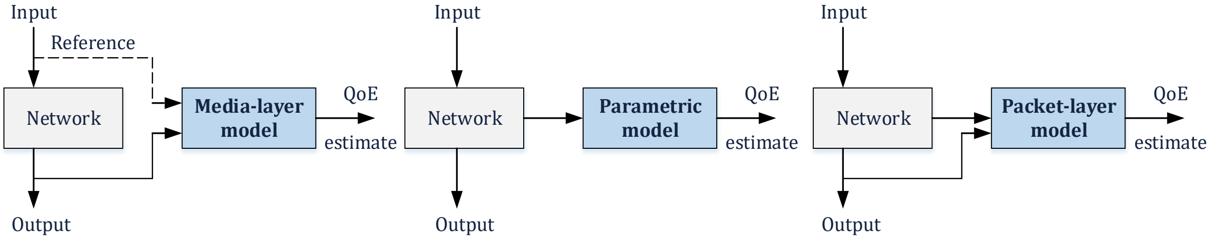


Figure 16: QoE model evaluation method (based on [34]).

- **Model mode:** The signal evaluated by an objective QoE model may either be a specific signal injected into the network exclusively for test purposes or a signal really used for communication purposes. According to this discrimination, we distinguish the *intrusive (active)* and *non-intrusive (passive)* modes, respectively. Intrusive models have the disadvantage of occupying additional network resources for no actual communication purposes; however, they allow for a better control and comprehension of the relationships between system input and achieved output quality.
- **Model timeframe:** Dependent on their time of implementation, *offline* models refer to pre-service or post-service evaluation methods, whereas *online* refer to in-service, hence real-time, quality evaluation.
- **Usage purpose:** This criterion refers to the aim of QoE modeling. For instance, it may be targeted for network planning, lab-testing, real-time service monitoring, optimization, benchmarking, etc. Also, different models target different applications, such as: audio, video (audio-visual), data (web), graphics, text, live TV, VoIP, browsing, video-telephony, teleconferencing, real-time gaming, etc. The QoE models should be carefully used only within their scope.

Parametric QoE models are basically derived by conducting subjective experiments (lab or crowdsourcing) and then by performing statistical analysis (e.g., regression analysis) on the acquired evaluation results. The derived objective models may be then well-described by providing formulas for the direct computation of QoE based on specific input parameters. On the contrary, signal-based models are based on one-to-one comparison between the original source signal and the degraded destination signal, by exploiting knowledge from the area of Psychophysics.

Also, worth mentioning is a third category of QoE modeling, which lies between the subjective and objective ones. It operates in a hybrid fashion, namely it works as an automatic and objective quality estimator, relying however on prior available subjective scores. These hybrid methods are based on Machine Learning tools, and they are using subjective test scores as input to train a QoE model. This model then maps network parameters (e.g., codec used, packet loss rate, mean loss burst size, packetization interval, one–way delay, jitter, etc.) to MOS values and it can be further used for real-time quality prediction. Characteristic examples of this approach are the Pseudo-Subjective Quality Assessment (PSQA) method [47], the MLQoE, a modular algorithm for user-centric QoE prediction [48], and the Adaptive Neural Fuzzy Inference System (ANFIS)-based video quality prediction model [49].

What is more, some research works propose methodologies for the construction of objective QoE models. For instance, [50] describes basic principles for building a QoE model from scratch, which is based on the egress of parameterized mappings among three layers: the transport-layer, the service-layer and the user layer (bottom up). Similarly, in [51], the authors build a QoE estimation function based on a general regression model and prove its applicability to web browsing and file upload/download scenarios.

At the moment, most objective models account for the human factor in terms of their inherent characteristics, but the context and content of the tested service are considered only at a limited extent. Under this observation, more research and standardization work is needed for designing more accurate objective estimation models. Especially extra forces should be allocated towards the designing of new parametric QoE estimation models, since they are currently the most appealing candidates for quantifying QoE levels in an indirect and user-transparent way in mobile networks. Taking this into account, the dominant parametric QoE estimation models are studied in the next Chapter 5, and used throughout the thesis.

4.1.3 Comparative study of QoE models

Focusing on the challenging issue of QoE model selection, in Table 4 below, QoE estimation models are classified based both on the subjectivity involved (i.e., subjective, objective and hybrid models) and the evaluation method used (i.e., media-layer, parametric planning, packet-layer models, as well as models following the QoS-to-QoE mapping logic). Furthermore, advantages, disadvantages and obstacles/challenges in adopting them for practical QoE estimations are discussed. Finally, characteristic examples of either standardized or non-standardized QoE models are provided per category.

Following the previous classification and comparison, in Table 5 below, representative QoE estimation models per evaluation method are described in terms of their applicability to mobile cellular networks (last column). To elaborate on this, information on each model’s logic/technique, required input, produced output and purpose is also provided, namely:

- **Logic / technique:** Here, the type of method that is adopted by each QoE model is briefly described.
- **Input:** Here, the information that is used by each model as an input is presented. This input may be a signal, one or more key parameters, information extracted from IP packets, user feedback, etc.
- **Output:** Here, the produced output by the respective category is described. This might be, for instance, a MOS score.
- **Purpose / usage:** This entry describes the primary target of the respective QoE model. A model may be used for network planning or for monitoring and performance evaluation, amongst others. Moreover, a model may be able to work proactively, i.e., to predict a bad QoE value and improve it before quality degradation is perceived by the users or it may respond reactively to an alarm indicating a decreased QoE.
- **Applicability to mobile cellular networks:** An overall comment about the feasibility of each model in mobile cellular networks is given. The pros and cons in this direction are therefore included. Some overall conclusions about requirements towards the applicability of QoE models to mobile cellular networks are summarized in Section 4.5.

Table 4: Classification and comparison of QoE models.

Evaluation method	Advantages	Disadvantages	Obstacles	Model examples
SUBJECTIVE MODELS				
Subjective evaluation process	+ The most reliable QoE measurement procedure, highly accurate and valid + Subjective models ensure uniformity between subjective scores from different laboratories	- Not real-time (require lab setting) - Not reproducible on demand - Time consuming and expensive - May be biased by user opinion, assumptions, unconscious factors - Users may be greedy on their QoE demands, and hence, QoE evaluations - Possible users' tiredness and lack/loss of concentration/interest	▪ Experiments need to be conducted under strict requirements and highly controlled conditions: isolated sound rooms, dedicated equipment, suitably selected panel, specific duration and configuration of tested signals, etc. ▪ User impediment to discriminate between e.g., “Bad” and “Poor” MOS	ITU-T P.800 & P.880
OBJECTIVE MODELS				
Media-layer, Full Reference	+ Do not require a-priori knowledge or assumptions about the underlying network + Highly accurate and robust	- Very high computational effort - Do not enable insight into the internal system functionality and problem causes (black-box) => diagnosis impossible - Neglect human dimensions, pure technical approach - May measure one-way speech quality only => interactivity not tested - May not test full-length call quality	▪ Require the reference signal that is not commonly available (intrusive approach) ▪ Prototype or simulation of the transmission channel is necessary ▪ Practically impossible to implement at network midpoints ▪ Pre-processing is necessary for level- and time-alignment of the input and output signals	ITU-T PSQM (P.861), PESQ (P.862), POLQA (P.863)
Media-layer, No Reference	+ Estimate QoE in a passive way (non-intrusive approach) + It can be used at any arbitrary location in the transmission chain	- Complex to implement - Only measure the effects of one-way speech distortion and noise (listening) - Impairments related to two-way interaction are not reflected	▪ Require prior knowledge about specific technical characteristics of the system ▪ Distortions have to be known in advance	ITU-T P.563
Parametric planning	+ Ease of use and respect of privacy + Only technical network specifications required, not the reference signal (non-intrusive) + Quantifies the human factor and context of use, e.g., “advantage factor” + Mouth-to-ear complete transmission chain quality evaluation + No restrictions on the network (size, configuration, hierarchy, technology, components)	- Originally intended only for the planning phase of a system, unless extended - Accurate only under strict scenarios - Speech independent - Valid under a given planned transport configuration - Input parameters are considered constant during a voice session, which is inconsistent with the features of time-varying QoS transport networks [33]	▪ Difficult in practice to include all model parameters online ▪ Some a-priori information is also required ▪ New subjective tests followed by new regression analysis need to be performed for different conditions to derive new models	ITU-T G.107 & G.1070

Packet-layer	+ Enables insight into the internal system functionality (glass-box) + Light computational effort + Multiple monitoring points and deep packet inspectors help identify the problem roots + Used both for speech quality predictions and for diagnosis + In-service, non-intrusive => respect of user privacy	- Assume a generic voice payload - Only concern impairments on the IP network (no end-to-end evaluation) - Large volume of QoE-related input data will be produced - The quality prediction refers only to the immediately preceding interval - One-way listening quality	▪ Not readily available: models need to be created that comply with the recommendations of standardization bodies ▪ Depend on packet visibility at the probe ▪ Models deployed require strict conformance testing ▪ Have to be validated using media-layer methods	ITU-T P.564
QoS-to-QoE mapping	+ These approaches provide a generic function that transforms available QoS parameters to QoE values + Exponential/Logarithmic relationships are revealed that enable QoE management based on QoS monitoring	- Only one QoS parameter can be mapped to a QoE value each time - The mapping function is not known in advance - In order to find the mapping function another QoE model may have to be used	▪ The mapping function needs to be validated per service and scenario before being adopted as valid ▪ Periodic re-validation is required	IQX hypothesis & WQL hypothesis
HYBRID MODELS				
Parametric planning, hybrid	+ Considers all type of media + Network, application and user dependent + Highly accurate + Can work in real-time	- Human involvement required in order to train and configure the model	▪ Initial configuration is required for selecting the influence factors and their ranges ▪ The model has to be periodically re-configured or re-validated, requiring human input again	PSQA

Table 5: Overview of characteristic QoE models.

Model type	**Logic - technique**	**Input**	**Output**	**Purpose / Usage**	**Applicability to mobile cellular networks**
Media-layer: **ITU-T P.862 (PESQ)**	Compares the original reference signal with the degraded output signal (Full Reference type)	▪ The output (degraded) speech signal ▪ The input (original) speech signal ▪ Perceptual and cognitive models are required to process the two signals	PESQ score mapped to objective MOS, determined by comparing the two signals at any system output point	▪ Laboratory monitoring ▪ Live testing of prototype or emulated networks ▪ Quality benchmarking ▪ Codec evaluation and selection	- The reference signal is not readily available in real time => impossible for real network monitoring - Computationally heavy - Raises privacy issues - Handles the network as a “black-box” + May be used for network planning purposes

Media-layer: **ITU-T P.563**	Requires a preprocessing stage, a distortion estimation stage and a subsequent perceptual mapping stage	▪ Output speech signal ▪ Types of calculated signal parameters are: basic speech descriptors, vocal tract analysis, speech statistics, static SNR, segmental SNR, interruptions/mutes	MOS on the Absolute Category Rating (ACR) Listening Quality Opinion scale (MOS-LQO)	Non-intrusive speech quality assessment, live network monitoring and testing with unknown speech sources at the far end side	- The model may not be accurate when part of the output signal is missing due to network errors - Listening-quality evaluation only + Live end-to-end quality monitoring is feasible + Not necessarily restricted to narrowband applications
Parametric planning: **ITU-T G.107**	Provides a well-defined computational formula with specific input parameters	▪ Quality parameters from various network constituents and the terminals, e.g., SNR, delay, equipment impairment (codec), echo, loudness, packet losses ▪ Advantage (compensation) factor	Transmission Rating factor (R), mapped to a MOS value	▪ Performance evaluation ▪ Network planning, application and terminal design ▪ Helps avoid “over- and under-engineering” phenomena	- In its original format the scope of the model is not for in-service evaluation - Extra signaling required to gather and transport the model’s input to the QoE evaluation point - Accurate only for specific recommended application scenarios + Can be simplified to transport-level metrics for simple real-time monitoring + The analytical model formulas are not very difficult to implement
Parametric planning, hybrid: **PSQA**	Implements statistical learning tools from the Random Neural Network (RNN) area	Network-wide and user equipment, e.g., codec, error correction, offset, packet loss rate, mean loss burst size, packetization interval, one–way delay, jitter, etc.	MOS score may be used as output	To dynamically optimize quality by understanding the relations among quality affecting parameters and perceived QoE	- The model’s real-time applicability is ambiguous due to human involvement for training and validation - The implementation of RNNs in the network invokes extra complexity + Can capture the human-context-technical factors’ influence
Packet-layer: **ITU-T P.564**	Extracts information from the packets travelling in the network using DPI techniques	▪ Available at any mid-network monitoring points (probes, etc.), e.g., time-stamps, sequence numbers from packet headers, and/or payload ▪ Extra information about e.g., endpoints transferred through control packets	MOS on the ACR listening quality scale	▪ Non-intrusive (passive) quality monitoring ▪ Live assessment for reactive QoE control ▪ SLA support	- Models have to be designed based on this Rec. (not readily available yet) - Testing is required using a detailed conformance test methodology - Extra processing load for the network midpoints to analyze packets (big data volume is produced) + Help diagnose network problems at any point + Light model
QoS-to-QoE: **IQX & WQL**	▪ IQX: Translates generic network-level QoS to QoE (exponential dependency) ▪ WQL: Describes the effect of physical stimuli on human perception based on Psychophysics (logarithmic dependency) [19]	Any single QoS degrading parameter (one per mapping function), e.g., packet loss rate, setup time of a wireless connection, waiting and response times, etc.	▪ IQX: A mapping function between QoS degradation and QoE value ▪ WQL: The relationship between QoS-dependent stimulus and perception	▪ QoE monitoring and proactive QoE control ▪ Already considered for VoD quality performance assessment (e.g., YouTube)	- The mapping function has to be estimated beforehand and regularly validated - Depend on QoS parameters’ availability: ○ For IQX: network-level parameters ○ For WQL: application-level parameters + Enable real-time QoE control mechanisms that build on QoS monitoring + Enable smarter resource control (allocate resources when users will really perceive the difference)

4.2 Key quality influence factors

According to [6], an influence factor is "*any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user*". The most important influence factors for the users' QoE are depicted in Table 6. The improvement of one or more of these factors indicates that QoE will be also improved, and thus, network engineering targeted on these factors should be conducted. From another perspective, these factors may be seen as the major "impairment factors" for the quality degradation of a provided service [35]. Moreover, most of these factors may constitute KPIs for QoE assessment.

Table 6: Major QoE influence factors.

Aspect	Quality influence factors
Service-independent	
Transport/ Network layer	Round trip / one-way delay, jitter, packet loss ratio, delay burstiness distribution, loss burstiness distribution, congestion period, packet size
Physical layer	Signal to Noise Ratio (SNR) / Signal to Interference Ratio (SIR) / Signal to Interference plus Noise Ratio (SINR), throughput, bottleneck bandwidth, bit rate, Block Error Ratio (BLER), outage probability, packet / symbol / bit error probability, outage capacity, ergodic capacity / rate / throughput, diversity order / coding gain, area spectral efficiency
Equipment factors	Codec, de-jittering buffer characteristics (overflow, delay), Voice Activity Detection (VAD) / temporal clipping, echo cancellation, noise suppression artefacts, Packet Loss Concealment (PLC) algorithm, Talker Echo Loudness Rating (TELR)
Mobile networks' additional factors	Transient loss of connectivity (e.g., due to handovers), battery consumption, session establishment delay, accessibility, availability, reliability, GoS, QoR
Common factors	Charging policy and cost, service support, privacy, security, fidelity, conversational task, usability, accuracy, efficiency, context of use (environment, etc.), ambient noise level and variation, equipment brand, service provider reputation, comfort
Service-dependent[1]	
Video specific	Frame rate, video bit rate, video content (almost static / high motion, etc.), packet loss visibility, re-buffering, Group of Pictures (GoP) size and structure, video and audio synchronization, terminal type, monitor specifications, display size, type and resolution, ambient luminance, codec type and implementation, video resolution and video format, key frame interval, freshness, blocking
VoD	Video streaming: Number and duration of stalling events, total video duration, initial delay (start-up delay) / For HAS: time on highest layer, frequency and amplitude of switches, chunk size, buffer size, etc.
Download-type services	Web browsing: web page download time / For file download: data rate, file download time, delivery synchronization
Voice	Service-independent factors apply (e.g., packet loss ratio, delay, codec, coding rate), call setup success ratio / blocking probability, call setup time, call cut-off ratio, start-up time, response time

[1] More details in Chapter 5.

Although the factors listed in Table 6 are correlated with QoE, it is important to emphasize that the exact impact of this correlation on QoE may be possible only through the use of specific QoE evaluation and estimation schemes.

4.3 Quality metrics

The most common measure of QoE based on subjective testing is the MOS, which ITU-T defines as "*the mean of opinion scores, i.e., of the values on a predefined scale that subjects assign to their opinion of the performance of the telephone transmission system used either for conversation or for listening to spoken material*" [52]. Although this definition makes reference specifically to the telephone system, the MOS score is adopted in the evaluation of a variety of services. Typically based on an ordinal five-point numerical scale, ranging from 1 to 5 to denote an increase in QoE, MOS scores are sometimes assigned a textual description (Table 7), as different MOS notations are often used depending on the employed evaluation method. It is also worth noting that besides the absolute MOS scale on which most tests rely, relative scales can be used when testers are required to perform a comparison between two samples [26]. Even though MOS was originally used for measuring the subjective quality of voice/video/data, objective models also yield a MOS score, either directly or via mapping a different score (e.g., the "Rating factor") to MOS.

Table 7: The MOS scale.

Rating	Label
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

The following quality metrics (or scales) can be also used during subjective tests:

- DSIS (Double-Stimulus Impairment Scale)
- DSCQS (Double Stimulus Continuous Quality Scale)
- PC (Pair Comparison)
- SSCQE (Single Stimulus Continuous Quality Evaluation)
- ACR (Absolute Category Rating)
- ACR-HR (Absolute Category Rating with Hidden Reference)

For media-layer FR models, the PSNR (Peak Signal to Noise Ratio) metric is also used for video quality assessment. PSNR is defined using the Mean Square Error (MSE) [53]:

$$PSNR = 20\log_{10}\left(\frac{255}{\sqrt{MSE}}\right) \tag{4-1}$$

$$MSE = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}[f(i,j) - F(i,j)]^2}{MN} \tag{4-2}$$

where $f(i,j)$ is the original signal, $F(i,j)$ is the reconstructed one, $M \times N$ is the picture size, and 255 is here the maximum luminance value. A PSNR value ranging from 30 to 40 characterizes a medium to high quality video.

Similarly, the SSIM (Structural Similarity Index) metric can be used. It is calculated on various windows of an image. The measure between two image windows x and y of common size $N \times N$ is given by the following formula [53]:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{({\mu_x}^2 + {\mu_y}^2 + c_1)({\sigma_x}^2 + {\sigma_y}^2 + c_2)} \tag{4-3}$$

where μ, σ^2 and σ_{xy} represent the average, variance and covariance respectively, while c_1 and c_2 are two constant variables. Based on PSNR and SSIM, more complex formulas have arisen in the literature for FR video quality assessment. Some characteristic metrics are the Video Quality Metric (VQM) and the Moving Pictures Quality Metric (MPQM) [53].

4.4 QoE estimation tools

Table 8 and Table 9 list commercially implemented and open source tools for QoE estimation, respectively. The lists are not exhaustive, since the ever-increasing interest for QoE monitoring pushes forwards the emergence of more such tools and solutions.

Table 8: Commercial QoE monitoring solutions.

Name	Online reference[2]
VQuad	https://www.gl.com/vquad.html
Witbe QoE robots	http://www.witbe.net
OPTICOM	www.opticom.de
VQmon	http://www.telchemy.com/vqmon.php
QoE Systems	http://www.qoesystems.com
Elecard Video Quality Estimator	http://www.elecard.com/en/products/professional/analysis/video-quest.html
SIGOS	http://www.sigos.com/use-cases/quality-of-experience-testing-qoe
Vasona Networks	http://www.vasonanetworks.com/
FIWARE QoE-API	https://forge.fiware.org/plugins/mediawiki/wiki/fiware/index.php/ Quality_of_Experience_(QoE)_API_Specification

Table 9: Open source tools for objective QoE estimation.

Name	Online reference[3]
Evalvid	http://www2.tkn.tu-berlin.de/research/evalvid/fw.html
PESQ	http://www.mathworks.com/matlabcentral/fileexchange/33820-pesq-matlab-wrapper
VQMT	http://mmspg.epfl.ch/vqmt
VQone	http://www.helsinki.fi/~tiovirta/Resources/VQone/index.html
NS3QoE	https://github.com/aphirak/qoe-monitor
NS3 Evalvid	https://gitlab.com/gercom/evalvid-ns3/
VQM	http://www.its.bldrdoc.gov/resources/video-quality-

[2,3] All online links accessed on 7/11/2017.

	research/software.aspx
QoE-RNN	https://code.google.com/archive/p/qoe-rnn
SMRT	https://github.com/MuSAELab/SRMRToolbox

4.5 Conclusions on QoE modeling requirements in mobile cellular networks

A QoE model can be appealing for integration in mobile cellular networks as long as it has certain characteristics. Specifically, it needs to be:

- An objective model, namely a model that does not require the human factor input at any stage.
- A no-reference model, namely a model where the original signal is not required at the QoE measurement location. In this way overhead and complexity are significantly reduced.
- A parametric or packet-layer model, so that the complexity is not very high and so that the information may be acquired inside the network using a simple mechanism.
- An online model, for in-service use, i.e., to support real-time QoE measurement.
- A passive (non-intrusive) model to avoid injecting pilots into the system just for QoE testing purposes and waste resources for that matter. It is preferred to exploit information already available inside the network under regular operation, which refers to actual, realistic communication scenarios.
- The input parameters used by this model should be easily and readily available.
- These parameters should ideally be accessed by the network side, in the sense that user agents at terminals will not be necessary.
- Finally, a general guideline is that the selected models (which will feed the QoE-Monitor block of Section 3.3.1) are of low complexity, well-standardized, and able to be implemented in real-time on top of existing network infrastructures.

As a conclusion, the use of media-layer models is not recommended for quality estimation in mobile cellular networks, due to the complexity or even impossibility of setting them up. On the contrary, parametric or packet-layer models enable the acquisition of already available information through various network nodes. However, packet-layer models are not well-standardized yet, and the collection and exploitation of packet header information requires a lot of original work. Thus, presently, parametric models seem to be the perfect candidates for real-time QoE management. In addition, they require less overhead and are capable of monitoring communication sessions through heterogeneous transport infrastructures, which is ideal for modern mobile environments [33]. Therefore, the next chapter presents standardized and well-known literature-based parametric QoE models.

As a final remark, it is worth mentioning, that the selective generalization of the previous model selection guidelines to other network types, such as Wi-Fi or Ethernet, is not excluded; however, their analysis is not in the scope of this thesis.

5. PARAMETRIC QoE ESTIMATION FOR POPULAR SERVICES

As we are moving forward to the 5G era, we are witnessing a transformation in the way networks are designed and behave, with the user placed at the epicenter of any decision. The shift from QoS to QoE service provisioning paradigms paves the way for flexible service management and personalized quality monitoring. This can be enabled by exploiting QoE assessment models, and especially parametric, i.e., formula-based QoE estimation methods.

Current literature on the topic of QoE modeling mainly offers classifications of existing standards and focuses on one specific service at a time. For instance, [29] studies speech quality estimation and provides a detailed taxonomy of standardized objective speech quality prediction models. Similarly, [33] conducts a thorough survey of QoE assessment approaches for VoIP services, while [32] focuses on IPTV. However, these considered service types are just a subset of the plethora of services available in current networks. With the availability of 4G and with 5G on the horizon, which allows the co-existence of multiple parallel resource-hungry requests, applications like video streaming and VoD constitute the prevalent traffic over a network.

What is more, survey papers on QoE estimation usually focus on models that require the originally transmitted signal or part of it to deduce the QoE at the receiver side, not targeting in this way at real-time network management application (e.g., [41]). On the contrary, parametric QoE models are appropriate for this type of scenarios; however, a handy collection of these models for different types of services is currently missing from the literature.

In this chapter, recognizing this gap in the literature between the lack of a proper manual regarding the objective QoE estimation and the ever-increasing interest from network stakeholders for QoE intelligence, we provide a comprehensive guide to standardized and state of the art parametric quality assessment models. More specifically, we identify and describe parametric QoE formulas for the most popular service types (i.e., VoIP, online video, video streaming, web browsing, Skype, IPTV and file download services), indicating the KPIs and Major Configuration Parameters (MCPs) per type. Throughout this chapter, it is revealed that KPIs and MCPs are highly variant per service type, and that, even for the same service, different factors contribute with a different weight on the perceived QoE. This finding can strongly enable a more meaningful resource provisioning across different applications compared to QoE-agnostic schemes. Overall, this chapter is a self-contained repository of QoE assessment models for the most common applications, becoming a handy tutorial to parties interested in delving more into QoE network management topics. The described QoE models are the ones also used throughout this thesis, for the purposes of QoE assessment and thus, QoE management.

5.1 Standardized parametric QoE estimation

In this section, we present two basic standardized parametric models, namely ITU-T G.107 for VoIP and ITU-T G.1070 for online video. We present the most substantial parts of these models, the full versions of which may be found in the ITU-T portal, while we also study the impact of their key parameters on a user's QoE.

5.1.1 Parametric QoE estimation for VoIP services: ITU-T G.107 (E-model for VoIP)

In VoIP applications, the QoE is expressed in terms of how clearly the user can listen and understand his or her interlocutor's speech, and how easy or not the communication is, due to potential arrival delays of speech Internet packets. Because of this, the models for this service are divided into the following categories: listening-only and conversational. Subjective assessment methods in VoIP services, are based on four testing axes [54]: Comprehensibility tests, Multi-dimensional test, Listening Quality and Conversational Quality tests. The MOS is the most extensively used measurement scale for observations of this kind. Concerning parametric objective methods, the ITU-T Rec. G.107, a.k.a. the “E-model” [55],[56] is the most reliable and representative approach.

5.1.1.1 The basic rating factor

The E-model provides a formula that can be used for the computation of the transmission quality of voice communications by estimating the mouth-to-ear conversational quality as perceived by the user at the receive side, both as listener and talker (Figure 17). It is a parametric model that takes into account a variety of transmission impairments producing the so-called Transmission Rating factor (R factor).

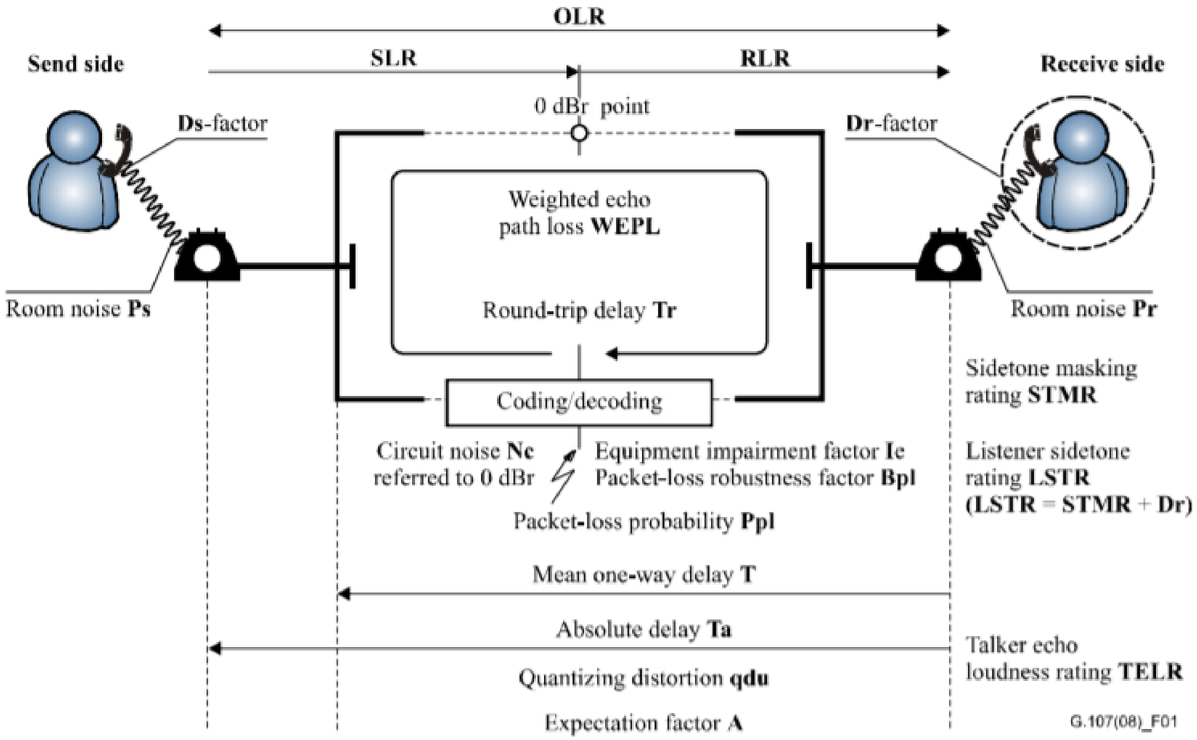

OLR
SLR
RLR
Send side
Receive side
0 dBr point
Ds-factor
Dr-factor
Weighted echo path loss WEPL
Round-trip delay Tr
Room noise Ps
Room noise Pr
Coding/decoding
Sidetone masking rating STMR
Circuit noise Nc referred to 0 dBr
Equipment impairment factor Ie
Packet-loss robustness factor Bpl
Listener sidetone rating LSTR (LSTR = STMR + Dr)
Packet-loss probability Ppl
Mean one-way delay T
Absolute delay Ta
Talker echo loudness rating TELR
Quantizing distortion qdu
Expectation factor A
G.107(08)_F01

Figure 17: Reference connection of the E-model [55].

The conversational quality is estimated by means of this rating factor R, scaling from 0 (worst) to 100 (best):

$$R = R_0 - I_s - I_d - I_{e-eff} + A \tag{5-1}$$

where:

- R_0 represents in principle the basic signal-to-noise ratio, including noise sources such as circuit noise and room noise.
- I_s is a combination of all impairments which occur more or less simultaneously with the voice signal.
- I_d represents the impairments caused by delay.
- I_{e-eff} represents impairments caused by low bit-rate codecs (effective equipment impairment factor). It also includes impairments due to randomly distributed packet losses.

- The advantage factor A allows for compensation of impairment factors when the user benefits from other types of access. For instance, the maximum value of A is 5 for offering mobility by cellular networks in a building, 10 for mobility in a geographical area or moving in a vehicle and 20 for access to hard-to-reach locations, e.g., via multi-hop satellite connections.

A simplified version of this model that enables real-time quality monitoring purposes is presented below.

5.1.1.2 Online adaptation of G.107 E-model

A methodology for QoE monitoring of VoIP applications in a network is described in [57], and it is presented in Figure 18. According to this, the E-model is reduced to transport-level parameters only, which can be easily measured within the network. The main idea is to combine transport-level measurements such as delay and packet loss with architectural-specific parameters such as the de-jitter buffer at the receiver side to get an estimation of the effective equipment impairment factor I_{e-eff}. This methodology can be therefore directly used for VoIP conversational quality measurement and monitoring.

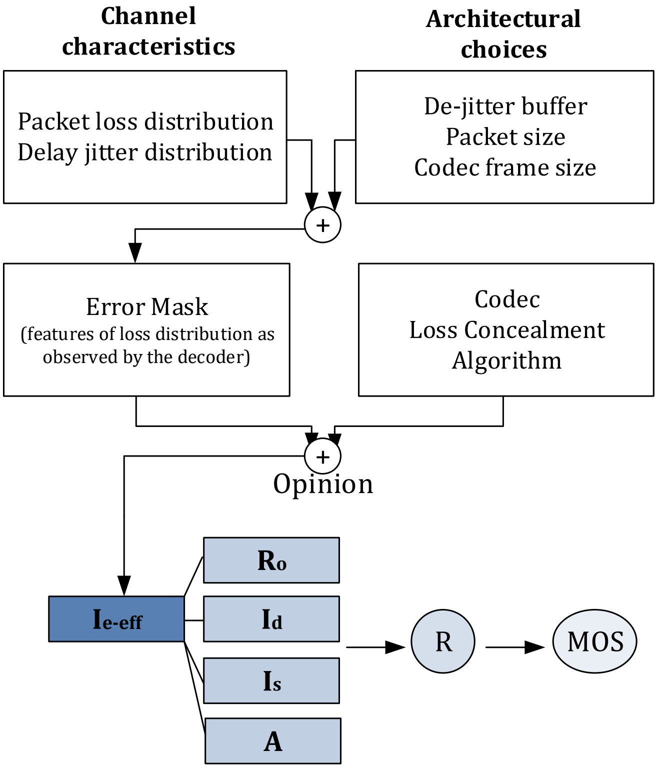


Figure 18: Methodology for VoIP quality measurement according to G.107 (based on [57]).

The extended version of the E-model may be simplified under specific assumptions according to [57]. These are:

- The existing model will be reduced to transport-level metrics.
- It will be used for monitoring the conversational voice quality (“online” use).
- Echo cancellers are properly working.
- The G.729a codec is used.
- Packet loss is random and up to 16%.
- Packet size is 20 msec.

- The "Advantage factor" is neglected.

In the case of the baseline scenario where no network or equipment impairments exist, the R factor is given by:

$$R = 94.2 - I_d - I_{e-eff} \tag{5-2}$$

Focusing on parameters that depend on the wireless part of the communication, i.e., transmissions between base stations and users, it holds that:

$$I_d = 0.024d + 0.11(d - 177.3)H(d - 177.3) \tag{5-3}$$

where:

$$H(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \tag{5-4}$$

(i.e., the Heaviside function) and d is the average packet delivery delay. Also, assuming that the codec G.729a is used, the packet loss rate, referred here as p, affects the parameter I_{e-eff} as follows:

$$I_{e-eff} = 11 + 40\, ln(1 + 10p) \tag{5-5}$$

By substituting these values to Eq. (5-2) above, we get a simplified expression for R, i.e.:

$$R = 94.2 - 0.024d - 0.11(d - 177.3)H(d - 177.3) - 11 - 40\, ln(1 + 10p) \tag{5-6}$$

The R factor can be used as an assessment value; however, we may transform it to MOS values to retrieve results comparable with results provided by subjective methods. The transformation formula is as follows:

$$MOS = \begin{cases} 1, & if\ R < 0, \\ 1 + 0.035R + R(R - 60)(100 - R) \cdot 7 \cdot 10^{-6}, & if\ 0 \leq R \leq 100, \\ 4.5, & if\ R > 100 \end{cases} \tag{5-7}$$

A simplified version of the E-model is available for using the G.711 codec too (see [57]).

Below, we graphically present the impact of average packet delivery delay and packet loss rate on the I_d and I_{e-eff} components, respectively, as well as on QoE. We observe a "knee" on the plot of Figure 19 at a delay value of 177.3 msec, after which MOS starts to degrade significantly. With respect to the impact of packet loss rate on QoE, we observe at Figure 20 that higher packet loss values cause a monotonical increase to the I_{e-eff} and a monotonical decrease to MOS.

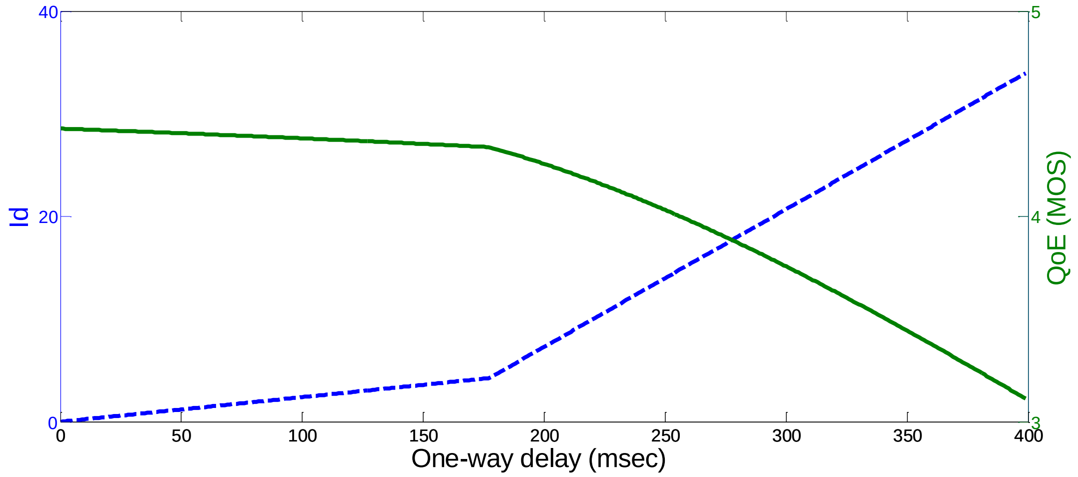


Figure 19: Impact of average packet delay on I_d and QoE.

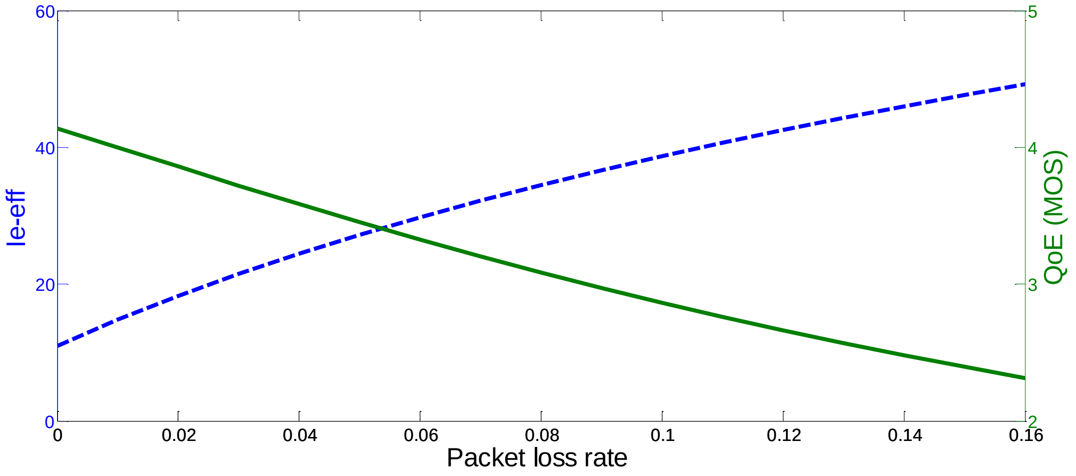
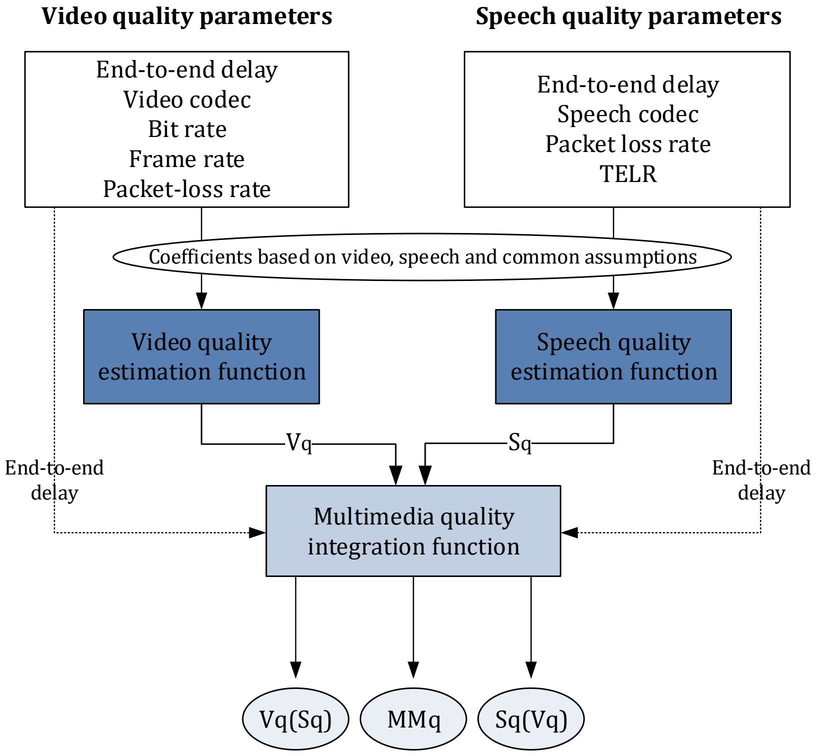


Figure 20: Impact of packet loss rate on I_{e-eff} and QoE.

5.1.2 Parametric QoE estimation for online video: ITU-T G.1070 (E-model for video)

The Recommendation ITU-T G.1070 [58] describes "*a computational model for point-to-point interactive videophone applications over IP networks that is useful as a QoE/QoS planning tool for assessing the combined effects of variations in several video and speech parameters that affect the QoE*". This recommendation assumes videophone applications using dedicated videophone terminals, desktop PCs, laptop PCs, Personal Digital Assistants (PDAs) and mobile phones, and it describes a parametric model applicable to online multimedia services over IP, such as a video conference.



Video quality parameters
Speech quality parameters
End-to-end delay
Video codec
Bit rate
Frame rate
Packet-loss rate
End-to-end delay
Speech codec
Packet loss rate
TELR
Coefficients based on video, speech and common assumptions
Video quality estimation function
Speech quality estimation function
Vq
Sq
End-to-end delay
End-to-end delay
Multimedia quality integration function
Vq(Sq)
MMq
Sq(Vq)

Figure 21: Methodology for multimedia quality assessment according to G.1070 (based on [58]).

The model provides three output quality metrics in the MOS scale, named the multimedia quality (MM_q), the video quality influenced by speech quality ($V_q(S_q)$), and the speech quality influenced by video quality ($S_q(V_q)$). Different formulas are provided for each one of them. The degradation caused by pure delay is considered only in the multimedia quality integration function. Note that various implementations can be found for a coding technology (e.g., MPEG-4 codecs) due to variations in coding-parameter settings and decoder characteristics. Therefore, the coefficients of video and speech quality estimation functions in this model were determined by referring to tables prepared in advance for each video and speech codec.

The framework and methodology of G.1070 are presented in Figure 21, where the key influence parameters for each one of the three aforementioned dimensions are presented (multimedia / video / speech). By mapping video, speech and common assumptions into specific coefficients, the impact of e.g., terminal type, monitor characteristics, environmental noise and conversational task on the multimedia quality is quantified.

Network, application and terminal quality parameters of high importance to QoE/QoS planners are incorporated into this model. Quality benchmarking and monitoring are not originally objectives of this recommendation, because some of the parameters required as input for the model are not readily available in real-time.

5.1.2.1 Video quality estimation function

Taking specific speech-related, video-related and task-related assumptions into consideration, as these are documented in the Recommendation G.1070, specific formulas have been derived for each one of the three aforementioned functions. Below, the video quality estimation function (V_q) is described, which takes values between 1 (worst) and 5 (best). The following notation is used:

- Fr_V is the video frame rate (fps).
- Br_V is the video bit rate (kbps).
- P_{pl_V} is the video packet loss rate (%).

As long as these three parameters are known, V_q can be estimated as described next.

The function that provides the video quality (V_q) is:

$$V_q = 1 + I_{coding} \cdot I_{transmission} \tag{5-8}$$

where:

- I_{coding} represents the basic video quality affected by the coding distortion under a combination of video bit rate and video frame rate:

$$I_{coding} = I_{Ofr} exp\left\{-\frac{(\ln(Fr_v) - \ln(O_{fr}))^2}{2D_{Fr_V}{}^2}\right\} \tag{5-9}$$

  Note than when $Fr_V = O_{fr}$ then $I_{coding} = I_{Ofr}$.

- $I_{transmission}$ represents the video quality affected by the transmission process:

$$I_{transmission} = exp\left\{-\frac{P_{pl_V}}{D_{P_{pl_V}}}\right\} \tag{5-10}$$

and:

- O_{fr} is an optimal frame rate that maximizes the video quality at each video bit rate:

$$O_{fr} = v_1 + v_2 Br_v, \quad 1 \le O_{fr} \le 30, \quad v_1, v_2{:}\, constants \tag{5-11}$$

- I_{Ofr} represents the maximum video quality at each video bit rate:

$$I_{Ofr} = v_3 - \frac{v_3}{1+\left(\frac{Br_V}{v_4}\right)^{v_5}}, \qquad 0 \le I_{Ofr} \le 4, \qquad v_3, v_{4,} v_5 \colon constants \tag{5-12}$$

- D_{Fr_V} is the degree of video quality robustness due to frame rate:

$$D_{Fr_V} = v_6 + v_7 Br_V, \quad 0 < \ D_{frV}, \qquad v_6, v_7 \colon constants \tag{5-13}$$

- $D_{P_{pl_V}}$ expresses the degree of video quality robustness due to packet loss:

$$D_{P_{pl_V}} = v_{10} + v_{11} exp\left(-\frac{Fr_V}{v_8}\right) + v_{12} exp\left(-\frac{Br_V}{v_9}\right), \\ 0 < D_{P_{pl_V}} \qquad v_8, v_9, v_{10}, v_{11}, v_{12} \colon constants \tag{5-14}$$

- The coefficients $v_1 - v_{12}$ are dependent on the codec type, video format, key frame interval and video display size. Their provisional values for specific configurations may be found in the Appendix of the aforementioned recommendation or may be derived using a standard methodology, also described in the Annex of the recommendation. For completion, for the case studies defined in Table 10, we depict in Table 11 the values of the $v_i, i = 1,2, \dots 12$ coefficients.
- Formulas are also available for the speech quality estimation function as well (not presented here though for simplicity). Finally, the multimedia quality MM_q can be calculated using the speech quality S_q, the video quality V_q, as well as the speech and video delays.
- Extensions to this model have been proposed in literature, so that they are more realistic, and include extra factors such as: the packet loss pattern (different from random), the influence of video content, the effect of buffering [59], etc.

Table 10: Case studies for the derivation of the $v_i, i = 1,2, \dots 12$ coefficients.

Factors	#1	#2	#3	#4	#5
Codec type	MPEG-4	MPEG-4	MPEG-2	MPEG-4	ITU-T H.264
Video format	QVGA	QQVGA	VGA	VGA	VGA
Key frame interval (sec)	1	1	1	1	1
Video display size (inch)	4.2	2.1	9.2	9.2	9.2

Table 11: Values for the $v_i, i = 1,2, \dots 12$ coefficients.

Parameter	#1	#2	#3	#4	#5
v_1	1.431	7.160	4.78	1.182	5.517
v_2	2.228×10^{-2}	2.215×10^{-2}	1.22×10^{-2}	1.11×10^{-2}	1.29×10^{-2}
v_3	3.759	3.461	2.614	4.286	3.459
v_4	184.1	111.9	51.68	607.86	178.53
v_5	1.161	2.091	1.063	1.184	1.02
v_6	1.446	1.382	0.898	2.738	1.15
v_7	3.881×10^{-4}	5.881×10^{-4}	6.923×10^{-4}	-9.98×10^{-4}	3.55×10^{-4}
v_8	2.116	0.8401	0.7846	0.896	0.114

v_9	467.4	113.9	85.15	187.24	513.77
v_{10}	2.736	6.047	1.32	5.212	0.736
v_{11}	15.28	46.87	539.48	254.11	-6.451
v_{12}	4.17	10.87	356.6	268.24	13.684

5.1.2.2 Online adaptation of G.1070 E-model

In [60], an extension to the G.1070 model is presented, so that is also becomes a valid tool for online quality monitoring. The objective is to calculate the frame rate (Fr_V), bit rate (Br_V) and packet loss rate (P_{pl_V}) directly from the received video bitstreams. As long as these three parameters are known, and since the 12 aforementioned coefficients can be found, the video-alone quality estimation function V_q can produce real-time output.

The proposed model requires that specific desired features or data from video bitstreams encapsulated in each network packet are extracted, and then used to create statistics over an N-frame sliding window. These statistics are subsequently used as input to the above G.1070 model's formulas. The basic input parameters are estimated as follows:

$$P_{pl_V} = \frac{\#lost\ packets}{\#lost\ packets + \#received\ packets} \tag{5-15}$$

The number of lost packets may be calculated using the recorded discontinuities in the packet sequence numbers. The frame rate and bit rate are estimated as shown in the next equations:

$$Fr_V = \frac{reference\ clock\ frequency}{time\ interval\ between\ two\ adjacent\ frames} \tag{5-16}$$

$$Br_V = Fr_V \frac{\#bits\ received}{N(1 - P_{pl_V})} \tag{5-17}$$

All parameters are estimated during an N-frame sliding window, so that each output value depends on the N preceding frames.

We have conducted a parameter-study to graphically present the impact of video frame rate, video bit rate, and packet loss rate on the video quality V_q. Figure 22 demonstrates that V_q (and consequently, QoE) increases as the video frame rate increases up to a certain point, after which it starts decreasing (except for very high bit rates where it reaches an upper threshold). This pattern highlights that the video bit rate (i.e., the Br_V), and in sequence, the serving rate of the network, imposes a "bottleneck" to the number of frames per second that can be transmitted over the network.

Then, Figure 23 shows an exponentially decreasing MOS trend, as the packet loss rate increases, for most of the case studies. The slope of this decrease is a function of the case study considered for the derivation of the $v_i, i = 1,2, \ldots 12$ coefficients (see Table 10). This figure shows that the selected codec type, video format and display size play a combinatorial role in the final user perception, not easily leading to more conclusions. However, comparing case study 1 and 2, we may observe the phenomenon that smaller device screens (e.g., smartphones) seem to offer a better video quality to users than larger screens, for any packet loss rate. Moreover, comparing case study 4 and 5, we observe that ITU-T H.264 is less robust to high packet losses than MPEG-4, which might be the case because of its high compression rate.

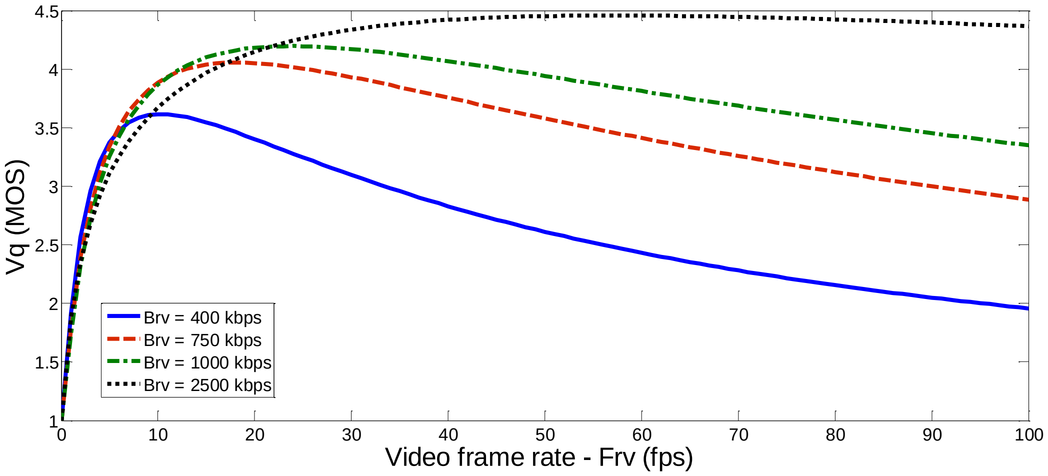


Figure 22: Impact of video frame rate on MOS.

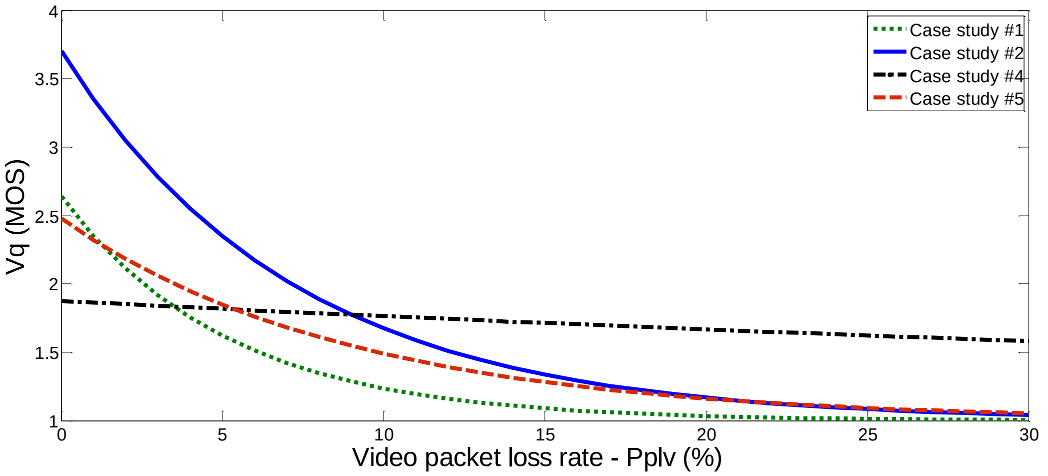


Figure 23: Impact of video packet loss rate on MOS.

Having studied standardized parametric QoE estimation for VoIP and real-time (online) video, next we move on to literature-based estimation methods for other popular services (namely, file download, web browsing, IPTV, video streaming - both conventional and adaptive, and Skype). Since standardization efforts are still ongoing for these common services, we present and analyze well-cited models from the research literature.

5.2 Literature-based parametric QoE estimation

In this section, we describe non-standardized parametric QoE models which can be used for a reliable estimation of QoE for various types of services, namely File Transfer Protocol (FTP) services, web browsing, lossy video streaming (IPTV), lossless video streaming (conventional and adaptive) and Skype applications. Per service type, we evaluate the impact of certain KPIs on the QoE.

5.2.1 Parametric QoE estimation for FTP services

The main characteristic of FTP services is that there is no need for a continuous and in-sequence packet arrival. Taking into account that the delay expected by the user is proportional to the size of the downloaded file and the fact that the FTP service is not adjusted in the application layer, the data rate is the dominant factor that affects the QoE level. More specifically, the model that provides MOS for an FTP service is as follows [61]:

$$MOS_{FTP} = \begin{Bmatrix} 1 & u < u^- \\ b_1 \cdot log_{10}(b_2 \cdot u) & u^- \le u < u^+ \\ 5 & u^+ \le u \end{Bmatrix} \tag{5-18}$$

where u represents the data rate of the correctly received data, i.e., $u = R \cdot (1 - P_{error})$, where R is the data rate and P_{error} the error ratio. The values of the b_1 and b_2 coefficients are obtained from the upper (u^+) and lower rate (u^-) expectations for the service. For instance, for $u^- = 8kbps$ and $u^+ = 315\ kbps$, it holds that $b_1 = 2.5037$ and $b_2 = 0.3136$, while the estimated MOS values are depicted in Figure 24.

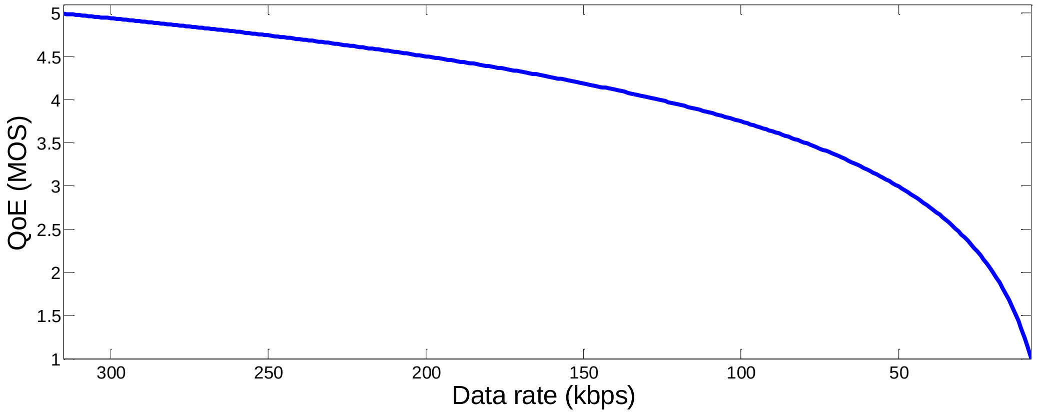


Figure 24: Estimated MOS for data rates in the range from 8 to 315 kbps.

5.2.2 Parametric QoE estimation for web browsing

The main observation for web browsing services is that the delay is the key QoE performance indicator. A long waiting time for the response of web will make users lose patience and negatively affect their perception for the provided service. Taking this into account, the model described in [62] is a suitable candidate for QoE estimation. This model is based on subjective validation tests, and the resulted empirical formula is as follows:

$$MOS_{web} = 5 - \frac{578}{1 + (11.77 + 22.61/\tau)^2} \tag{5-19}$$

where τ is the response time. In Figure 25 the estimated MOS for different response times is depicted.

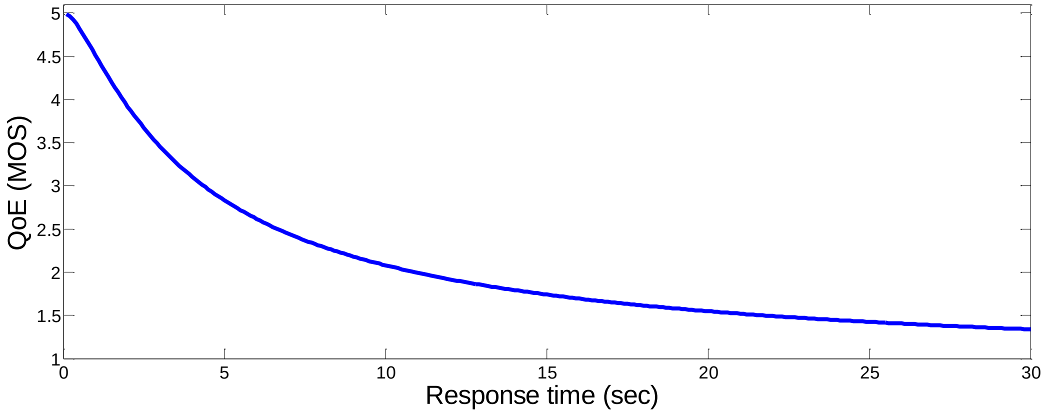


Figure 25: Estimated MOS for various response times.

5.2.3 Parametric QoE estimation for video streaming

Due to their increasing popularity, video services cause the majority of traffic over the Internet, while they are characterized by high resource requirements. Therefore, the estimation of QoE for video streaming applications becomes of notable importance. In this section, we study two different types of video streaming: a) IPTV, which is a lossy type of service, and b) VoD, which is a lossless service type. For the latter, we focus on the paradigm of YouTube, and further study it in two versions, namely adaptive and non-adaptive streaming over HTTP.

5.2.3.1 IPTV

IPTV is a common video streaming service. It is User Datagram Protocol (UDP)-based and is therefore prone to packet losses. In [63],[64] a MOS prediction formula is proposed for three video content types, named “Slight movement (SM)”, “Gentle walking (GW)” and “Rapid movement (RM)”. This formula considers the objective parameters Send Bitrate (SBR), Frame Rate (FR) and Packet Error Rate (PER):

$$MOS_{IPTV} = \frac{a_1 + a_2 \cdot FR + a_3 \cdot \ln(SBR)}{1 + a_4 \cdot PER + a_5 \cdot (PER)^2} \tag{5-20}$$

where the coefficients: a_1, a_2, a_3, a_4, a_5 are obtained by linear regression of the proposed model with the training set of video sequences. More specifically, the values depicted in Table 12 have been experimentally calculated.

Table 12: Typical values for the coefficients per video content type.

Coefficient	SM	GW	RM
a_1	4.5796	3.4757	3.0946
a_2	-0.0065	0.0022	-0.0065
a_3	0.0573	0.0407	0.1464
a_4	2.2073	2.4984	10.0437
a_5	7.1773	-3.7433	0.6865

In Figure 26 we consider the values included in Table 12 and $FR = 30\,fps$ to estimate MOS for various PER and SBR values. As depicted in this figure, there is a decrement of the MOS value for increasing PER values, while this decrement is strongly correlated with the SBR value. More specifically, as it can be observed in Figure 26, for the selected FR value ($FR = 30\,fps$) a stabler performance is achieved when $SBR = 25\,kbps$. For higher SBR values (e.g., $SBR = 50\,kbps$) a linear degradation of the MOS is observed, leading to low MOS performance for high PER values. Additionally, for lower SBR values (e.g., $SBR = 10kbps$) the degradation of the MOS as the PER increases is exponential, leading very fast to low MOS values.

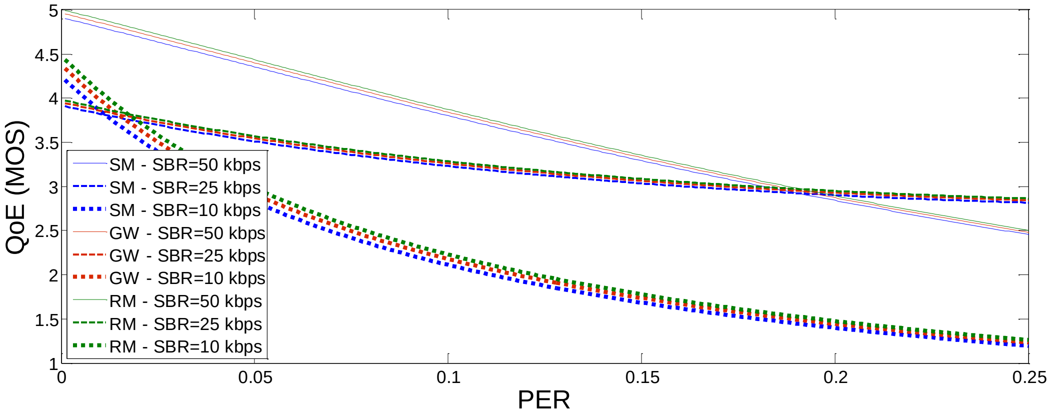


Figure 26: Estimated MOS for various SBR values and increasing PER.

5.2.3.2 YouTube - conventional streaming

Another type of video content delivery that deserves attention is that of streaming pre-encoded video, i.e., VoD. YouTube is the most popular paradigm in this category. YouTube is not subject to packet losses, since the connection is Transmission Control Protocol (TCP)-based. The following description regarding YouTube QoE analysis is based mostly on [65], the authors of which demonstrate an extensive literature on the topic. The most popular models of YouTube QoE are built based on subjective experiments, conducted either in controlled laboratory environments or using crowdsourcing tests and field studies. Through these well-designed experiments, the system-level key influence factors that affect the YouTube video delivery quality may be found, which are:

- Number of stalling events, N, where the term stalling refers to the interruption of video playback that occurs when the playout buffer runs out.
- Duration of stalling events, L.
- Total video duration, T (significant is the total stalling duration over the whole video duration).
- Initial delay, which refers to the inherent delay at the beginning of each streaming session, i.e., the video start-up delay, and it is necessary in order to fill a part of the buffer up to a threshold after which, playback starts.

Regarding these influence parameters, some important findings, that follow up from these subjective experiments are:

- The number of stalling events together with the stalling length (i.e., the stalling pattern) are clearly dominating the user perceived quality.
- Initial delays have almost no influence on MOS for videos of duration 60 sec and 30 sec, namely they are tolerated up to a reasonable level.
- User ratings are statistically independent from video characteristics such as resolution, video motion, content type, encoding scheme and video bit rate. The stalling pattern is what really influences the user's experience.

Especially for YouTube, the mapping function proposed in [65] follows the IQX hypothesis and has the form:

$$MOS_{YouTube} = \alpha \cdot e^{-\beta(L) \cdot N} + \gamma \tag{5-21}$$

where α, γ and $\beta(L)$ are coefficients derived from the experimental process. More specifically, $\beta(L)$ has a linear relation with the stalling duration L, which is defined as: $\beta(L) = 0.15 \cdot L + 0.19$. Typical values for α and γ coefficients are $\alpha = 3.5$ and $\gamma = 1.5$.

In Figure 27, the MOS is depicted for various numbers of stalling events and stalling durations. As it can be observed from this figure, the number of stalling events is the dominant factor that affects MOS. Also, for a stalling duration of up to a point, the degradation of MOS is fast, while after that point the degradation of MOS is slower. This is a reasonable result, since when the MOS reaches a very low value of about 1.5, the effect of a longer stalling duration is hardly considered by the users anymore.

QoE metrics tailored to the YouTube application have also been defined. These metrics can be used instead of the MOS scale to get an indication about the quality of the user viewing experience. For instance, the reception ratio, ρ, is calculated as follows [65]:

$$\rho = \frac{Download\ throughput\ or\ Bottleneck\ capacity}{Video\ encoding\ rate} \tag{5-22}$$

Although the reception ratio cannot be directly related to QoE, it is a good indicator about whether there are problems in the network. If $\rho > 1$, the video has good quality, otherwise poor.

Moreover, rate λ gives a good indication of YouTube video delivery quality according to [66], and its value should ideally be zero or close to zero:

$$\lambda = \frac{total\ stalling\ time}{total\ video\ elapsed\ time} \quad (5\text{-}23)$$

Following a similar logic, [67] proposes the following metric:

$$\rho' = \min_{t \in (0,T)} \frac{b(t)}{B}\frac{\Delta}{t} \quad (5\text{-}24)$$

where:

- t is the instantaneous download time.
- T is the total download time.
- Δ is the video duration.
- B is the total video size in bytes.
- $b(t)$ are the bytes downloaded so far.

In this case, $\rho' > 1$ indicates a stalling-free video, while $\rho' < 1$ implies a non-seamless video session.

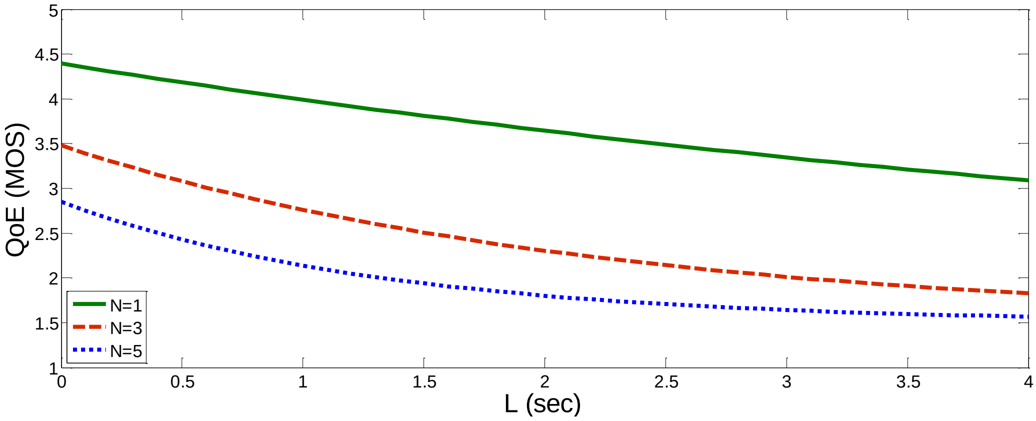


Figure 27: Estimated MOS for various number of stalling events and stalling durations.

5.2.3.3 YouTube - adaptive streaming

In adaptive streaming scenarios, a video file is broken into multiple segments, while each segment is available at different quality levels. These levels may differ in video bit rate or in the video resolution, etc. Then, each user independently requests the next segment in a specific quality level, based on the user's current perception of available bandwidth for this session.

For the case of HAS, [68] proposes a simple but highly accurate QoE model, that is:

$$MOS_{HAS} = 0.003 \cdot e^{0.064 \cdot t} + 2.498 \quad (5\text{-}25)$$

where t is the percentage of time that the video was being played out at the highest layer.

Based on this formula, the QoE of HAS applications depends mainly on the fraction of time that the highest layer is being played out over the total viewing time. Moreover, as it can be seen in Figure 28, the MOS is bounded by the quality that can be achieved by the highest and lower layers (4.3 and 2.498 respectively). Moreover, the percentage of time at each quality layer that the user spent watching a video is another meaningful KPI, strongly correlated to the resulting video bit rate [69].

Another important influence factor of HAS QoE based on [68] is the “adaptation amplitude” (or “altitude”), which refers to the gap between two subsequent quality levels. In the case that the highest and then the lowest quality levels are sequentially selected (or vice versa), the amplitude will be high and the QoE impression will be low; if, however, such intense switches are refrained, the amplitude will be lower. The higher the amplitude, the worse the perception of the overall quality at the user.

Moreover, some additional quality influence factors, with lower impact though, are the frequency of switches (i.e., adaptation events) and their direction. Last but not least, it has been shown that the buffer length of the user’s application and the size of the segment encoded at the server’s side play a significant role on QoE [70].

For the case of HAS, additionally the “activity factor” metric proposed in [67] applies:

$$a = \frac{total\ time\ of\ actual\ data\ download}{total\ time\ elapsed\ for\ complete\ video\ download} \tag{5-26}$$

If this metric is close to 1, it means that the client was “struggling” to download each segment on time; however, if this factor is much lower than 1, it means that the client had sufficiently available bandwidth and could even afford higher video resolutions, if those were available. Note that gaps in the video download are occurring because the client is not buffering the full content at once but is just targeting to maintain an acceptable buffer threshold.

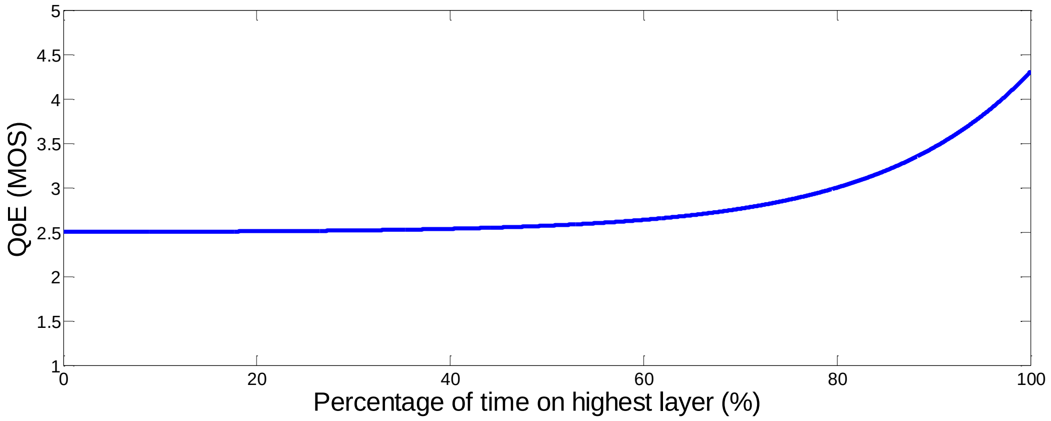


Figure 28: Estimated MOS for different percentages of time on highest HAS layer.

5.2.4 Parametric QoE estimation for Skype

For Skype applications, a practical QoE estimation approach can be found in [71]. The proposed model has been derived by measurements conducted on Skype video calls. It has been found that three types of resolutions are available, namely 160x120, 320x240 and 640x480. Moreover, the maximum frame rate is 35 fps.

Then, the MOS level for this service type is as follows:

$$MOS_{skype} = \left\{ \begin{array}{cc} 1 & res = 160x120 \\ 2 & res = 320x240 \\ 3 + \dfrac{FR}{35fps} + (2 \cdot I - 1) & res = 640x480 \end{array} \right\} \tag{5-27}$$

where I is the image quality ranging from 0 (worst) to 1 (best) and FR is the Frame Rate.

In Figure 29, we evaluate the relationship between Skype QoE and the image quality. Moreover, we vary the FR to study its impact. As expected, MOS degrades linearly while the image quality is reduced, while the FR also has a significant influence on the perceived QoE.

Based on this model, the authors in [71] also propose an adaptation mechanism of the Skype application to poor network conditions. Assuming a maximum acceptable threshold of the packet delay, if this threshold is reached, the Skype application starts to gradually degrade first the frame rate, then the image quality and finally, if required, the resolution. This adapting behavior helps sustain a viable and meaningful communication between two Skype applications, compromising on the quality though.

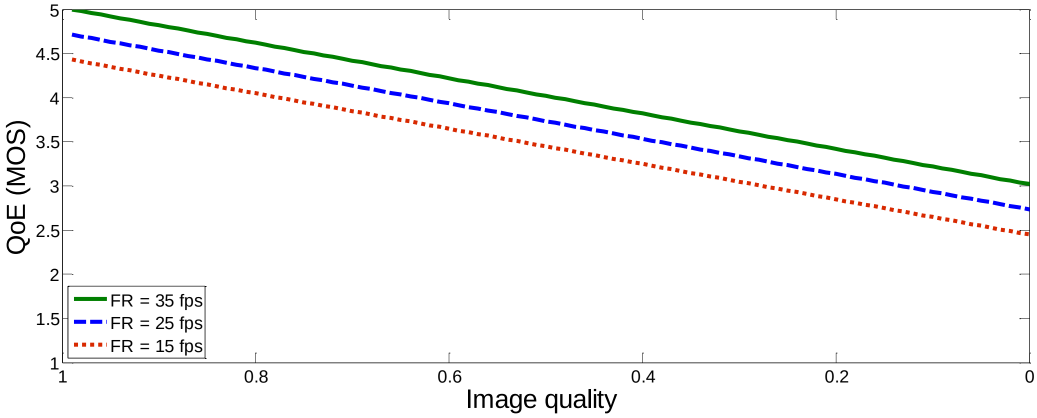


Figure 29: Impact of image quality and frame rate on Skype MOS.

Having described both standardized and literature-based QoE estimation models, next we move on to summarized results and potential research and exploitation directions.

5.3 Summarized results and exploitation directions

The parametric QoE estimation models described above define a major set of formulas that can be exploited by academia and industry to understand how the users perceive the quality of a provided service. Summarizing the study in the previous sections, in Table 13 we indicate the key parametric QoE estimation models available in the literature and list the MCPs and KPIs that affect the QoE performance per service type.

Table 13: Parametric QoE estimation per service type.

Service type	QoE estimation model	MCPs and KPIs
File transfer	Data rate-based formula [61]	Data rate, expected upper and lower data rate
Web browsing	Response time-based formula [62]	Response time
Skype	Skype-specific formula [71]	Frame rate, image quality, resolution
VoIP	ITU-T Rec. G.107, E-model [55][56][57]	Packet loss ratio, delay, codec, coding rate
Video streaming	IPTV model [63][64]	Data rate, frame rate
	YouTube (conventional) model [65]	Number of stalling events, duration of stalling events, video duration
	YouTube with adaptive streaming model [68][70]	Time on highest layer, amplitude, frequency of quality switches

Online video	ITU-T Rec. G.1070, E-model [58]	Packet loss ratio for audio and video packets, relative delay between video and audio packets, data rate, frame rate, monitor size

From an academic and research perspective, through a clear collection of parametric QoE models, an easy and straightforward “translation” of QoS research works to QoE vocabulary may be applied. To be more specific, a potential research direction that can be aided by this study is the direct quantification of the impact of existing research works on QoE. This may be possible either via the realization of appropriate QoE estimation models (column 2 of Table 13) or by quantifying a potential improvement on specific MCPs and KPIs per service (column 3 of Table 13). Furthermore, the collection of KPIs helps identify the specific influence factors that play the most important role on the user’s perceived quality, guiding in this way future works towards devising network and application mechanisms that target at improving exactly those factors.

Regarding the impact of explicit QoE parametric models on the industry sector, this is twofold. On the one hand, it can help operators design their networks in a QoE- rather than QoS-meaningful way. That is, the operators are guided to give emphasis on designing and maintaining their networks in such a way that requirements regarding the KPIs per service are met (e.g., through a QoE-meaningful resource provisioning, a proper positioning of network servers and gateways - e.g., close to the user, etc.). Furthermore, attention on the per-service KPIs has to be given during the network management process, i.e., during the network’s real-time operation. Mechanisms such as scheduling, mobility management and power control can be tuned so that proper weight is given on the actual QoE impact factors per service. In this way, an indirect QoE improvement will be achieved through the targeted enhancement of carefully selected QoS parameters. What is more, if we consider the recently emerged paradigm of “User Provided Networking” (UPN), like the one proposed in [72], a massive potential is unlocked. According to this paradigm, users are actively involved not only in service evaluation tasks by providing feedback about the experienced quality either passively (e.g., device capabilities, response times, context of use) or actively (e.g., MOS feedback), but they can also participate in the service provisioning loop by becoming “micro-providers”, given the proper incentives.

Another important, even though less obvious capability exposed by the collection of the different KPIs per service, is the opportunity to achieve a more meaningful cross-service resource provisioning towards a) higher QoE, and b) higher resource utilization. All services are currently competing for the same resources on an equal basis; nevertheless, it would make more sense to allocate the limited resources in a service-dependent rather than in a service-oblivious (i.e., blindly fair) way. This may be possible a) by performing the scheduling process on a per-flow basis (e.g., prioritize a more delay-critical service with respect to another), or b) by optimizing the sum QoE in a cell by taking appropriate cross-service management decisions. Regarding the latter, a potential enabler is to exploit the adapting behavior of Skype or HAS applications and provoke a deliberate quality degradation at specific Skype or HAS flows, so that resources are moved to other applications, with QoE/KPIs currently at a critical level. The goal would be to keep all users’ QoE above a critical threshold, or, to achieve a maximum possible summed QoE. Note, that average QoE values per cell are not definitely appropriate indicators of quality though (e.g., a cellular MOS of 3 may be a result of user1’s MOS=1 and user2’s MOS=5); on the contrary, each flow should be treated independently, or, at least, standard deviations should be considered as well (see [73]).

5.4 Conclusions

As we are moving closer and closer to future network generations, the human factor is becoming the epicenter of attention and the driving force for the network design. Thus, the comprehension and, in extension, the control of the provisioned QoE to the users has become a necessity for network operators. Parametric QoE estimation models are a prerequisite for this purpose. They constitute the ideal tools towards live network quality monitoring and, hence, QoE management. Nevertheless, despite the increased interest from academia and industry to push towards a QoE service provisioning model, a clear/comprehensive manual on the available parametric models and the critical QoE performance parameters per service type is currently missing. Identifying this gap, this chapter aspires to become a thorough and handy “manual”, currently absent from the literature, that identifies and describes appropriate parametric models for popular services nowadays, such as YouTube, Skype and IPTV, as well as describes and studies standardized ones. Therefore, the current study may become a stand-alone, useful tutorial both for researchers and operators, who are interested in moving from the pure technical QoS-domain to a more meaningful QoE-domain, so that they can understand and influence the impact of their network decisions on the final recipient, the end-user.

6. QoE-DRIVEN DEVICE-TO-DEVICE COMMUNICATIONS

Device-to-Device (D2D) communications are planned to become an indispensable part of future mobile cellular networks. A lot of attention has been paid to this new communication paradigm, due to the important benefits it brings for both cellular operators and users. Under this perspective, and realizing that the main asset of D2D is the potential enhancement of the user experience, we propose a QoE-driven framework for the management of this type of communications. With this objective, the QoE management cycle described in Chapter 3 is customized to serve a QoE-driven version of D2D communication setup. Simulation results show that this framework is able to capture and enhance the overall experience of mobile users, and, thus, allow for proportionate financial benefits for network operators.

6.1 Introduction to D2D

The concept of D2D has emerged over the last years as a promising add-on feature not only of 4G mobile networks (mainly LTE/LTE-A), but also of future 5G technologies. It refers to the new communication paradigm, where two cellular UEs exchange data directly, without the intervention of the base station (eNB). Nevertheless, the exchange of control information is traditionally handled by the operator's central devices [74].

D2D communication may be classified using various criteria, as illustrated in Figure 30. Depending on the type of spectrum used, we distinguish inband and outband D2D, which utilize licensed and unlicensed spectrum, respectively [75]. Furthermore, D2D may work as an underlay to the standard cellular operation by reusing resources with standard cellular users, unlike the overlay mode, where specific dedicated resources are either statically or dynamically assigned exclusively for D2D operation. Regarding the level of control of the operator in the D2D setup procedure, we find autonomous and controlled schemes. Moreover, regarding the initiation of the D2D communication request, there are two options: Either the D2D request is fully transparent to the user, whose communication is automatically switched from cellular to D2D mode by the operator (network-originated), or the D2D mode is originally (explicitly) requested by the user (user-originated). Finally, D2D transmissions may be unicast or multicast/broadcast, where the former case describes peer-to-peer links for direct communication or relaying links (e.g., for coverage extension purposes), while the latter would be more appealing for social and commercial applications, such as proximity-based advertisement or public safety scenarios.

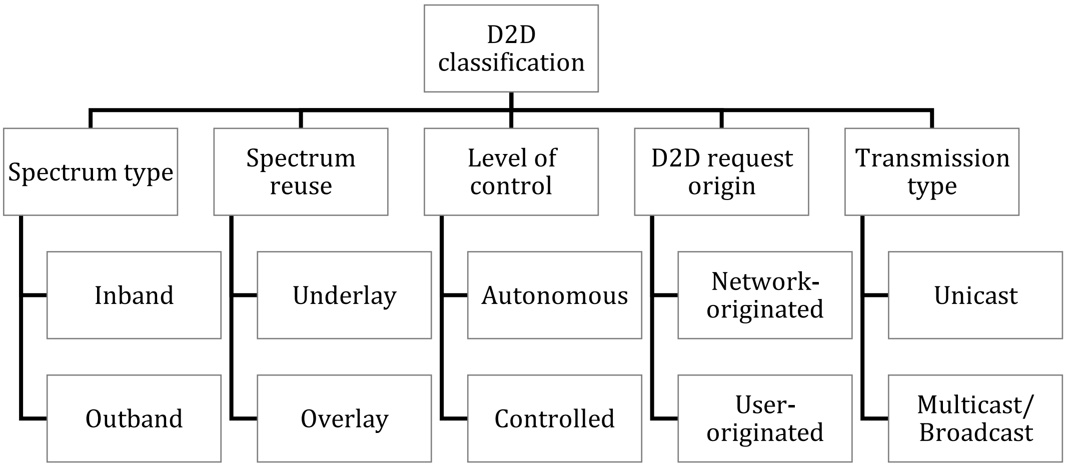


Figure 30: D2D classification types.

On the one hand, D2D communications are driven by the operators' need to utilize their current infrastructure more efficiently in terms of spectrum, processing resources, and network load. Another major driving force is the operators' need to find a profitable Peer-to-Peer (P2P) competitor to the popular free-to-use Wi-Fi Direct, with D2D being an appealing candidate for that purpose.

On the other hand, this new technology brings proportionate benefits for the users as well. D2D may result in a more efficient reuse of network resources, which in turn guarantees higher data rates and increased total user capacity. Moreover, bypassing the eNB, during the direct user data exchange between two UEs, enhances the transmission quality and reduces the communication delays, not only due to the devices' physical proximity, but also because only one directional transmission needs to be scheduled, instead of both uplink (UE to eNB) and downlink (eNB to UE) directions. Finally, the UEs use less battery power, as a result of the communicating entities' proximity, which is a crucial issue for mobile handsets.

The first goal of the current study is to investigate whether the possible gain of switching from cellular to D2D operation, is also reflected to QoE terms. Encouraging results in terms of QoE improvements indicate a huge marketing asset for operators, who can then advertise D2D technology as an experience-enhancing service and charge it accordingly. A specific QoE-based charging model is therefore proposed as well in Section 6.4.

The second objective is to develop and examine a QoE-aware management framework for controlling the transition of cellular links to D2D links or vice-versa, driven by the user's benefit. The proposed D2D network management framework is integrated into an LTE-A system, in accordance with recent standardization activities.

6.2 QoE-centric network management

Recent research works have turned their attention to QoE-centric approaches of network control. For the case of mobile networks, QoE-driven management techniques include radio resource allocation, mobility management, battery consumption optimization, service optimization, etc. For instance, in [76], a QoE-aware handover scheme for seamless and optimized support of users running multimedia applications in heterogeneous networks, called “QoE Hand”, is proposed. This approach ensures “always-best” connectivity, which has a significant impact on the user perceived QoE, especially during congestion periods. Similarly, in [77], a QoE-driven mobility management technique exploits QoE-awareness to initiate or assist vertical handover decisions, in the context of Mobile IP.

Service management approaches that rely on QoE-awareness also spread in other functionalities, such as a) network routing functions, where adaptive routing protocols enhance the customer experience while optimizing network resources' usage [78], b) new power allocation techniques that maximize the overall QoE subject to the total transmit power constraint [79], and c) advanced CEM techniques through enhanced charging schemes that also account for QoE-based intelligence [80]. A QoE-driven selection mechanism for controlling the mode of operation of the links inside a cell is herein proposed to contribute to the area of QoE-centric network management. The proposed mechanism harnesses benefits both from a network- and a user-centric perspective.

D2D communications are planned to become one of the major components of future cellular networks, but standardization efforts are still ongoing. Presently, the most dominant criterion considered for the switchover decision between D2D and cellular mode is *throughput*, e.g., [74]. However, there is no linear dependency between QoS factors, such as throughput, and the perceived QoE, as already elaborated in Chapter 2. Hence, a D2D scheme triggered by QoE values, unlike or complementary to existing

QoS criteria, is the closest to the user's benefit decisive factor for selecting between cellular and D2D operation modes.

Under this perspective, we provide a techno-economic framework for QoE-based D2D support inside the network, considering both technical and business issues. First, in Section 6.3, we describe the system model requirements, focusing on the network entities involved, their operations and signaling. Also, we discuss the QoE management cycle in terms of data collection, modeling and management. Next, in Section 6.4, we propose a charging scheme suitable for this framework, which is both fair for the users and profitable for the operators. Finally, in Section 6.5, we provide simulations to show the validity and benefits of this model, followed by the conclusions in Section 6.6.

6.3 Technical system requirements

6.3.1 System model

D2D communications operate as an add-on layer to standard cellular communication networks, in the sense that they are tightly integrated in the existing infrastructure and utilize licensed spectrum resources [74]. Normally, when two UEs located in the same cell want to communicate, e.g., UE_1 and UE_2 in Figure 31, all the control and user data of their entire communication have to pass through the eNB (uplink-UL). Afterwards, they enter the Evolved Packet Core (EPC) nodes Serving Gateway (S-GW) and Packet Data Network Gateway (PDN-GW), they then follow the reverse route back to this eNB and any other eNBs located in the same tracking area, before the target receiver (UE_2) is finally paged and starts to receive the data (downlink-DL) [81]. This inevitable waste in access and core network resources (signaling, spectrum, energy, network load, processing and memory requirements), which derives from the fact that the user data have to follow this entire route despite the sender's-receiver's proximity, is exactly what has triggered the interest for the introduction of D2D communications, which allow direct data exchange (e.g., $D2D_1$-$D2D_2$, Figure 31).

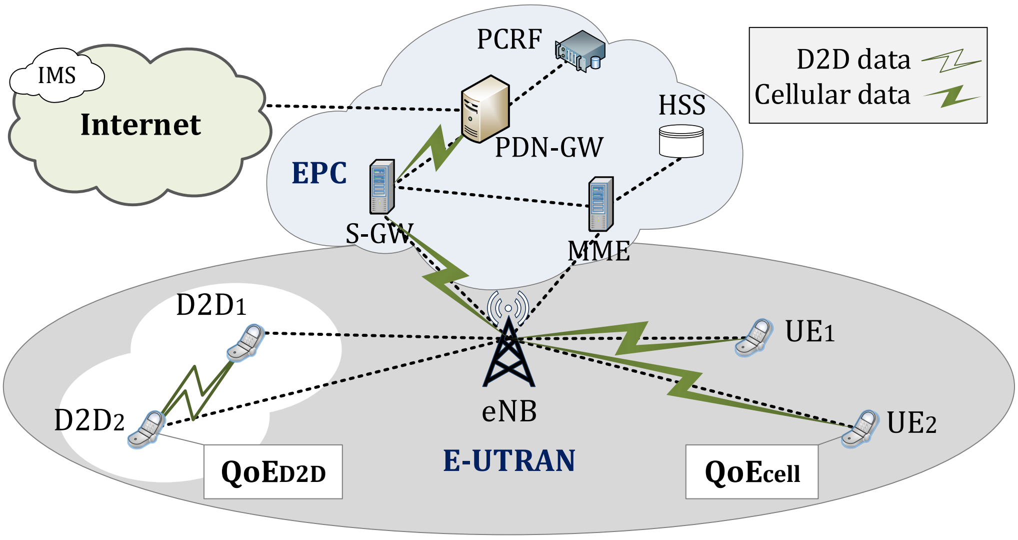


Figure 31: The D2D communication paradigm.

D2D devices are standard LTE-A UEs, enhanced to support the D2D mode. Enhancements are also needed in the core network (PDN-GW), as well as the eNBs to allow for D2D communication setup and management. The PDN-GW is responsible for sniffing network traffic (IP headers of arriving packets) in order to identify data transported between UEs belonging to the same or even neighboring cells, indicating a potential D2D link, while the eNB is responsible for triggering a D2D link establishment check [74]. As long as a D2D switchover is successful, all user data are exchanged via a direct path, bypassing the eNB and EPC. Nevertheless, the D2D link might break anytime if no longer considered advantageous, in which case the communication seamlessly continues via a traditional cellular link. Either mode of operation (cellular or D2D) manages to achieve a different QoE score (Figure 31).

D2D links may raise interference issues to the standard cellular operation and to other parallel D2D links. To avoid this, either advanced interference management schemes need to be deployed, which also foster the utilization of resources, or dedicated spectrum has to be devoted per link. We adopt the latter solution, since spatial spectrum reuse is not a target of this work. Hence, we consider inband, i.e., licensed D2D communications, operating as an *overlay* to the LTE-A network. Even though D2D UEs could potentially transmit at maximum power, this would negate the energy efficiency gain that comes with D2D. Thus, it is preferable that D2D devices use lower power that enables local connections. Specifically, we select a transmission power of -19dBm that supports a D2D range of around 50m.

6.3.2 QoE-driven D2D mode selection

Although the criteria used to switch on D2D links may vary, in this study we consider the receiver's *QoE* as the decisive factor. Then, the proposed QoE-aware D2D management framework consists of the next steps, also depicted in Figure 32:

1. Standard cellular communication is initiated: the UE transmitter makes a scheduling request and the eNB assigns resources, used for the uplink transmission. This communication path goes through the EPC network.
2. The QoE of the existing cellular link, referring to a preceding time interval, is estimated and reported to the attached eNB. Since QoE-awareness is expected to become an integral functionality of future systems, we implement it periodically during standard cellular operation.
3. Potential D2D traffic is identified by the PDN-GW and indicated to the eNB. The PDN-GW ensures at this point that the policies regarding this communication type are respected (via the Policy and Charging Rules Function - PCRF); for instance, whether the user has paid for this service and thus is allowed to use it.
4. A proximity discovery procedure is triggered by the eNB, to judge the feasibility and potential advantage of establishing a D2D link. For this purpose, the eNB orders that, prior to their next transmission, the two communicating entities perform a D2D test. Thus, it instructs the UE sender to transmit an eNB-determined signature (pilot packets) at indicated resources and the target UE to listen for this signature at the defined resources, in order to conclude on the reception quality.
5. The QoE of the potential D2D communication is indeed estimated using the directly exchanged pilot packets between the two users and is reported back to the eNB.
6. If the D2D test reveals a higher QoE for the receiver than the one reported in step 2, then the cellular link switches to D2D. So, a new D2D bearer is established, upon eNB's request, while still maintaining the original bearer linking the UEs to the GW. The eNB informs the PDN-GW that the D2D link is feasible, not only for the bearer establishment procedure, but also for continuous validation of the charging and policy requirements.
7. After this point, the scheduling of the data is still controlled by the eNB, but the user data are directly exchanged on the *uplink* direction, as described in [82].
8. Periodic D2D QoE monitoring is performed by running quality estimations on the packets exchanged directly via D2D, corresponding to real communication traffic.

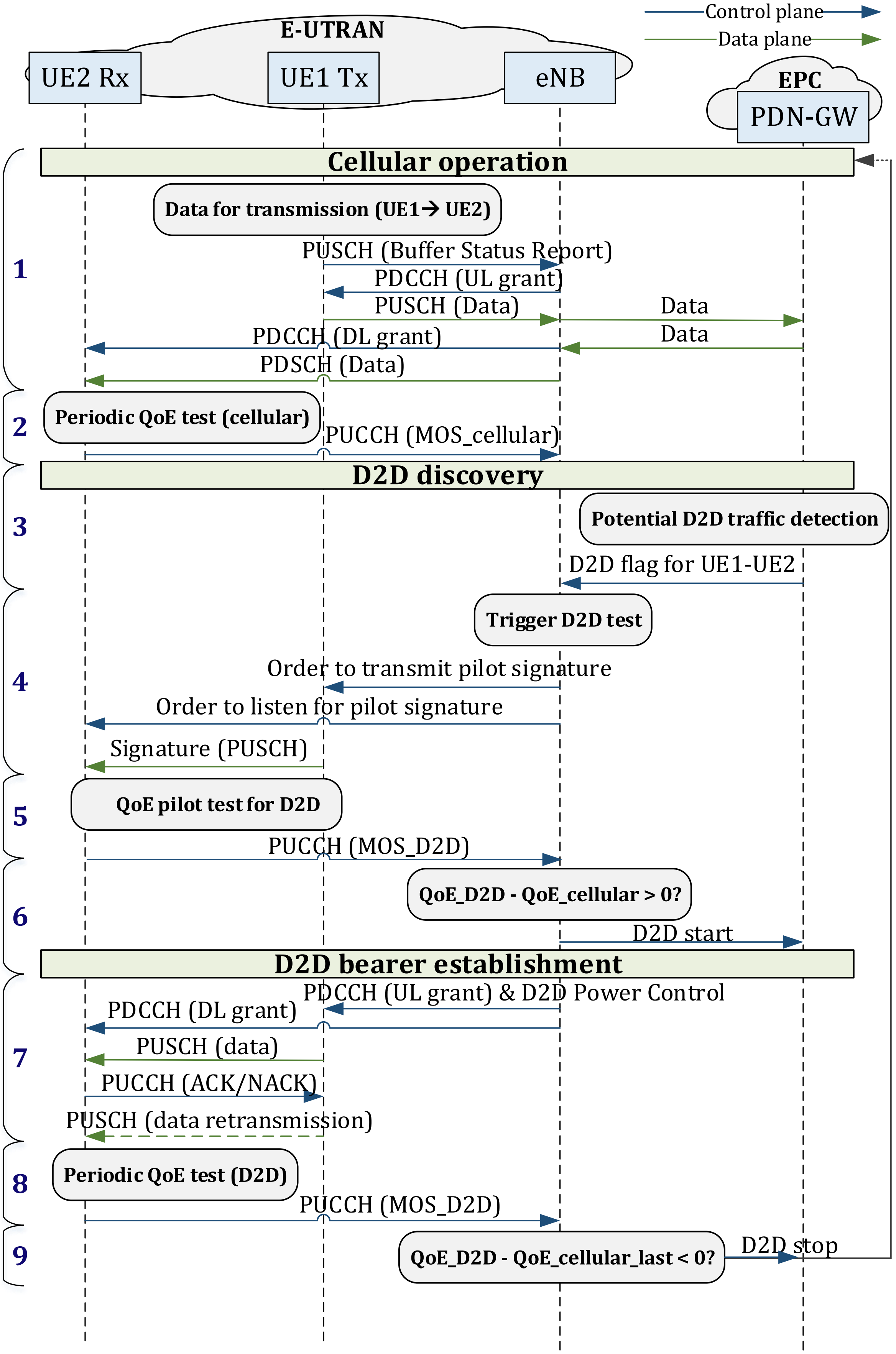


Figure 32: LTE-A signaling for QoE-driven D2D management.

9. If, at any point, the QoE of the direct connection drops below the last recorded cellular QoE (during step 2), D2D is considered no longer viable and the link falls back to cellular mode.

In Figure 32, the previous steps are incorporated into an LTE-A system [83]. Hence, Physical Downlink/Uplink Control Channels (PDCCH/PUCCH) are used for transporting the control information (scheduling, QoE reports, etc.), while Physical Downlink/Uplink Shared Channels (PDSCH/PUSCH) are used for carrying user data.

6.3.3 QoE-management supporting framework

As elaborated in Chapter 3, QoE provisioning requires the implementation of three major functions that comprise the QoE management cycle, as in Figure 33. For convenience, these functions and respective components (mapped to Figure 12) are summarized next.

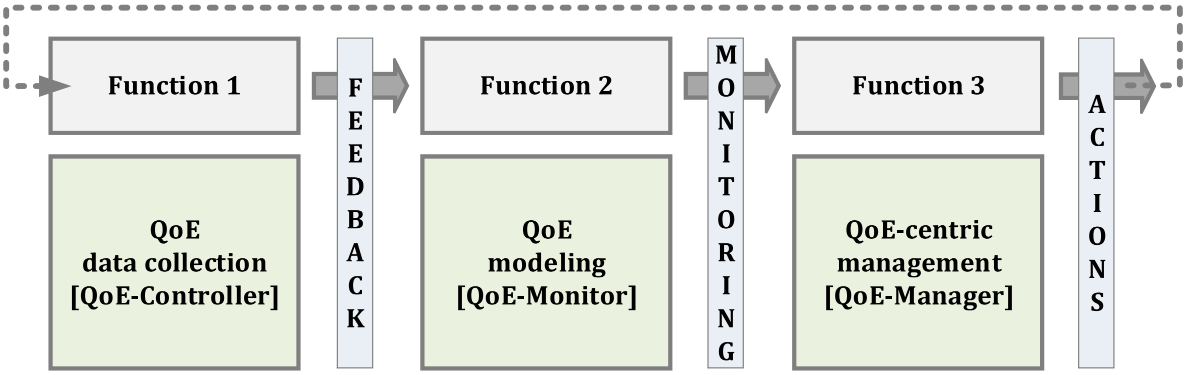

Function 1
QoE data collection [QoE-Controller]
FEEDBACK
Function 2
QoE modeling [QoE-Monitor]
MONITORING
Function 3
QoE-centric management [QoE-Manager]
ACTIONS

Figure 33: Generic QoE provisioning framework.

- **QoE data collection (in QoE-Controller):** The collection of QoE-related metrics is a vital function that provides the required input for quality estimation. The acquired information needs to be transferred as feedback to another entity inside the network, the QoE-Monitor, where the QoE modeling function is implemented.
- **QoE modeling (in QoE-Monitor):** This function implements the logic of the quality estimation function. The model output, commonly measured using the MOS scale, needs to be constantly monitored against recommended values.
- **QoE-centric network management (in QoE-Manager):** This function decides and triggers the network's corrective mechanisms that will improve the provisioned QoE. Any control actions have to be disseminated back to the network and be delivered to the affected nodes (here: D2D link setup).

Below, details on these functionalities are described under the proposed QoE-driven management framework for D2D communications.

6.3.3.1 QoE data collection

The collection of QoE-related information in a mobile cellular environment, such as LTE-A, imposes several challenges, which include the positioning of probes in the network, the type of input information collected, the delivery of this input to the quality estimation model, the periodicity of this procedure, etc. Under the proposed QoE-driven D2D management framework, the collection of information is implemented via passive probes inside the user terminals, i.e., exclusively in the Evolved UMTS Terrestrial Radio Access Network (E-UTRAN). This is a good practice, because unlike core network-based measurement approaches, the degradations caused by the wireless hop are also considered, thus providing an end-to-end indication of the achieved quality.

Two parameters are collected, the selection of which is justified later. The first is the average delay associated with the transmitted packets. The receiving UE can indeed examine timing information extracted by received packets in order to calculate the average packet delay during data transfer. The second parameter is the packet loss rate. This is estimated as the number of erroneously received packets over the aggregate number of transmitted packets, throughout the QoE reporting period. Erroneously received packets produce Negative ACKnowledgments (NACK) by the receiver, as in Figure 32. At short time intervals, these two parameters are converted to a MOS at the receiver side, using the QoE model described next.

6.3.3.2 QoE modeling

For our analysis and for the purposes of real-time quality monitoring, we assume VoIP traffic and select the ITU-T G.107, i.e., the E-model. Specifically, we adopt the E-model’s simplified version, as described in Section 5.1.1.2, which provides a formula for *R* that can be used for the online, i.e., in-service computation, of VoIP transmission quality. This formula is repeated here for convenience:

$$\begin{gathered} R = 94.2 - [0.024d + 0.11(d - 177.3)\mathrm{H}(d - 177.3)] \\ -[11 + 40\ln(1 + 10p)] \end{gathered} \tag{6-1}$$

where d is the average packet delay, p is the packet loss rate and $H(x)$ is the Heaviside step function ($H(x) = 0, if\ x < 0$ and $H(x) = 1, if\ x \geq 0$).

The reason for selecting this model is first of all its simplicity and suitability for real-time quality monitoring of interactive VoIP applications. Moreover, the input required can be easily collected from network entities located at the E-UTRAN (the UEs), as described in the previous section, without the need for complex signaling mechanisms.

The derived reports on quality (MOS reports) need to be signaled to the eNB on a periodic, eNB-defined basis. Therefore, we propose that QoE reporting about each connection comprises an extra procedure to the already standardized UE feedback procedures, namely the Channel Quality Indicator (CQI), Rank Indicator (RI) and Precoding Matrix Indicator (PMI). Regarding QoE reporting during the proximity detection phase for potential D2D establishment, this procedure will be aperiodic, triggered by the eNB.

6.3.3.3 QoE-centric network management

This procedure implements any control actions carried out by the eNB, as these have been already described in Section 6.3.2. In brief, the network management function consists of triggering the D2D pilot tests, collecting and processing the periodic QoE reports provided by the users, as well as controlling the transition of the communication from cellular to D2D mode or vice versa. The decision for switchover to D2D is indirectly delivered to the involved UEs via the UL and DL resource allocation grants, i.e., a map regarding where to transmit and receive respectively, as well as via the D2D power control order (Figure 32, Step 7).

Due to this scheme, and specifically due to the D2D tests and QoE reports, extra load is imposed on the network. Therefore, in order to support D2D communications, the network must be able to withstand this overhead in both the control and data planes. An effort must be made to ensure that the eNB collects enough up-to-date data to guarantee optimal mode switching decisions, while at the same time trying to minimize this overhead. More advanced mode switching plans (not considered here though) should also account for the possible ping-pong effect between the two modes of operation, in situations where the devices happen to experience roughly equal connection conditions in both modes. A potential drop in QoE due to this issue could be handled, for instance, by enforcing a no-switching time window after each mode selection, ensuring that it cannot occur again for a given amount of time. Alternatively, a weighted average MOS over larger time windows could be used, to account for both past and current MOS values.

6.4 QoE-aware charging model

Devising a proper charging model is a prerequisite for launching the D2D technology into the market, and thus, in this section, we present a model for charging the D2D users. More specifically, we propose the addition of a charging functionality to the proposed QoE support framework described previously. This framework, and particularly the QoE-centric network management component, is responsible for feeding the charging model with the required input, i.e., the excess QoE offered to users operating on D2D mode. This information can be provided through the “D2D start/stop” messages (depicted in Figure 32) sent from the eNB to the EPC, where the charging estimations take place.

Charging for D2D is justified because an enhanced user experience is offered, in terms of throughput, QoE, battery consumption, etc. Somehow, these overall advantages need to be quantified and charged accordingly. Possible charging may be of “pay-as-you-go” type, namely based on the data volume or duration of the communication session. Alternatively, fixed pricing schemes are possible, e.g., on a monthly basis. However, such schemes do not reflect the enhanced D2D-caused user experience.

Therefore, we propose the adoption of a QoE-based scheme, as a fair mechanism of charging for D2D. The price may be estimated using the difference between the QoE score actually offered through the D2D link minus the QoE score that would be offered by the cellular link, provided that this difference is nonnegative (if it were nonnegative, a D2D link would not have been setup in the first place). So:

$$Charge = f(QoE_{D2D} - QoE_{cellular}) \mid QoE_{D2D} > QoE_{cellular} \tag{6-2}$$

Hence, using the characteristic example of Figure 34, the user’s charge is quantified using the shadowed area created between the two curves, i.e., until the instant t_2. In fact, even fairer would be a scheme where providers charge only for the add-on MOS that exceeds both the offered cellular QoE (dashed curve) and a threshold (dotted line) that represents the minimum acceptable QoE (in Figure 34, equal to 3.5). In this way, providers will not inflict charges for improving the QoE up to this threshold through D2D, in the unfortunate event that the originally offered quality was less.

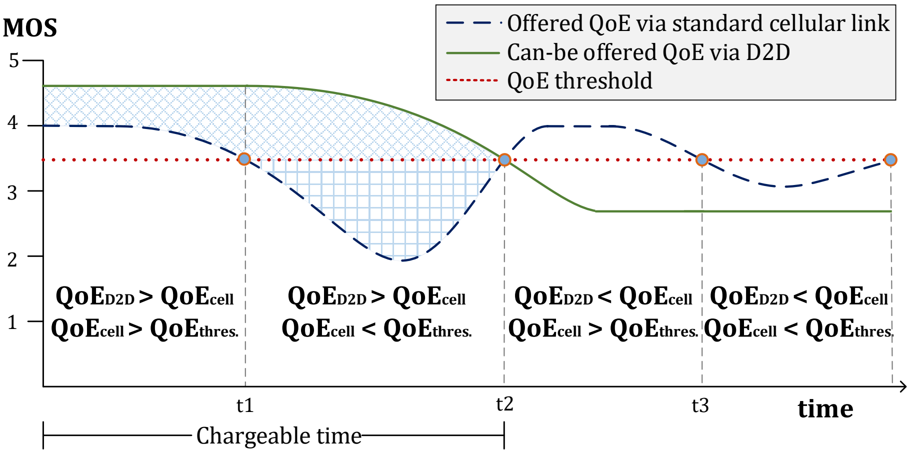


Figure 34: Charging model for QoE-driven D2D management.

Nevertheless, some aspects need further consideration. First of all, we face the fact that the customer is not beforehand aware of the total incurred cost, since QoE depends on a multitude of factors changing in real-time. Actually, the user can only be aware of the maximum possible charging unit per second. Therefore, this model assumes that the customer is willing to pay. Furthermore, we assume that the users' willingness-to-pay is not negatively affected by the fact that the operators also gain from successful D2D links, e.g., in terms of offloading.

Another issue to be considered is QoE's idiosyncrasy to serve both as input and output of the charging process, as discussed in Section 2.5.2. As input, QoE affects the overall charge, whereas as output, it is influenced by the final price. For instance, a hypothetical user, who is paying for a service, is more sensitive to quality disturbances, while a user receiving it for free tends to be more accepting. The E-model, if required, compensates for this case by adding up an "advantage factor" to the *R* factor.

6.5 Simulation

For the purposes of simulation, the "LTE-Sim" framework [36] has been used and significantly extended so as a) to support D2D connections, and b) to realize the proposed QoE-driven D2D management framework. The basic input parameters of the simulations are shown in Table 14.

Table 14: Basic simulation parameters for QoE-driven D2D scenario.

Parameter	Value
Topology	1 macro-cell of 500 m (EPC ignored)
eNB's TX power	43 dBm
Cellular UEs' TX power	eNB-regulated (23 dBm max)
D2D UEs' TX power	-19 dBm fixed (≈50m distance)
UE mobility pattern	Random speed and direction
Traffic load per UE	1 VoIP call
Packet size	20 bytes
Source data rate	8 kbps
VoIP codec	G.729
Channel bandwidth	5 MHz Uplink / Downlink
Duplex mode	FDD
Scheduling algorithm	DL: Proportional fair, UL: Round Robin
QoE assessment model	G.107 E-model, simplified
QoE reporting interval	2 sec (on D2D mode), 10 sec (on cellular)
Pathloss model	$L = 128.1 + 37.6\, log_{10}\, d$

To analyze the user and operator gains, we consider the case of 40 UE pairs, with a configurable percentage of them being within D2D range of their peers. We measure the overall achieved QoE and throughput in the cell and quantify their improvements when applying the proposed scheme, compared to a reference scenario, i.e., where D2D mode is not available. Regarding QoE for instance:

$$MOS\ gain = 100 * (MOS_{D2D} - MOS_{cellular})/MOS_{cellular} \tag{6-3}$$

We plot these improvements against the D2D users' percent in the cell for both near-eNB and near-edge scenarios. The obtained simulation results demonstrate that QoE increases linearly with the percentage of D2D links, as depicted in Figure 35. The average user QoE gain (measured in MOS) goes up to 35% for the extreme case of a cell with only D2D users. This increase is a direct result of the short distance between the communicating entities, providing a better channel with less delays and lower packet loss ratios. Similarly, the better signal propagation conditions between D2D UEs allow for the selection of a higher MCS for the transmissions, resulting in higher throughput. For either metric, the increase is higher for UEs near the cell's boundaries, since users at those locations tend to experience worse channel conditions while on cellular mode, and can therefore benefit more from D2D.

In fact, these results are slightly underestimated, because a) any delays incurred due to the cellular data flowing in the EPC are neglected, b) devices carry out their transmissions over dedicated resources and therefore a higher scheduling delay is imposed compared to a case that resources were shared, and c) VoIP traffic, due to its light-weight nature, does not significantly overload the network. Consequently, even higher QoE values might be expected if this scheme was implemented in a real system.

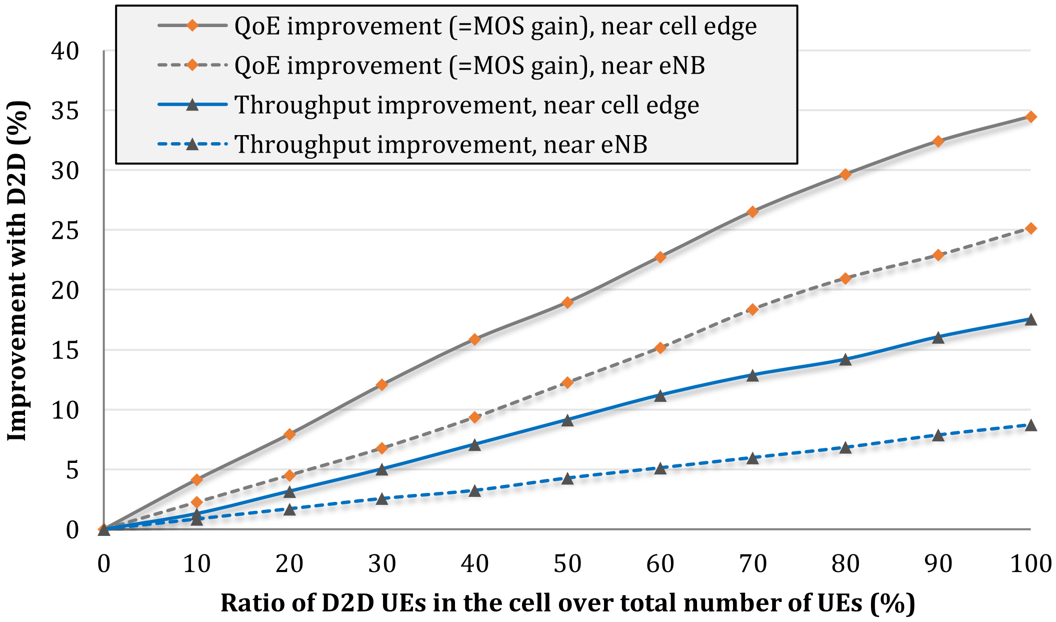


Figure 35: Network improvements when using D2D.

Next, in order to study the coverage area that D2D transmissions may span, Figure 36 presents measurements for different D2D transmission powers derived from simulations where D2D transmitters slowly move out of D2D range. In all five scenarios, we observe that there is a point up to which the receivers steadily measure high MOS values, despite the fact that the distance from the transmitters increases. Beyond that point, however, there is a rapid decrease of MOS values, indicating that devices at some point fall back to cellular mode.

Note in Figure 36, that by changing the D2D transmission power from -19 to -15dBm (i.e., an increase of 153%), the senders can gain about 20m more coverage, (i.e., an increase of roughly 44%). Moreover, a 900% power increase (from -25 to -15dBm) offers an increased range by 132%. Thus, we observe an important trade-off when selecting the D2D design parameters. On the one hand, smaller D2D ranges guarantee large power savings, adding up to the total energy savings due to disengaging from the eNB transmissions and EPC nodes' involvement. On the other hand, the lower the D2D range, the lower the probability for UEs to be found in proximity of each other so as to exploit a direct D2D connection.

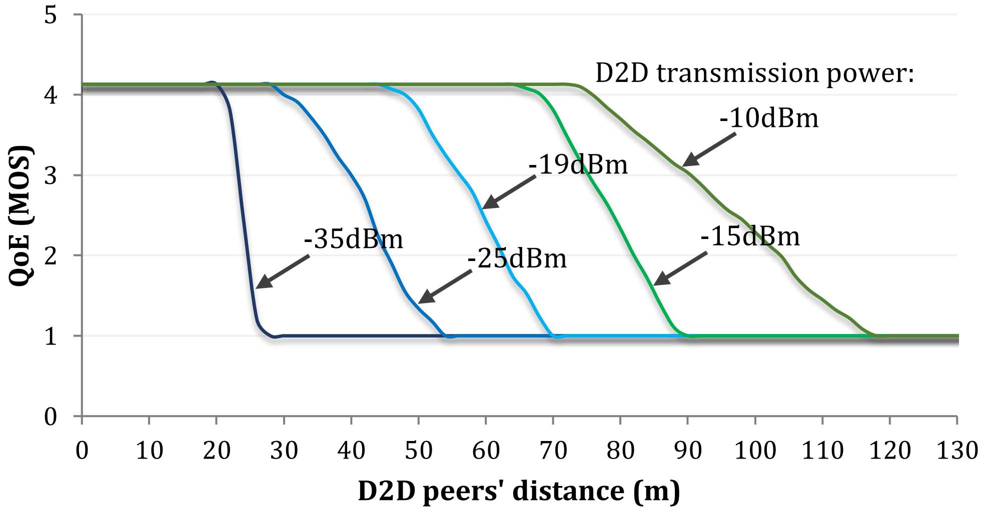


Figure 36: QoE for various D2D ranges and transmission powers.

6.6 Conclusions

In this chapter, we have presented a network management framework, which exploits QoE awareness for controlling the operational mode of mobile users in LTE-A networks with D2D support. Simulations have shown the expected benefits of this mechanism, both for the users (increase in MOS) and the operators (increase in offered throughput). Hence, we envisage that such a QoE-driven scheme may become the enabler for introducing D2D into the market, by allowing operators to qualify for justified and acceptable user charges, when provisioning this new technology.

Future work will include the adaptation of the described model to a multi-cell architecture, allowing also for D2D spatial spectrum reuse and power control. Such an approach would provide an even more efficient system, while it would have to deal with the challenges of new intra- and inter-cell interference situations, as well as the possibilities of establishing D2D links between UEs in neighboring cells.

7. QoE-INSPIRED CONSISTENCY IN RADIO-SCHEDULING

Radio scheduling is a well-studied problem that has challenged researchers throughout the last decades. However, recent findings that stem from the QoE domain come to give a new perspective to traditional radio scheduling approaches. In this study, we take advantage of recent subjective results regarding the impact of throughput fluctuations on the QoE of interactive applications and revisit well-known scheduling algorithms. By quantifying the impact of traditional radio schedulers on user-perceived QoE, we manage to draw new conclusions regarding the radio scheduling problem, such as the importance and impact of consistency of the resource allocation decisions on the users' QoE. As main result, fair algorithms inherently seem to be more consistent than greedy ones, providing less throughput fluctuations and, thus, better QoE. Based on this outcome, we propose a new scheduling approach, which further improves users' QoE by moderating throughput fluctuations.

7.1 Introduction

Nowadays, the QoS and the resulting QoE of mobile users keeps improving thanks to the development and roll-out of new network technologies and standards. In this context we witness a trend towards rising importance of a new quality criterion: network stability in terms of *consistent* performance experienced by the user. This prioritization is visible in the Next Generation Mobile Networks (NGMN) 5G White Paper [84] in which "consistent user experience" forms an integral part of the 5G vision. This is not surprising, since the trend towards ever rising peak rates (as enabled by new Radio Access Network (RAN) technologies) also increases the probability of wireless users experiencing larger network performance fluctuations. Moreover, due to the diversity in the RAN technologies and the heterogeneity within the cellular infrastructure (e.g., overlaying femtos, etc.), the phenomenon of throughput fluctuations becomes even more intense.

Furthermore, fluctuations have a noticeable impact on subscribers' QoE. In [85], it is demonstrated on behalf of subjective user testing results that throughput fluctuations have a significant negative impact on the user experience. Focusing on interactive networked applications, it is suggested that novel downlink-throughput related KPIs have to be developed for proper QoE-based traffic analysis in mobile networks. For the domain of QoE-based network management these results imply that avoiding throughput-related quality fluctuations leads to significant QoE gains.

In general, consistency (in terms of fluctuation avoidance) can be achieved using two different strategies: a) by mitigating the application-level impact of throughput fluctuations, or b) by smoothing throughput on the network-level itself. Each one of these strategies corresponds to a different business case.

On the one hand, application-level strategies are driven by Over-The-Top (OTT) players, who have the means and the interest to control their customers' QoE by handling application level parameters that they can control. A prominent example is HAS, which in essence dynamically changes the media quality (or bitrate) of video segments requested in order to avoid playout buffer starvation. In the HAS context, switches among different layers (i.e., "fluctuations" of media quality) have been identified as an important QoE influence factor (e.g., [86]), and thus video adaptation algorithms with a smoothness logic have been proposed (e.g., [87]).

On the other hand, network-level strategies are driven by Mobile Network Operators (MNOs), who only have the means to control the QoE of their customers through lower-layer parameters. For instance, network-aware bit-rate adaptation schemes have been proposed for UDP/Real-time Transport Protocol (RTP)-based streaming (e.g., [88]), as well as joint rate adaptation and admission control schemes that control how long and by how much the predicted video quality fluctuates/falls below a certain threshold [89].

Future business models even describe the collaboration between OTTs and MNOs, where cross-layer approaches can be envisioned. For instance, [90] proposes a way to mitigate temporal quality fluctuations using lower layer information (e.g., channel quality) and application layer information (e.g., application utility in term of MOS).

In this chapter, motivated by the MNOs' need to provide good QoE to their customers, without relying on OTT players to achieve that, we give our focus on the second strategy, namely on network-level fluctuation mitigation. By giving a solution at network level, we do not depend on different application implementations to solve the same problem, but rather provide a catholic and centralized solution to the MNO's interest (i.e., an application- and device-independent solution).

One promising network-level QoE management approach that can help increase QoE is to ensure stable amounts of bandwidth available to each user. In this context, schedulers play a vital role as they directly influence the radio resource allocation per user. This study adopts this technique, and quantifies the impact of radio scheduler behavior on QoE. More specifically, the study's contribution lies in:

a) evaluating current state of the art schedulers regarding the throughput fluctuations they cause and the respective QoE performance based on fluctuations-aware KPIs, and

b) proposing an *inherently* fluctuations-avoiding scheduler, that further improves the QoE of users.

Our proposal mainly concerns real-time interactive applications, namely web browsing, google maps, IPTV, video-conferencing, etc., where fluctuations are mostly observable; but the implementation itself is application-unaware (so OTT-cooperation is not required). This way we help MNOs better understand and improve the experience of their customers without relying on other parties.

The remainder of this chapter is organized as follows: In Section 7.2, QoE models and metrics are provided that can be useful to evaluate QoE in throughput fluctuation situations. In Section 7.3, the background on traditional radio schedulers is given, explaining their design objectives. Using the said QoE models, scheduling algorithms are compared in terms of QoE and fairness in Section 7.4, while the importance of accounting for fluctuations when designing new radio schedulers is revealed. Section 7.5 describes a novel radio scheduler that inherently accounts for throughput fluctuations, while Section 7.6 evaluates it. Our conclusions are presented in Section 7.7.

7.2 Models and metrics to evaluate QoE

When discussing throughput fluctuations we have to distinguish between two different cases: a) firstly, throughput may fluctuate as a consequence of the normal behavior of an application and/or the natural usage pattern of the user; for example, the YouTube downlink throughput presents a very clear on/off fluctuation pattern as a consequence of the chunk-based flow control of the application, and the downlink throughput pattern of a web browsing session is highly dependent on how fast a user browses a site and goes to the next one. In the second case, b) throughput fluctuates as a consequence of variations in the bandwidth of the corresponding network connection. In mobile networks, the bandwidth of a connection can vary for multiple and very different reasons, such as fast and slow fading, interference, changes in coding and modulation scheme, scheduler algorithm, resource constraints, contention with other users, handovers, etc. As also implied in the introduction, here we focus on this second case, which represents the undesirable and uncontrollable fluctuations stemming from the network.

This study has been triggered by the research outcomes of [85]. The authors of that paper present a complete study of the QoE undergone by 52 mobile users in controlled subjective lab tests, using different mobile applications such as YouTube, web browsing and Google Maps. Their results suggest that novel downlink throughput related KPIs must be defined for QoE-based traffic analysis in mobile networks. The common approach to consider only average throughput values has been found to be insufficient to describe subjectively perceived network quality in the case of news site browsing and browsing Google Maps. Whereas a constant bandwidth of 2 Mbit/s for browsing a Google Map led to a MOS of ≈ 4, an alternating bandwidth of 0 and 4 Mbit/s (average throughput is also 2 Mbit/s) led to a MOS of ≈ 2.6 (see Figure 37). For the case of YouTube, the difference between the two MOS values (constant bandwidth vs. fluctuating bandwidth but identic average throughput) is even bigger: MOS of ≈ 4.5 vs. MOS of ≈ 2.5 [85].

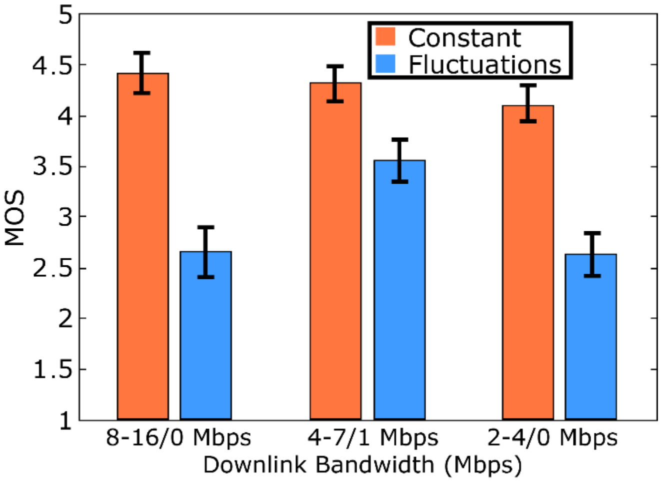


Figure 37: Subjective QoE results for Google Maps browsing, with constant and fluctuating bandwidth of the same mean value [85].

Hence, in [85] a first approach is presented regarding how to define fluctuation-specific KPIs by considering the amount of time in which the throughput is below a certain threshold. With this approach, the so called Effective Average Download Throughput (EADT) can be determined and utilized to calculate the realistic MOS value by multiplying the plain Average Download Throughput (ADT) by a model-dependent Correcting Factor (CF), namely:

$$QoE = f(EADT), \qquad where\ EADT = CF * ADT \tag{7-1}$$

Continuing the previous work of [85], [91] discusses and evaluates five models to derive the EADT. In the first one (LTD, Low-Throughput Duration), the CF is determined by the fraction of time that the throughput is below a certain downlink bandwidth threshold. The second model (SLTD, Selective Low-Throughput Duration) is similar to the first one, but it assumes that short time bandwidth drops are not perceived by the users. Instead of using a fixed download bandwidth throughput, the third model (TJ, Throughput Jitter) uses a moving average-based threshold, e.g., a sliding window length of 5 seconds. The fourth model (AREA, Area-based model) does not only consider the time below a threshold, but also accounts for how deep the corresponding throughput gap is. The fifth model (DOUBLE) is similar to LTD but considers two different bandwidth thresholds.

In the same work, these models are evaluated via empirical user studies. The optimal model selection depends on the scenario (browsing Google Maps, News Site, etc.) and the specific fluctuation pattern (progressive outages with disconnections and subsequent recoveries vs. fast bandwidth changing environments vs. high/low bandwidth profile with fast short-scale variations). Overall, a set of first generic throughput fluctuation models is proposed that allows for quantifying the impact of throughput fluctuations on QoE.

7.3 Background on radio scheduling

7.3.1 Traditional objectives of scheduling algorithms

Radio scheduling is the problem of allocating spectrum resources to competing user requests. Since, commonly, these requests exceed the number of available resources, intelligent radio schedulers need to be designed. Radio schedulers, as of today, are designed to meet four objectives [92]:

Increase spectral efficiency: This objective guarantees the efficient utilization of the radio spectrum, commonly expressed in bit/s/Hz. This can be achieved by accounting for the channel conditions between the base station and the various users in a cell, while taking scheduling decisions. As a consequence, users with better channel conditions get more spectrum resources and hence, achieve higher data rates. In this way, the sum cell throughput is also increased.

Increase fairness: If spectral efficiency was the only criterion for radio scheduling, users with bad channel conditions (e.g., at cell edge) would starve. Therefore, fairness guarantees that even those users receive a decent service in the long run.

Satisfy QoS guarantees: Different flows may have different QoS requirements and constraints, such as a minimum Guaranteed Bit Rates (GBR), maximum acceptable packet delays, etc. QoS-specific schedulers have been designed in order to respect such special requirements.

Achieve low complexity and good scalability: This requirement guarantees that scheduling decisions can be actually taken in real-time, so that they can be implemented into a real base station.

In the literature, a plethora of proposed schedulers can be found that take into account the previous factors. Since, however, these four objectives actually compete with each other (e.g., spectral efficiency vs. fairness, QoS guarantees vs. complexity, etc.), trade-offs need to be made in their design.

7.3.2 State of the art scheduling algorithms

Radio scheduling is the problem of allocating K resources to N users. Its solution is based on estimating a “priority weight” or “metric” in favor of allocating resource k to user j. A comparison of these weights leads to the decision about which resource will be allocated to which user. The rule is that resource k is allocated to user j among all users i, if the following metric is the highest one, namely:

$$m_{j,k} = max_i\{m_{i,k}\} \quad \forall\, i = 1..N, k = 1..K \tag{7-2}$$

One scheduling decision is taken per available resource (i.e., per spectrum unit) per Transmission Time Interval (TTI) and per base station. Some of the most popular scheduling algorithms are presented below:

***Resource Fair (RF) or Blind Equal Throughput (BET)*:** The metric that is estimated by this scheduler is the following:

$$m_{i,k} = \frac{1}{r^i(t)} \tag{7-3}$$

where $r^i(t)$ is the achieved data rate at current time t due to the resources already allocated to user i during the same TTI. Thus, the only objective of this scheduler is to achieve fairness in the resource distribution. It is worth noting that the current decision of this scheduler depends on its previous decisions.

***Maximum Throughput (MT)*:** The metric used in this case is the potentially achieved data rate by each user, if this user is indeed scheduled with the examined resource, namely:

$$m_{i,k} = d_k^i(t) \tag{7-4}$$

where $d_k^i(t)$ is the expected data rate when assigning resource k to user i. This expectation relies on feedback from the users to the base station about the experienced channel conditions. Therefore, according to this metric, users with better channel conditions will get more resources, since they will be able to take better advantage of the channel and support the reception of more bits per second (Downlink). It is interesting, that this scheduler depends on current channel estimations only.

***Proportional Fair (PF)*:** This scheduler is a compromise between the previous two, and is widely used today:

$$m_{i,k} = \frac{d_k^i(t)}{r^i(t)} \tag{7-5}$$

The Proportional Fair scheduler tries to find a balance between spectral efficiency (i.e., maximum throughput in the system) and fairness among the users.

As regards the Resource Fair and Proportional Fair schedulers, a “fairness window“ in the past can be also applied, in which case fairness is targeted over a longer timeframe. In this case, predefined weights are given to the past window and current timeframe.

Towards a new “consistency” objective: It becomes evident that, throughput fluctuations are not considered by these state of the art schedulers (or their variations, thereof). Taking, however, into account that throughput fluctuations directly affect QoE, we here introduce a fifth, new objective for the radio schedulers’ design i.e., a “consistency factor”. This new objective may be added to the list of the four objectives presented above. In this context, we provide the following definition:

Definition: A scheduler is characterized as “consistent” if it minimizes the occurrence or the amplitude of throughput fluctuations. This may be possible for instance by providing highly constant available bandwidth levels to each user.

Next, we are going to investigate how traditional state of the art schedulers perform from a QoE-perspective, using a selection of the models introduced in Section 7.2.

7.4 Comparison of traditional schedulers

7.4.1 Fluctuations-specific comparison

As elaborated before, fluctuations play a crucial role in the perceived QoE. However, existing scheduling algorithms have not been designed with this in mind, and thus, their impact on QoE is unknown. The purpose of this section is therefore to compare current schedulers based on the QoE models described in Section 7.2 and to draw conclusions regarding their efficiency into mitigating throughput fluctuations.

The evaluation of these algorithms has been performed using the LTE-A Downlink System Level Simulator (v1.8 r1375) [93], using the input parameters of Table 15. In LTE-A the scheduling interval (TTI) corresponds to 1 msec.

Figure 38 below shows the instantaneous experienced throughput at a random LTE-A user during 10 sec. Throughput values are smoothed over a 50 msec window instead of being presented for each scheduling interval of 1 msec, for higher readability. The observed fluctuations are a result of the uncontrollable, instantaneous channel conditions, but also of the scheduling algorithm decisions. Since, however, we have used the same channel conditions across all schedulers in this experiment, the resulting differences in the fluctuations' magnitude are caused solely by the scheduling algorithms themselves. This shows that the selection of the scheduler has a strong influence on the resulting fluctuations.

Table 15: Basic simulation parameters for schedulers' comparison.

Parameter	Value
Macro-cell radius	0.5 km
eNB	1 eNB, omnidirectional
eNB TX power	43 dBm
Number of users	Configurable
Distribution of users	Uniform
Traffic load per user	Full buffer
Duplex mode	FDD (focus on downlink)
Channel bandwidth	5 MHz
Number of resource blocks	25
Flow duration	30 sec
Scheduler implementations	[94]
QoE estimation models	LTD, AREA, constant
QoE formula	$0.45 * ln(ADT) + 2.48$

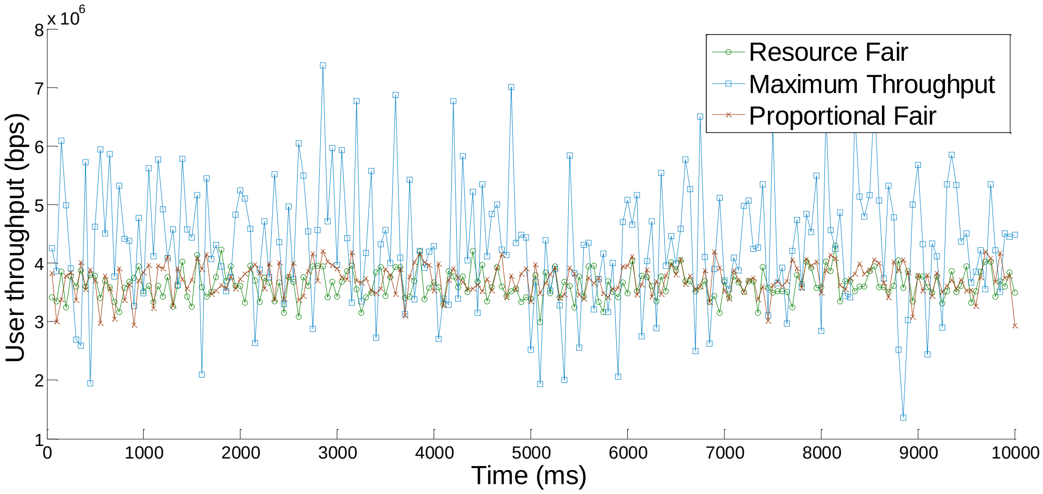


Figure 38: Fluctuations experienced by a random LTE-A user by the three state of the art schedulers.

Regarding the fluctuations' impact of each scheduler, it is shown in Figure 38 that the Proportional Fair and Resource Fair schedulers lead to lower throughput fluctuations. The Maximum Throughput scheduler, on the contrary, leads to significant fluctuations. Therefore, there seems to exist some correlation between the number/magnitude of fluctuations and the *fairness* of the scheduler. That is, the fairer the scheduler, the less the fluctuations. We, therefore, would expect to measure higher QoE values for fairer schedulers, something that we are going to investigate and quantify next.

7.4.2 QoE-specific analysis

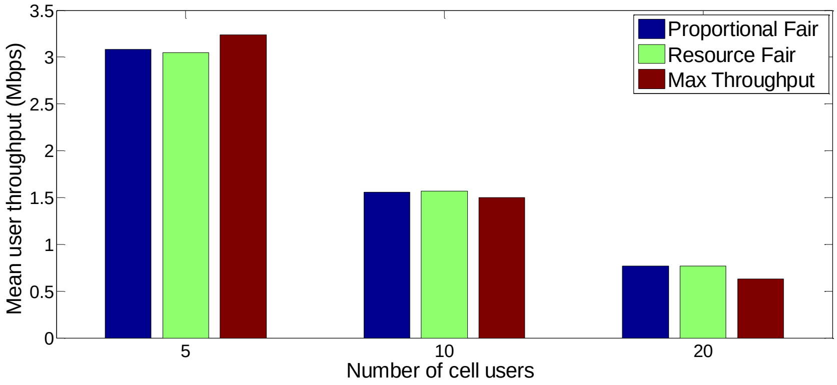
In this section, we compare the aforementioned schedulers in terms of a) the average and b) effective average download throughput that they achieve (i.e., ADT and EADT, respectively), c) the average QoE that they offer, d) the distribution of MOS scores for all users in the cell, e) their fairness, and f) the QoE model used. For the QoE estimations, we have used the LTD and AREA models, as well as a model that ignores fluctuations, namely assumes constant throughput (which is the currently standard approach). The collected results are presented in Figure 39.

First of all, comparing Figure 39a and Figure 39b we observe that the average throughput is much higher than the EADT, while the latter better correlates to the real user QoE, i.e., to the MOS values in Figure 39c. Moreover, the Proportional Fair and Resource Fair schedulers provide better QoE than the Maximum Throughput scheduler. This observation actually reveals the significance of designing and evaluating a network on a QoE- rather than a QoS-basis, and, in the context of this study, it emphasizes the need for scheduling on a QoE-basis.

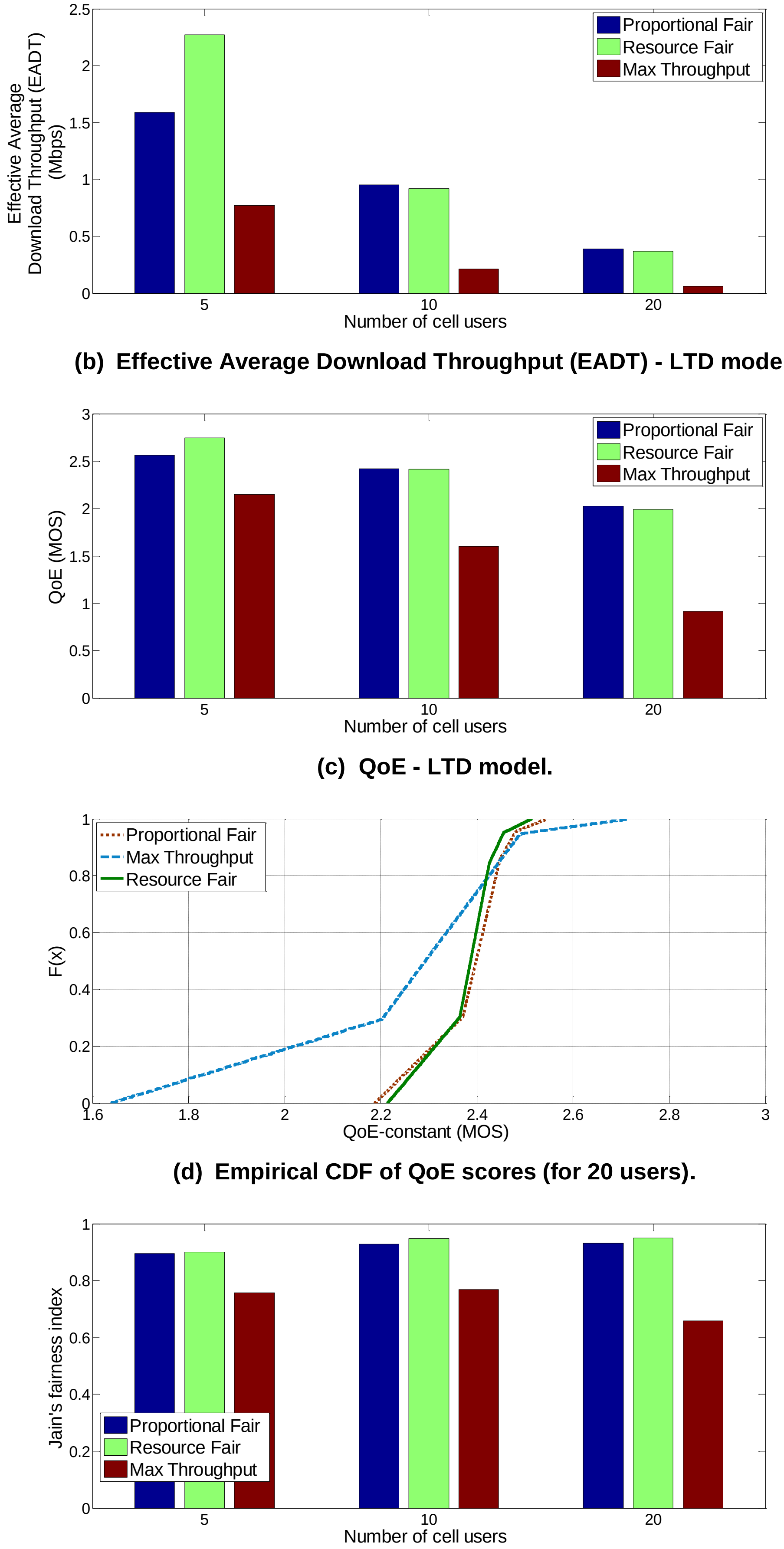
Similarly, what is validated from Figure 39c-Figure 39e is that those schedulers that perform better in terms of QoE are also the fairest ones. This is revealed by the Jain's fairness index presented in Figure 39e (this index takes values 0..1, where 1 represents the fairest), but also by the empirical Cumulative Distribution Function (CDF) plots in Figure 39d. Through the CDF plots it is depicted that fairer schedulers do not cause a high deviation among the MOS scores of different users in the cell (so CDFs are steeper). This is an indication of network stability and consistency in the radio scheduling decisions, which is only achieved by fairer schedulers (i.e., Proportional Fair and Resource Fair).

Finally, examining Figure 39f in terms of the different QoE models implemented, we observe that those models that consider a mean constant throughput actually overestimate the experience of the users. LTD or AREA models give QoE estimations closer to reality (see Section 7.2), since they account for the impact of throughput fluctuations on QoE.

In the next section, we are going to take advantage of the previous conclusions and propose a more consistent, fluctuations-avoiding scheduler.



(a) Average Download Throughput (ADT).

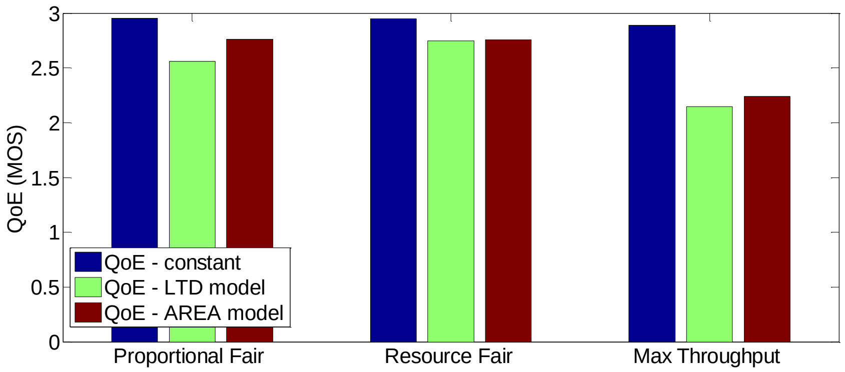


(b) Effective Average Download Throughput (EADT) - LTD model.

(c) QoE - LTD model.

(d) Empirical CDF of QoE scores (for 20 users).

(e) Fairness.



(f) QoE models' comparison (for 5 users).

Figure 39: Comparison of state of the art schedulers.

7.5 Designing a consistent scheduler

In the previous sections, the impact of throughput fluctuations on QoE has been revealed and it has been shown that the fluctuations' effect can be *indirectly* moderated at some extent by using fairer schedulers. However, a more efficient way to achieve that is to design new schedulers that explicitly mitigate these fluctuations.

We therefore propose a fluctuations-aware, consistent scheduler. This scheduler takes into account the evolution of the achieved throughput over time (per user) and the impact of this evolution on the user QoE, an aspect not currently addressed by any state of the art schedulers.

The fluctuations' effect may be moderated, i.e., smoothed out, by introducing a new metric that tries to capture and mitigate the magnitude and occurrence of fluctuations. The purpose of this metric is to quantify the gap between the average throughput value over a time window in the past (say $\overline{R^{\iota}}(t-1)$) and the expected data rate for the current time interval for each user (the sum of all $d_k^i(t)$). The goal is to minimize this gap, namely to minimize the amplitude of the resulting fluctuations. The larger this amplitude, the less the favoring of giving resource k to user i. Since a decision needs to be taken jointly for all users and for all the available resources, user j will be allocated with k only if:

$$m_{j,k} = max_i\{m_{i,k-fluct}\} \quad \forall\, i = 1..N, k = 1..K \tag{7-6}$$

Overall a complex optimization problem needs to be solved, with the objective to find the minimum number of resources per user that minimize this user's deviation from his past average throughput. The optimal solution will provide the best possible combinations of resources that minimize the fluctuations for all users at the same TTI. However, in order to find a solution that works in real-time (sub-optimal though), we introduce the following metric:

$$m_{i,k-fluct} = \frac{1}{\overline{R^{\iota}}(t-1) - r^i(t) - d_k^i(t)} \tag{7-7}$$

where:

$$\overline{R^{\iota}}(t-1) = \frac{\sum_{\tau=t-1-W}^{\tau=t-1} r^i(\tau)}{W} \tag{7-8}$$

and W is the window length over which the average throughput is estimated, while $r^i(t)$ and $d_k^i(t)$ have the same meaning as for the state of the art schedulers.

The way this scheduler works is graphically depicted in Figure 40 (abstract example). Say there are three users in a cell competing for a total of six available resources during one TTI (also known as Resource Blocks - RBs). Each RB will result in a different data rate when allocated to a different user, subject to the user's channel quality. $\overline{R^1}$ to $\overline{R^3}$are the past average throughput values per user, which the scheduler tries to maintain in order to avoid fluctuations. Therefore, the decisions will be as shown in Figure 40. Each decision is taken per RB, and it is based on minimizing the gap between $\overline{R^\iota}$ and the data rate progressively achieved per user in the current scheduling interval. (The achieved data rate, r, is progressively increased every time a user gets another RB in the current TTI).

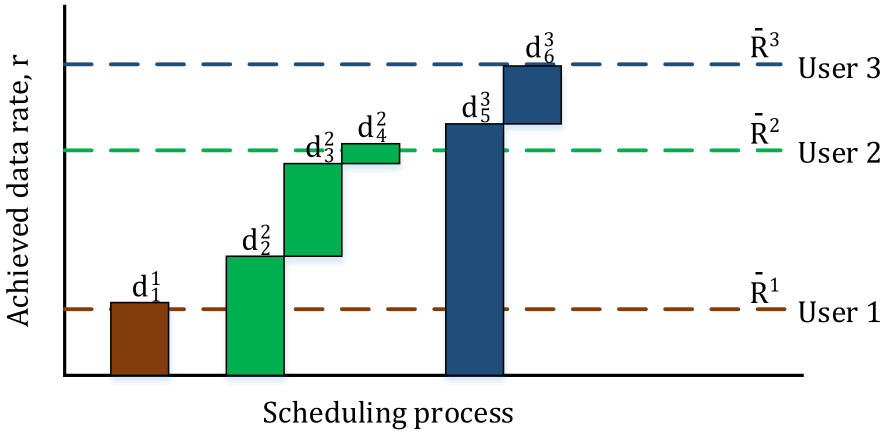
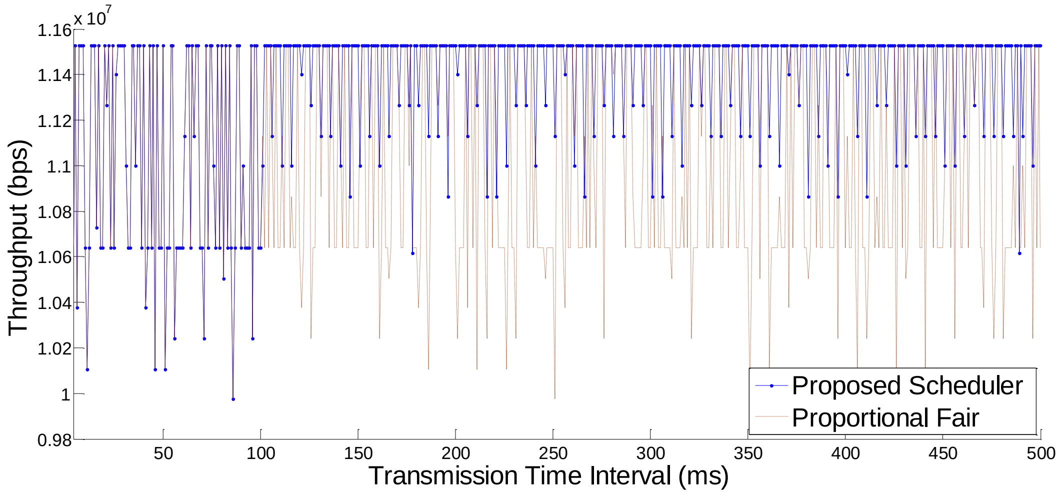


Figure 40: Scheduling logic of the proposed consistent scheduler.

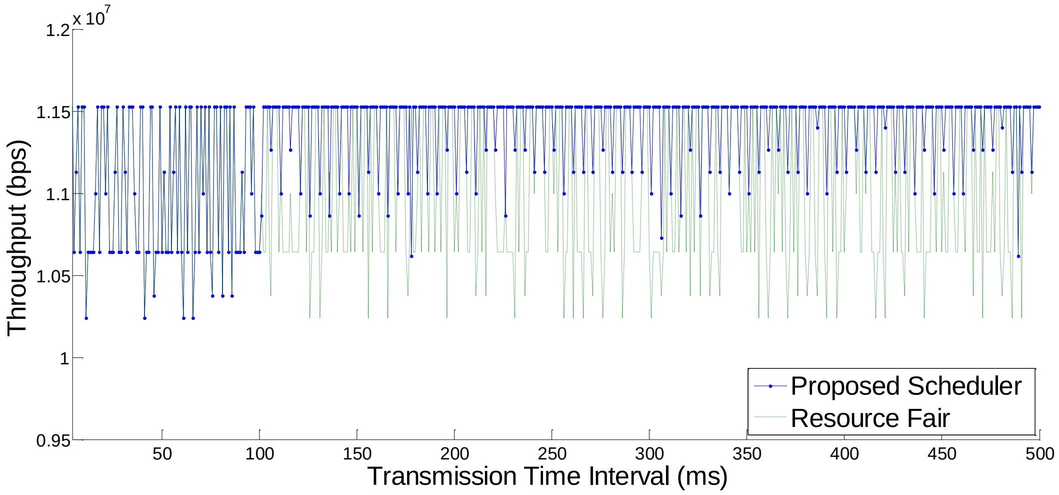
7.6 Evaluation study

For the purposes of evaluation, we implement the proposed scheduling algorithm into the LTE-A simulator of [93]. In the first evaluation study, we aim to prove the concept of the proposed metric for a specific user in the cell.

The results are shown in Figure 41, where we can visualize the successful fluctuations' mitigation. A comparison is done with the Proportional Fair and Resource Fair schedulers, while for the Maximum Throughput scheduler the differences are much higher. Note that we have used the Proportional Fair and Resource Fair schedulers for the 100 first TTIs (Figure 41a and Figure 41b respectively), after which the proposed fluctuations-avoiding scheduler is activated ($W = 100$ TTI).



(a) Comparison with the Proportional Fair scheduler.



(b) Comparison with the Resource Fair scheduler.

Figure 41: Proof of concept of the proposed scheduler.

Next, we compare the CDF of the proposed scheduler with the state of the art schedulers, for the case of 20 users uniformly distributed in the cell. The results are presented in Figure 42. We can observe that the proposed scheduler (blue line): a) is very fair, as shown by the steepness of the CDF, b) that the achieved *minimum* MOS values are higher than for the other schedulers (CDF shifted to the right), while c) the larger MOS values are comparable to the other schedulers. This behavior is explained by the fact that the resource allocation procedure of the proposed scheduler is greedy in some sense. By trying to minimize the gap between the average throughput values and the potentially achieved data rates jointly for all the users, eventually this scheduler manages to first satisfy the low-throughput users. This happens, because the lower the average $\overline{R^{\iota}}(t-1)$, the lower the difference to the achieved data rate $d_k^i(t)$ and thus the higher the scheduling priority. However, the low-throughput users do not necessarily take the "best" RBs, and therefore higher-throughput users are also served well.

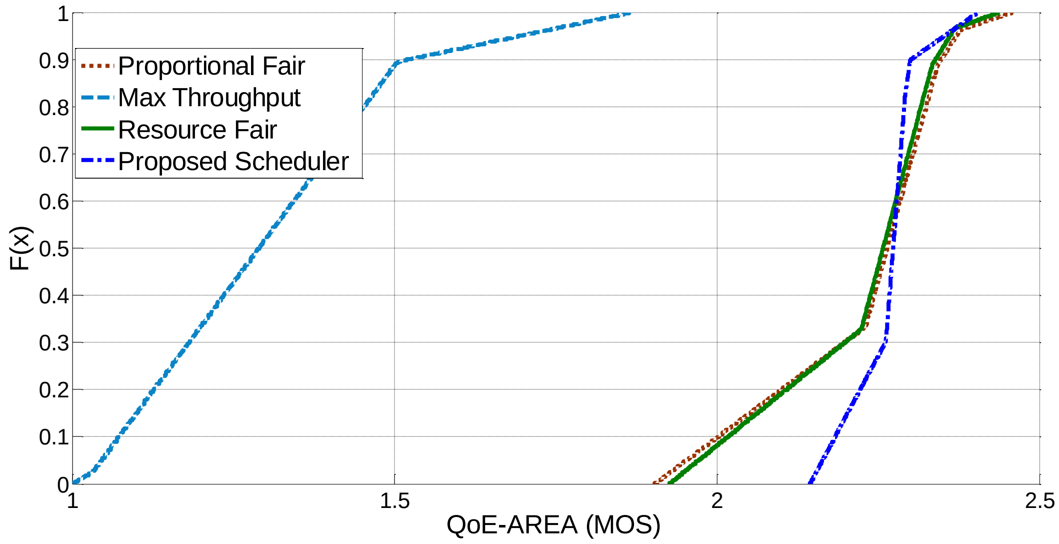


Figure 42: AREA-MOS CDF for standard schedulers and the $m_{i,k-fluct}$ metric.

7.7 Conclusions

One aspect that has only recently been acknowledged regards the impact of throughput fluctuations on the perceived user QoE. This is the reason why "consistency" is an aspect lacking appropriate attention in current state of the art radio schedulers. The study described in this chapter tries to cover this gap by explaining the meaning and significance of taking consistent radio scheduling decisions, proposing in parallel this novel research direction for future works.

With this in mind, we have evaluated exemplary scheduling algorithms in a realistic LTE-A network simulator. We have reached the conclusion that fairness inherently favors consistency, which is a valuable attribute among different users, but also regarding a single user. On the one hand, consistency among different users is desirable so that the expectations of users co-located in the same cell are similar. On the other hand, consistency over time for a single user is also essential, as it has been revealed by the discussed studies that map per-user throughput fluctuations to QoE. In this chapter, we have validated this conclusion by demonstrating that fairer schedulers outperform maximum throughput ones in terms of QoE, as can be measured by proper KPIs.

Nevertheless, these fair exemplary schedulers only indirectly account for the per-user fluctuations. Having identified this deficiency, we have proposed a novel fluctuations-avoiding scheduler that explicitly smooths throughput fluctuations. The measured achieved QoE improvements demonstrate the potential of this scheduler as well as the significance of research towards that direction. Therefore, future work is required in order to design more sophisticated fluctuations-aware schedulers that optimize the decision-making process, considering in parallel real-time constraints.

8. ENRICHING HTTP ADAPTIVE STREAMING WITH CONTEXT AWARENESS

Video streaming has become an indispensable technology in people's lives, while its usage keeps constantly increasing. The variability, instability and unpredictability of network conditions poses one of the biggest challenges to video streaming. In this chapter, we analyze HTTP Adaptive Streaming (HAS), a technology that relieves these issues by adapting the video reproduction to the current network conditions. Particularly, we study how context awareness can be combined with the adaptive streaming logic to design a proactive client-based video streaming strategy. Our results show that such a context-aware strategy manages to successfully mitigate stallings in light of network connectivity problems, such as an outage. Moreover, we analyze the performance of this strategy by comparing it to the optimal case in terms of QoE-related KPIs for video streaming, as well as by considering situations where the awareness of the context lacks reliability. The collected evaluation results encourage further research on how context-awareness can be exploited to further enhance video service provisioning by OTT service providers.

8.1 Introduction

8.1.1 Motivation

The rising number of smart phone subscriptions, which are expected to reach 9.2 billion by 2020, combined with the explosive demand for mobile video, which is expected to grow around 13 times by 2019, accounting for 50% of all global mobile data traffic, will result in a ten-fold increase of mobile data traffic by 2020 [95]. This explosive demand for mobile video is fueled by the ever-increasing number of video-capable devices and the integration of multimedia content in popular mobile applications, e.g., Facebook and Instagram. Furthermore, the use of video-capable devices, which range from devices with high resolution screens to interactive head mounted displays, requires a further increase of the bandwidth, so that on-demand video playback can be supported, and differentiated expectations raised by the end video consumers can be satisfied.

In parallel, since most of the consumed video of a mobile data network is delivered through server-controlled streaming, the ability of traditional HTTP video streaming to support a fully personalized video playback experience at the user is questioned. To this end, this technique is gradually being replaced by client-controlled video streaming exploiting HAS. HAS splits a video file into short segments of a few seconds each, with different quality levels and multiple encoding rates, allowing a better handling of the video streaming process, e.g., by adapting the quality level of future video segments. HAS is a key enabler towards a fully personalized video playback experience to the user, as it enables the terminal to adapt the video quality based on the end device capabilities, the expected video quality level, the current network status, the content server load, and the device remaining battery, among others.

Following this immense interest for video streaming, mobile operators, ISPs and OTT players are very interested in understanding and, thereafter, improving the QoE of their customers. Conventionally, each one of these stakeholders makes use of their own available data and possible means of controlling the users' experience, intervening in parameters that reside in different OSI layers. For instance, networks providers can influence their customers' QoE by controlling QoS network parameters (e.g., implement packet prioritization, traffic shaping, etc.), while OTT providers can control higher-layer parameters (e.g., adapt the video resolution, encoding rate, etc.). In parallel, users have mechanisms to control their streaming experience, for instance using application layer techniques, such as HAS, as mentioned above.

Beyond these interventions of different stakeholders to isolated OSI layer parameters, the idea of designing cross-party and cross-layer mechanisms has also emerged [96]. The main challenge is to exploit the "context", referring to any type of information that raises the aforementioned isolation. More specifically, context-awareness may be based on information that a) is globally available or well-known (e.g., a map), b) can realistically be passed on from one interested party to another (e.g., information about network traffic or social context information), or c) can be acquired from different OSI layers or by other means (e.g., awareness of the signal strength or the user's speed at the application layer).

In this chapter, our objective is to investigate how context awareness in mobile networks can help not only understand but also enhance the user experienced quality during HAS sessions. We study a scenario where users travelling within a vehicle experience bad or no service at all (i.e., a service outage). In this or similar type of scenarios, the opportunity emerges to propose novel, preemptive strategies to overcome such imminent problems, for instance by proposing proactive adaptive streaming or buffering techniques for video streaming services. This scenario has been modelled, optimized and investigated by means of simulation.

Before presenting the problem under study, we first identify the need and the changes needed to move from a QoE-oriented to a context-aware network/application management.

8.1.2 From QoE-awareness to context-awareness

As discussed in the previous chapters, QoE is an inherently subjective indication of quality. Consequently, a significant amount of research efforts has been devoted to the measurement of this subjective QoE. The awareness of QoE in a network is valuable knowledge not only per se (namely for network monitoring and benchmarking purposes) but also as useful input for managing a network in an effective and efficient way. The "QoE-centric management" of a network can be performed as a closed loop procedure, which consists of three distinguishable steps, as it has been discussed in more detail in Chapter 3.

"Context" may refer to "*any information that can be used to characterize the situation of an entity*" [97]. In this way, context awareness can facilitate a transition from packet-level decisions to "scenario-level" decisions: Indeed, deciding on a per-scenario rather than on a per-packet level may ensure not only a higher user QoE but also the avoidance of over-provisioning in the network. This immense potential has been recently identified in academia and as a result, research works on context awareness and context-aware network control mechanisms are constantly emerging in the literature. For instance, in [98], a context aware handover management scheme for proper load distribution in an IEEE 802.11 network is proposed. In [99], the impact of social context on compressed video QoE is investigated, while in [100] a novel decision-theoretic approach for QoE modeling, measurement, and prediction is presented, to name a few characteristic examples.

If we now revisit the three-step QoE management loop described in Chapter 3 by also considering context awareness, then this is enriched as follows:

- **Context modeling:** Based on the discussion of Chapter 4 regarding the QoE modeling procedure, we may observe that the "System" as well as the "Human" influence factors are directly or indirectly taken into account in the subjective experiments' methodologies, e.g., [37]. Consequently, the impact of technical- and human-level characteristics is tightly integrated into the derived QoE models. Nevertheless, the "Context" influence factors are mostly missing in these methodologies or are not clearly captured. This happens because QoE evaluations are usually performed in controlled environments, not allowing for diversity in the context of use. Besides, context factors are challenging to control, especially in a lab setting, and new subjective experiment types would have to be designed. As a consequence, the mapping of context influence factors to QoE is absent from most QoE models that appear both in the literature and in standardization bodies. Therefore, novel context-aware QoE models need to be devised that are able to accurately measure and predict QoE under a specific context of use, as these context factors are (often) neglected. These context factors could either be integrated inside a QoE model directly, or, be used as a tuning factor of an otherwise stand-alone QoE model.

- **Context monitoring:** On top of QoE monitoring, context monitoring procedures could be implemented in the network. These procedures will require different input information from the ones used by traditional QoS/QoE monitoring techniques. The acquired context information may be used for enhancing the QoE of the users or for the prediction of imminent problems, such as bottlenecks, and may range from spatio-temporal to social, economic and task-related factors. Some of the possible context information that may be monitored in a network is the following (to give a few examples): the current infrastructure, which is more or less static (access points, base stations, neighboring cells, etc.), the specific user’s surrounding environment (location awareness, outdoors/indoors environment, terrain characteristics, presence of blind spots such as areas of low coverage or limited capacity, proximity to other devices, etc.), the time of day, the current and predicted/expected future network load, the current mobility level or even the predicted mobility pattern of users in a cell (e.g., a repeated pattern), the device capabilities or state (e.g., processing power, battery level, storage level, etc.), the user task (e.g., urgent or leisure activity), as well as application awareness (e.g., foreground or background processes), and social awareness of the users, among others. Moreover, charging and pricing can be included in the general context profile of a communication scenario. It needs to be noted here that context awareness does not necessarily rely on predicting the future (e.g., future traffic demand) but also on solid knowledge that is or can become available (e.g., time of day, outage location, etc.).
- **Context-aware management:** Three management possibilities emerge in a context-aware network. First, the network can take more sophisticated control decisions that are also influenced by context-awareness, such as a decision to relax the handover requirements for a user in a fast-moving vehicle or a decision to connect a device with low battery to a close WiFi access point. Second, the network can actualize control decisions exploiting the current context. For instance, it can exploit information about flash crowd formation to drive an effective Content Distribution Network (CDN) load balancing strategy [101] or, more generally, to take control decisions proactively based on context information about the near future. Finally, context-awareness can contribute towards taking decisions with the objective to increase the network efficiency as measured in spectrum, energy, processing resources or other requirements, and as a consequence to reduce operational expenses. For instance, context information could allow for a more meaningful distribution of the network resources among competing flows that refer to different communication scenarios.

Nevertheless, it needs to be noted that any context information service comes with certain costs in terms of privacy. A careful balance between those two objectives, i.e., preserving privacy and increasing the user’s QoE, would need to be found, but this is not currently under study in this work.

This chapter handles a characteristic use case of context-aware management and showcases its potential. More specifically, it studies a scenario where “context awareness” refers to awareness of the location and duration of a forthcoming outage, namely of a restricted area of very low or zero bandwidth (e.g., limited coverage due to physical obstacles or limited capacity due to high network congestion). Based on this knowledge, a proactive HAS strategy is devised that will enhance the viewing experience of a user travelling inside a vehicle towards this area.

8.1.3 Related work and contribution

Enhanced HAS strategies that account for future network conditions have lately emerged. A characteristic example is HAS strategies that use geo-location information (e.g., [102] and [103]), which evoke users to send measurements regarding their achieved data rates. These strategies rely on the collection of these device measurements in order to create a bandwidth lookup-service, which is then used to improve the prediction of future bandwidth availability. Our main differentiation with this approach is the exploitation of context-awareness in order to avoid the constant signaling to a bandwidth database, thus, we propose a context-aware rather than a predictive strategy. Moreover, [104] proposes a technique that identifies zero-bandwidth spatiotemporal events and triggers the HAS client to react accordingly. It demonstrates that by proposing a reactive “replace-request” method that substitutes higher quality segment requests with lower quality ones, stallings can be successfully prevented. However, in more bandwidth-challenging cases, *proactive* rather that *reactive* HAS strategies are required in order to sufficiently prepare for longer limited signal conditions.

Other HAS techniques rely on prediction as well, rather than context-awareness. For instance, [105] proposes an anticipatory HAS strategy, which requires prediction of the channel state in terms of Received Signal Strength (RSS) and proactively adjusts the user’s buffer. An optimization problem is formulated that minimizes the required number of spectrum resources, while it ensures the user buffer is better prepared for an imminent coverage loss. The authors even conducted a demo of this approach in [106] that serves as a proof of concept. Our difference with this approach, is that we rely on longer-term context-awareness rather than imminent channel prediction, and that instead of manipulating the user buffer size, we proactively adapt the video quality selection. Finally, [107] combines RSS information with localization sensors from the smart phones that reveal the user’s coverage state and help achieve a smoother and more stable HAS policy, called Indoors-Outdoors aware Buffer Based Adaptation (IOBBA).

In parallel, our proposed strategy is complementary to any other HAS strategy, since it can be activated at a specific instant of time, when the need arises.

This study’s contribution is summarized in the following:

- A proactive HAS strategy based on context-awareness is proposed, capable of avoiding stallings usually experienced by video streaming users under limited bandwidth conditions.
- Under a realistic scenario, the problem of preventing stalling events is formulated as a non-linear programming problem. To solve this, a close to optimal strategy in terms of QoE is proposed.
- The minimum advance time, when the enhanced HAS strategy should start running to guarantee a seamless video streaming experience, is estimated both analytically and via simulation. Constraints and dependency factors of this time parameter are investigated.

- A comprehensive discussion on the feasibility of the proposed approach into a real network is provided.
- An extensive evaluation process is followed, including users' QoE assessment through subjectively-validated HAS-compliant QoE models.

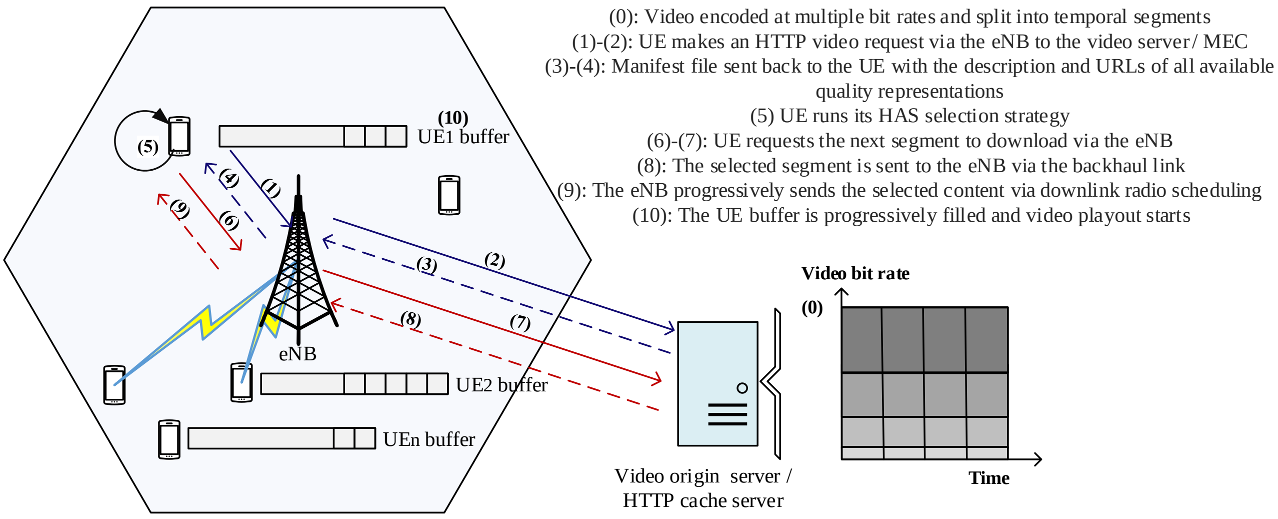
The remainder of this chapter is organized as follows. In Section 8.2 the system model is described and the problem under study is formulated. Also, the HAS logic in a mobile cellular network is briefly presented. The proposed context-aware HAS strategy and the optimal solution are described in Section 8.3. Evaluation results are presented in Section 8.4, while Section 8.5 concludes this chapter.

8.2 System model

8.2.1 HAS in a mobile cellular network

In this section, we briefly present the HAS logic within the context of a mobile cellular network.

In HAS, each video is encoded at the server side in multiple representations with a different quality level per representation (otherwise called "layer"). Different quality layers have differences in the video bit rate (bps) or in the video resolution, etc. Each representation is divided into "segments" of a few seconds each (around 2-10 sec each). The availability of these different layers becomes available to each user through a manifest file, before streaming starts. Then, each user requests the next segment that he wants to download, with the objective to eliminate any stalling and maximize the video bit rate. This decision is taken by each user independently based on information available at his side, namely: a) the manifest file, b) the user's current buffer level, and c) a "short-sighted", i.e., subjective perception of the network congestion, as this is independently and individually perceived by the throughput of the last downloaded segment(s). Namely, standard HAS relies on taking decisions in isolation from the rest of the network and unaware of the future network state.



(0): Video encoded at multiple bit rates and split into temporal segments
(1)-(2): UE makes an HTTP video request via the eNB to the video server/ MEC
(3)-(4): Manifest file sent back to the UE with the description and URLs of all available quality representations
(5) UE runs its HAS selection strategy
(6)-(7): UE requests the next segment to download via the eNB
(8): The selected segment is sent to the eNB via the backhaul link
(9): The eNB progressively sends the selected content via downlink radio scheduling
(10): The UE buffer is progressively filled and video playout starts

Figure 43: The HAS paradigm in LTE/LTE-A.

The HAS strategy followed by typical users is based on weighted perceived downlink data rate of previously downloaded segments. In a dynamic mobile environment, the achieved data rate is a result of: a) the scheduling algorithm combined with the Modulation and Coding Scheme (MCS), b) the user location in the cell, and c) the momentary load in each cell sector, as a result of competing flows' requests for bandwidth. The HAS operations in a cellular network environment are illustrated in

Figure 43, describing step by step the end-to-end logic of video streaming, starting from the user request for watching a video, up to the point that video playout starts at the user side. The notation used in this figure is that of an LTE/LTE-A network.

8.2.2 Problem description: The tunnel scenario

Consider a mobile user streaming video content over TCP (e.g., YouTube). Due to the unstable nature of the wireless medium, mobility, and physical obstacles, the channel quality may fluctuate significantly and, thus, the user may experience "coverage holes". The existence of a tunnel is a common example of a coverage hole in a cellular environment, meaning that users travelling through it will experience limited or no connectivity. This event is described as an "outage". For video streaming users, such an outage will potentially lead to a stalling event due to buffer depletion, i.e., to video freezing.

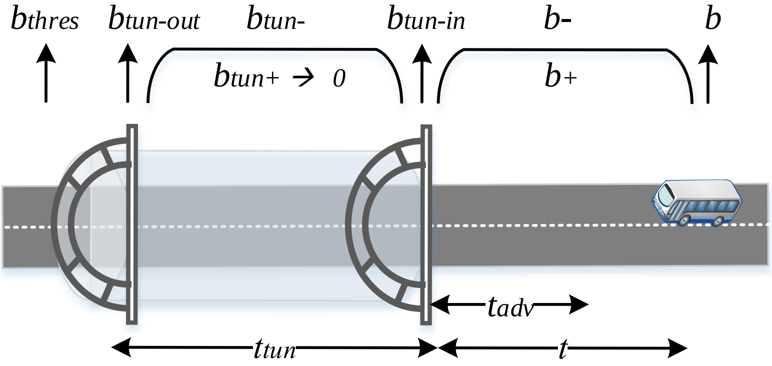


Figure 44: Problem description using buffer status information.

Assume a single streaming user inside a vehicle (e.g., a bus or train) travelling in a particular direction and with a specific speed (Figure 44). We assume, that the positioning and the length of an upcoming tunnel are known in advance (due to context awareness). Therefore, the remaining distance between the vehicle and the tunnel's entrance is also available at the client side. This distance corresponds to a travelling time of t, namely the time required until the user enters the outage region. Let b be the current buffer status of this user's HAS application. Then, during t, this buffer level will be boosted by b_+ but also reduced by b_-. Throughout the tunnel, the buffer will be reduced by b_{tun-}. Note that inside the tunnel there is negligible or no connection, so there is no buffer boost, i.e., $b_{tun+} = 0$. When the user enters (exits) the tunnel, the application's buffer level will be b_{tun-in} ($b_{tun-out}$), respectively, and it will hold that:

$$b_{tun-in} = b + b_+ - b_- \tag{8-1}$$

$$b_{tun-out} = b_{tun-in} - b_{tun-} \tag{8-2}$$

Then, we can express the objective of the proposed HAS strategy as the following:

$$b_{tun-out} \geq b_{thres} \tag{8-3}$$

which ensures that when the vehicle is exiting the tunnel, the buffer status of the HAS application will be at least equal to the minimum buffer threshold, b_{thres}, and so the video playout continues uninterrupted. Note that, a stalling always occurs when $b < b_{thres}$. Using equations (8-1) and (8-2) inside (8-3):

$$b_+ \geq b_{thres} + b_- + b_{tun-} - b \tag{8-4}$$

This condition answers the question about how much the buffer of the HAS application needs to be pro-actively filled during t, so that no stalling will occur. This should be achieved despite the imminent connection disruption. Note that all the parameters on the right-hand side of (8-4) are known to the client or can be easily estimated (b_{thres} is fixed, b is directly known to the client application, while b_{-}, b_{tun-} can be estimated).

In the next section, we estimate the minimum required time $t_{adv} \leq t$ to ensure a stalling-free video streaming.

8.2.3 Approaching the minimum required "advance time"

Based on the previous system model, at any point t, we can estimate the b_{+}, namely the required buffer boost (in bytes or in seconds) to avoid any stalling inside the tunnel. This measurement can be then further translated to a *minimum* required "advance time", t_{adv}, when the travelling user needs to start running the proposed proactive HAS strategy, *at the latest*, in order to avoid stalling. The t_{adv} should be sufficiently large so that the user has enough time to react; otherwise, a stalling will be inevitable, in which case the user can potentially be warned and be given the option to watch the video later. We assume that the users switch from any standard HAS strategy to the enhanced one exactly at t_{adv}.

We can express b_{+} as a function of t_{adv} as follows:

$$b_{+} = r * t_{adv} \qquad (8\text{-}5)$$

where r (bytes per sec) is the estimated experienced data rate by the client's application. Namely, r is the user's perception of the average available network bandwidth, as estimated by the HAS strategy. Therefore, the minimum required advance time in order to avoid any stalling would be:

$$t_{adv} \geq \frac{b_{thres} + \ b_{-} + b_{tun-} - b}{r} \qquad (8\text{-}6)$$

The t_{adv} will ensure that b_{+} will last for the whole zero-bandwidth tunnel duration. Since users, however, may travel at different speeds (u), it would make more sense to further translate this "advance time" to "advance distance". Then, the minimum distance (in meters) before which the user needs to be notified about the tunnel, x_{adv}, would simply be: $x_{adv} \geq u * t_{adv}$.

The crucial question left to answer is: What is the quality representation per video segment that has to be downloaded, namely what is the synthesis of b_{+} (which layers to be downloaded and in what order).

8.3 Context-aware HAS strategies

Standard-HAS approaches will inevitably lead to stallings in challenging network conditions (e.g., inside a tunnel or any other area of limited or zero bandwidth). This leads us to the proposed strategy that attempts to overcome the existence of stalling events even in zero connectivity conditions.

The main idea to achieve this is to pro-actively and deliberately decrease the quality layer of the requested segments for the video streaming application in advance (i.e., before the user enters the tunnel). As a result, the user's buffer when entering the tunnel will be kept at a higher level during video playback than it would have been without such a scheme. This idea is presented in Figure 45. In this figure, the "real time" axis represents either the time spent to download a segment (before tunnel start) or the time it takes to play a segment (after tunnel start). Note that the magnitude of these time values is not to be compared with each other in this illustration.

We now approach the problem of finding the appropriate quality segments that will sequentially fill the b_{+} a) as an optimization problem, and b) using a proactive HAS strategy.

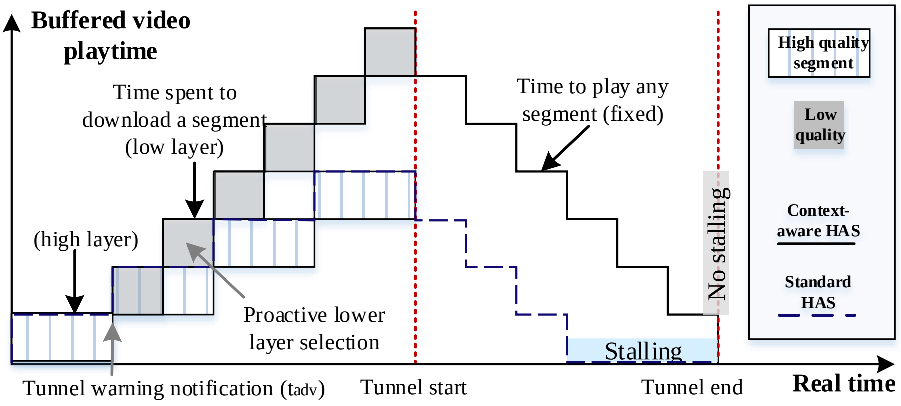


Figure 45: HAS scenario with and without context awareness.

8.3.1 Optimal HAS

The goal of this section is to formulate linear and non-linear programming problems that achieve optimal segment selection with respect to three different optimization objectives, described next. Each optimization problem is formulated using the following notation ([86] is used as a reference):

- τ is the length of each segment in seconds.
- T_0 is the initial delay of the video.
- D_i is the deadline of each segment i, meaning that this segment needs to be completely downloaded up to this point.

Then:

$$D_i = T_0 + i\tau, \qquad \forall i = 1, \dots, n \tag{8-7}$$

Also:

- n is the total number of segments that comprise the video.
- r_{max} is the maximum number of available layers/representations.
- x_{ij} represents segment i of layer j.
- w_{ij} is the weighting factor for the QoE of segment i of layer j (here, we use the quality layer value as weighting factor = {1,2,3}).
- S_{ij} is the size of segment i of layer j (e.g., in bytes).
- $b(t)$ is the total data downloaded until the point in time t. We assume perfect knowledge of $b(t)$.
- α is the weight for the impact of the quality layer and β for the impact of the switches $(\alpha + \beta = 1, \alpha > 0, \beta > 0)$.

QoE studies on HAS (e.g., [68],[70]) have revealed that major quality influence factors are in order of significance: a) the layers selected and especially the time spent on highest layer, and b) the amplitude, i.e., the difference between subsequent quality levels (the smaller the better). Other factors with less significance are: the number of quality switches, the recency time and the last quality level. Taking these findings into account, we focus on three distinct types of optimization objectives, which aim to maximize the positive impact of higher level selection, deducing the negative impact of quality switches and amplitude.

Three different versions of optimization objectives are thus formulated, as follows:

- Optimal strategy “W” accounts only for the weighted impact of the quality layers, trying to maximize their value, so that the highest layer will be favored over an intermediate layer, while an intermediate layer will be preferred over the lowest layer.
- Optimal strategy “W+S” additionally accounts for the number of switches, trying to minimize their occurrence.
- Optimal strategy “W+S+A” additionally accounts for the impact of the amplitude, trying to minimize the “distance” between subsequent layers, thus preferring direct switches e.g., from layer 1 to layer 2 rather than from layer 1 to layer 3 and vice versa.

This leads us to the following three different formulations of the optimization problem:

- **W:** Maximize the quality layer values:

$$maximize \sum_{i=1}^{n} \sum_{j=1}^{r_{max}} \alpha w_{ij} x_{ij} \tag{8-8}$$

- **W+S:** Maximize the quality layer values minus the number of switches (the term ½ is used so as to count each switch exactly once):

$$maximize \sum_{i=1}^{n} \sum_{j=1}^{r_{max}} \alpha w_{ij} x_{ij} - \frac{1}{2} \sum_{i=1}^{n-1} \sum_{j=1}^{r_{max}} \beta (x_{ij} - x_{i+1,j})^2 \tag{8-9}$$

- **W+S+A:** Maximize the quality layer values minus the number of switches and the amplitude difference:

$$maximize \sum_{i=1}^{n} \sum_{j=1}^{r_{max}} \alpha w_{ij} x_{ij} - \frac{1}{2} \sum_{i=1}^{n-1} \sum_{j=1}^{r_{max}} \beta \left[(x_{ij} - x_{i+1,j})^2 + \frac{(x_{ij} - x_{i+1,p})^2}{|p-j|} \right] \tag{8-10}$$

where:

$$p = \{1..r_{max}\} - \{j\} \tag{8-11}$$

Despite its complication, the terms in the last parenthesis of Eq. (8-10) represent the preference over switching between “neighbor” layers (i.e., after a layer 1 selection, the layer $p = 2$ will be preferred / after a layer 2 selection, either the layer $p = 1$ or $p = 3$ will be preferred / while after a layer 3 selection, the layer $p = 2$ will be preferred). A similar behavior will be observed if more than 3 layers are available.

All above optimization objectives are subject to the following constraints:

$$x_{ij} \in \{0,1\} \tag{8-12}$$

$$\sum_{j=1}^{r_{max}} x_{ij} = 1, \qquad \forall i = 1, \dots, n \tag{8-13}$$

$$\sum_{i=1}^{k} \sum_{j=1}^{r_{max}} S_{ij} x_{ij} \leq b(D_k), \qquad \forall k = 1, \dots, n \tag{8-14}$$

The three constraints in this problem are interpreted as follows: x_{ij} is a binary value (Eq. (8-12)) meaning that a segment is either downloaded or not, each segment has to be downloaded in exactly one layer (Eq. (8-13)), and all segments need to have been downloaded before their deadline, so that no stalling occurs (Eq. (8-14)).

Next, we add a set of outages Q_{outage} where the bandwidth is zero. An outage (l, m) starts at segment l and ends at segment m. In order to view the video until segment m throughout the outage duration, it needs to have been downloaded until l. This can be expressed as follows:

$$\sum_{i=1}^{m} \sum_{j=1}^{r_{max}} S_{ij} x_{ij} \leq b(D_l) = b(D_m), \qquad \forall (l, m) \in Q_{outage} \tag{8-15}$$

Compare $b(D_l)$ with b_+ in Eq. (8-4). Also, note that full knowledge of all parameters is necessary to solve this optimization problem. While this can hardly be achieved in a real scenario, partial knowledge may allow for sufficiently good heuristics.

It should be noted that stalling events are not considered in our model. Instead, the model works under the assumption that stalling can always be prevented by switching to a lower layer, otherwise the model is “infeasible”. For the sake of simplification, the initial delay is also ignored in our model.

8.3.2 Proposed HAS strategy

The proposed strategy needs to overcome the existence of stalling events during the outage, something which is extremely high likely to occur due to the very low network coverage. The main idea to achieve that is to pro-actively and deliberately decrease the quality layer of the requested segments for the video streaming application in advance (i.e., before the user enters this region). As a consequence, the buffer at the user side when entering the tunnel/outage region will be fuller than it would have been without such a scheme (see Figure 45).

As a result of this strategy, the user viewing experience will be less affected, not only because the video will continue to play without a stalling for a longer period of time (or hopefully will never stall depending on the outage duration), but also because the quality level will be *gradually* decreased (subject to the HAS strategy implementation) and thus the user will be better acquainted with lower quality levels. Such progressive quality degradations would be preferred in comparison to sudden and unexpected quality degradations, especially if the quality level is already very high (cf. the IQX hypothesis [20]). Overall, the main objective of the proposed strategy is to compute the optimal context-based quality level selection strategy to ensure the best QoE while avoiding any stalling events.

The HAS strategy is based on the estimation of the required buffer boost b_+ as this was described in Section 8.2.2. As for the estimation of the expected downlink rate (network bandwidth prediction), this is assumed equal to the segment rate. The segment rate estimation (in bytes per second) is done over a sliding window of the past k downloaded segments as follows:

$$r = (1 - w) * \frac{Size\ of\ last\ (k-1) segments}{Time\ to\ download\ (k-1) segments} + w * \frac{Size\ of\ segment\ k}{Time\ to\ download\ segment\ k} \tag{8-16}$$

where w is the weight (importance) given to the latest downloaded segment. Based on this rate estimation, the expected bytes that can be downloaded until the user enters the outage region is:

$$b_{+expected} = r * t_{adv}, \qquad (in\ bytes) \tag{8-17}$$

while the minimum required buffer playtime b_+ to exit the outage region and avoid a stalling is as in Eq. (8-4):

$$b_{+} = b_{thres} + b_{-} + b_{outage-} - b, \qquad (in\ seconds) \tag{8-18}$$

Therefore, the required bytes per segment are:

$$required\ video\ rate = \frac{b_{+expected}}{b_{+}}, \qquad (in\ bytes\ per\ second) \tag{8-19}$$

Note that the higher the outage duration, the larger the b_{+} and thus the lower the required video rate (lower layer selection). Based on the required video rate estimation, the HAS strategy will request the highest possible representation j that fulfills this condition:

$$\frac{S_{ij}}{\tau} \leq required\ video\ rate \tag{8-20}$$

Namely, the layer j that will be requested will be the highest one that yields a video bit rate less or equal to this estimation. The “required video rate” estimation may be updated each time in order to account for the most recently achieved data rate r. Alternatively, an average value may be calculated in the beginning (on t_{adv}) and assumed valid until entering the outage region. (Note: We assume that the client requests the lowest layer when initialized). In the case that the actual available data rate for this user is less than his subjective rate estimation, r, there is, however, a risk of stalling. Overall, this algorithm will determine the selection of the next video segment, proactively degrading the quality if required.

8.3.3 QoE models

The QoE models that are used in this work are the following (i.e., parametric models described in Chapter 5):

- A QoE model for HAS, where no stallings are assumed. This model, also discussed in Section 5.2.3.3, can be found in [86] and it can be described by the following formula:

$$QoE\ =\ 0.003 \cdot e^{0.064 \cdot t} + 2.498 \tag{8-21}$$

where t is the percentage of time that the video was being played out at the highest layer (here layer 3).

- A QoE model for video streaming over TCP, for the case that stallings occur. This model, also discussed in Section 5.2.3.2, can be found in [65] and it is described as follows:

$$QoE\ =\ 3.5 \cdot e^{-(0.15 \cdot L + 0.19) \cdot N}\ +\ 1.5 \tag{8-22}$$

where N is number of stalling events and L is the stalling length.

For the purposes of this scenario we combine the two aforementioned models, so that in case that no stalling has occurred, the former QoE model is used, while during and after a stalling event, we use the latter.

8.3.4 Realization in the network

Although we are not going to delve into details regarding the realistic application of the proposed framework into a mobile cellular network, we will give some insights. This discussion concerns the type of cross-layer and cross-party context information that is needed and how it may be acquired. The information, assumed to be known for this approach, is:

- The existence of a tunnel (or any other physical coverage hole), namely the tunnel starting and ending point (or, equivalently, its duration). This information is taken into account by the enhanced HAS strategy to find the appropriate activation time. The acquisition of such information is considered realistically possible, since terrain maps are/can be easily available at mobile phones. Alternatively, “Big Data” can empower the collection of such information.

- Information about the user’s direction and speed is required to predict whether the user will pass through a tunnel. This is also available via Global Positioning System (GPS) information (current location, speed and trajectory combined with a terrain map) and it may be estimated by the device itself by a path prediction algorithm.
- The minimum advance time t_{adv} or distance x_{adv} at which the user needs to activate the proactive HAS strategy. These estimations mainly depend on the tunnel duration, the user speed and the user’s perception of the network data rate. Since the user is capable of knowing about the existence of a tunnel a priori, he can estimate b_+ based on Eq. (8-4) and then t_{adv} based on Eq. (8-6). Therefore, the user is able to activate the enhanced HAS mode on t_{adv} without any network assistance, and hence, avoid/minimize the stalling occurrence.
- The standard information required for HAS is needed as well, namely information about the available video segments (acquired from the server), an estimation of the available network bandwidth for this user (estimated at the client as the size of downloaded segments over the time required to download those), and the current buffer state, which is also known at the client’s application.

As far as the need for “Big Data” mentioned before is concerned, this may take two forms: Either they could be data collected at the device itself, as a user usually has the same travel profile every day and, therefore, learns about any coverage problems on his way, or, the data are collected at a central network point (e.g., at a base station or a server) through measurements collected by any devices passing from there. Actually, in LTE networks, such measurements are already available via CQIs. CQIs report to the eNB the quality of the received signals (SINR) using values between 1 (worst) and 15 (best). Currently, CQIs are used only for real-time decisions such as scheduling; however, we may envision that CQIs may be collected by an eNB on a longer-term time scale (days or weeks) and be used in order to create a “coverage profile” of the cell. Following such past information, proactive measures could be taken at a cell for users travelling towards problematic areas (e.g., a physical tunnel ahead).

8.4 Evaluation

For the purposes of evaluation, we use Matlab simulation. We have implemented the client’s buffer using a queuing model, where the downloaded segments are considered as arrivals, and the played segments as departures. To simulate the network traffic, we rely on real traces recorded from a network, namely on a realistic traffic pattern recorded in a vehicular mobility scenario by [108]. Moreover, to simulate congestion we use the parameter “bandwidth factor” [86], which is a metric of the network congestion/traffic and takes values between 0 and 1 (the higher this factor the lower the congestion).

For the purposes of this simulation, the following parameter values have been considered (Table 16):

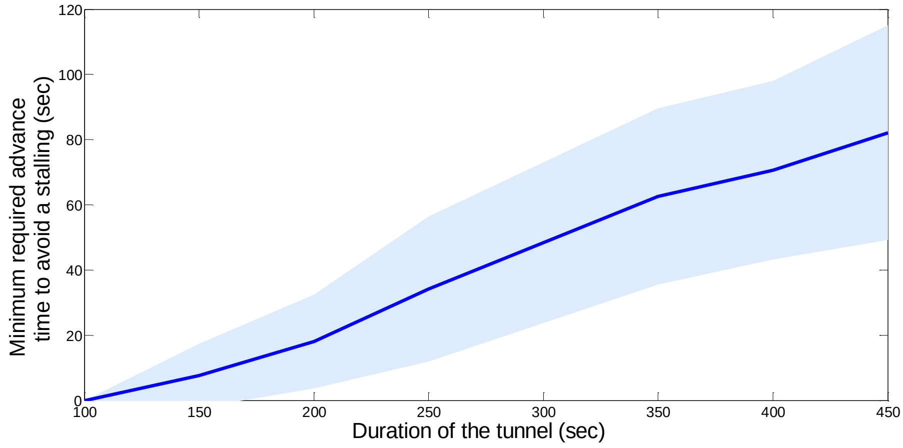
Table 16: Basic simulation parameters for tunnel HAS scenario.

Parameter	Value
Segment duration	2 sec
Number of video segments	350

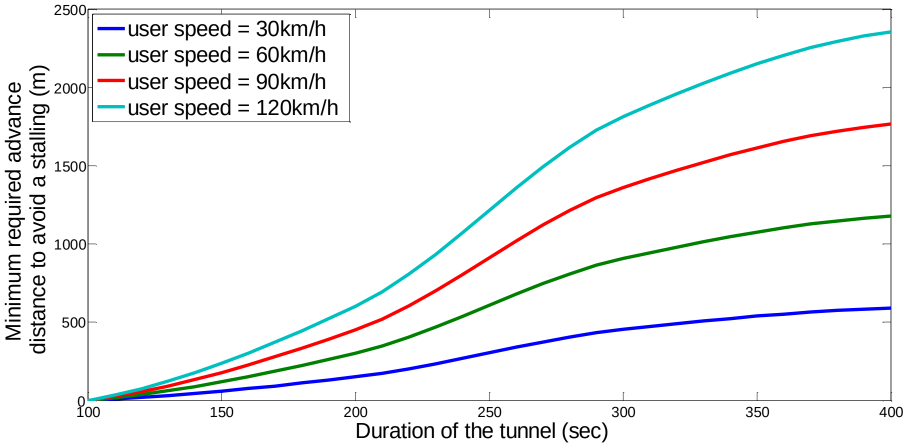
Available representations (layers) per segment	3
Buffer playout threshold (initial delay)	10 segments
Buffer size	Unlimited
Tunnel starting point	200 sec after sim start
Tunnel duration	[0..400] sec
HAS policy sliding window	50 segments
Bandwidth factor	0.8
Network traces used	30 different traces [108]
alpha (beta) coefficient	0.95 (0.05)

8.4.1 Estimation of the minimum required advance time

First of all, we practically estimate the advance time, t_{adv}, when the user needs to switch to the enhanced proactive HAS mode, and we study the impact of the tunnel duration on this metric. Please note that in the proposed system model, the user will execute the standard HAS strategy before t_{adv} and the context-aware HAS right after t_{adv}.



(a) Advance time (t_{adv}).



(b) Advance distance (x_{adv}).

Figure 46: Minimum required advance time and distance to avoid a stalling event inside the tunnel.

The respective simulation results are presented in Figure 46a. We validate the expected trend, namely that an increase in the tunnel duration mandates an earlier reaction by the user (a higher t_{adv}), so that the enhanced HAS strategy has more time to overcome the imminent network outage. It is also interesting to observe that the standard deviation also increases for higher tunnel lengths, which means that the level of uncertainty is higher in these circumstances. The reason is that any predicted estimation of the network data rate for the future is riskier when there is a lot of time ahead before entering the tunnel.

For a better understanding of the previous results, we plot the same scenario assuming different travel speeds of the users and present the required x_{adv}, to avoid a stalling (Figure 46b).

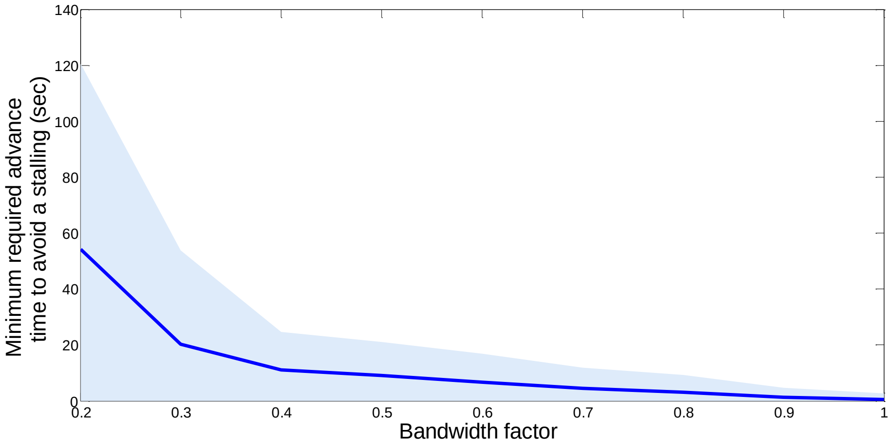


Figure 47: With context awareness: Minimum required advance time to avoid a stalling event in light of an outage event of 150 sec for various bandwidth factors.

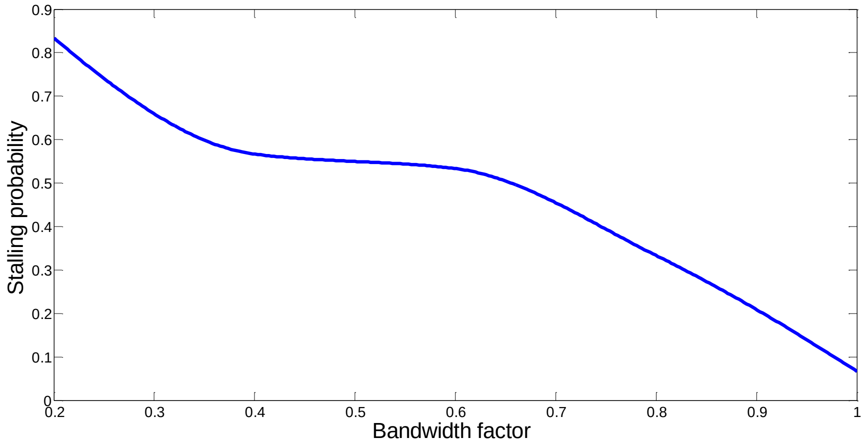


Figure 48: Without context awareness: Stalling probability for various bandwidth factors.

Next, we perform a study with respect to the availability of bandwidth, in order to evaluate how HAS performs in bandwidth-challenging scenarios. Since we use real traces as input information about the data rates in the network, we can indirectly enforce a network congestion by multiplying the measured bandwidth with the aforementioned bandwidth factor. Therefore, a low bandwidth factor emulates high network congestion, whereas higher values indicate low congestion.

The purpose of the first study with regard to the bandwidth factor is to investigate how it influences the minimum advance time t_{adv} in the case of context awareness. The results are presented in Figure 47. As demonstrated in this figure, for very low data rates (e.g., a bandwidth factor of 0.2), the minimum required advance time gets higher, as the user would need a much greater time margin to proactively fill the buffer in light of the outage, because the network is heavily congested. Moreover, the uncertainty in this case is also very high. On the contrary, the more relaxed the network conditions, the higher the margin for an early notification about the outage, while this practically gets zero seconds (i.e., no notification is needed) when the network conditions are very relaxed (bandwidth factor = 1).

Similar conclusions are drawn for the conventional context-unaware case with regard to the stalling probabilities for different bandwidth factors, namely the less this factor, the higher the stalling probability, as expected (Figure 48).

8.4.2 Proof of concept

Having estimated the appropriate advance times, next we conduct simulations that serve as a proof of concept of the effectiveness of the context-aware HAS strategy. The objective is to demonstrate how the proposed policy can indeed overcome an otherwise inevitable buffer depletion in light of a connection outage (here, a tunnel) and thus, prevent any stalling.

To prove that, we plot four different metrics: a) the client buffer size in bytes, b) the client buffer size in seconds (i.e., playtime), c) the HAS layers that the client has selected for each played out segment, and finally d) the QoE evolution in time for the travelling user (in MOS). For the latter, we make the assumption that the QoE models presented in Section 8.3.3 hold also in a real-time scale, and that the QoE model for HAS holds for the tested scenario where three different layers are available per segment. Real-time QoE estimation for a particular user means that QoE is estimated at every time instant t using as input accumulated information about the percentage of time that this user has already spent watching the video at layer 3 up to instant t, as long as no stalling has occurred yet, or information about the number N and duration L of stalling events since $t = 0$ up to instant t, as long as at least one stalling has occurred.

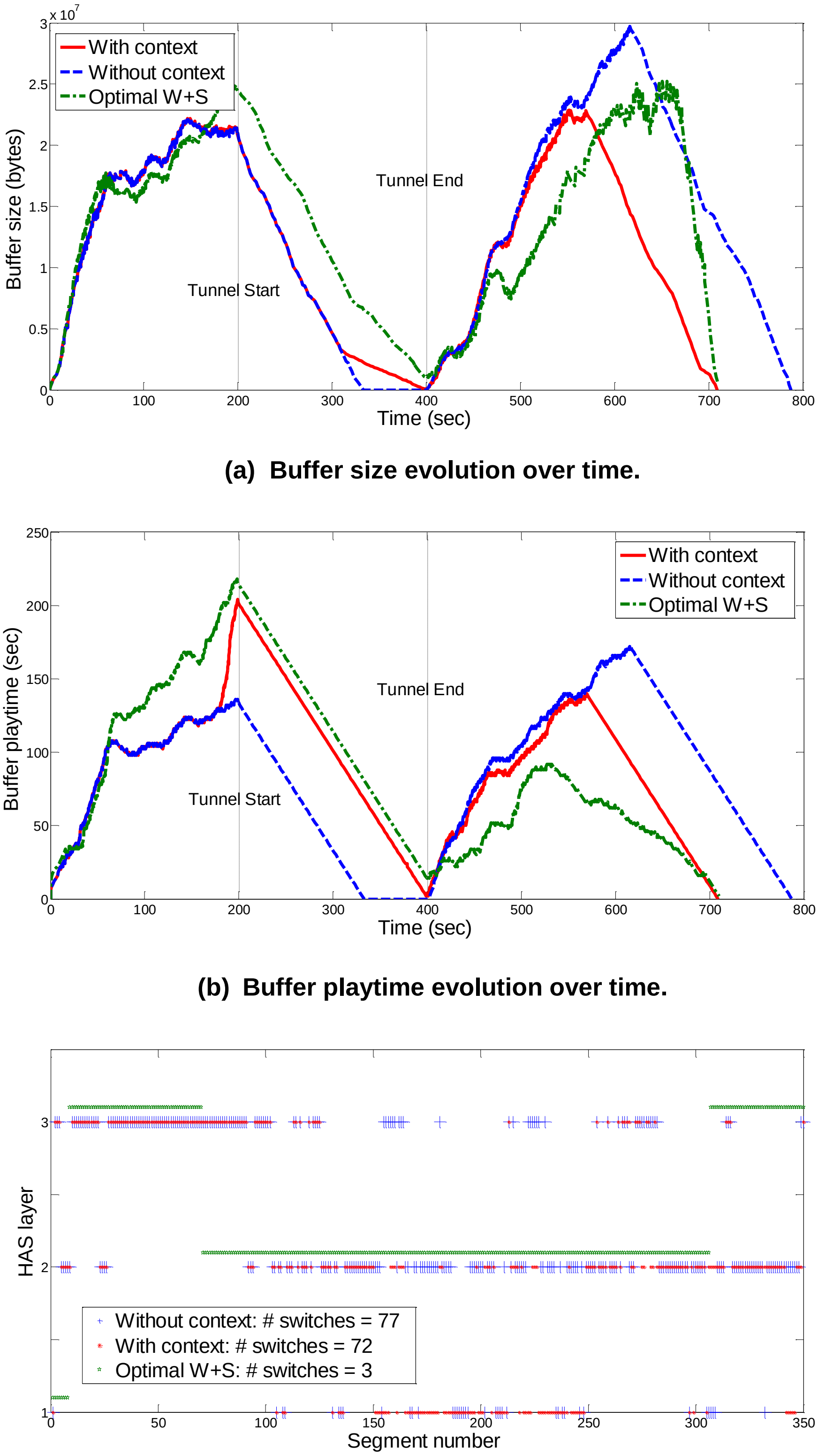
In Figure 49 we present: a) the conventional case, where no context awareness about the outage event is available, and consequently, the standard HAS strategy is continuously executed, b) the case where context awareness about the starting point and duration of the outage event is available, which automatically leads to the selection of the new HAS strategy after t_{adv}, and c) the optimal case “W+S” described in Section 8.3.1.

Looking at Figure 49a we can see that a stalling of around 80 sec is completely avoided when context awareness is deployed, or when optimal knowledge is assumed (buffer is never emptied). A similar conclusion is drawn by Figure 49b. The explanation behind the prevention of the stalling lies in Figure 49c: In the “without context” case higher HAS layers are selected as compared to the “with context” case. Having downloaded lower HAS layers in the “with context” case, the buffer of the client is fuller in terms of playtime than it would have been if higher HAS layers had been downloaded instead.

Compared to the optimal case “W+S”, Figure 49c shows that the number of switches in the proposed strategy are more. The reason is that this number is not a decisive factor in the proposed strategy (Section 8.3.2). However, as mentioned before, the impact of switches on the user experience is much lower than the impact of the “time on the highest layer”, which is the major QoE influence factor.

In fact, in terms of QoE, the proposed strategy performs very well (Figure 49d). However, QoE fluctuates more often, because, as shown in Figure 49c, layer 3 is not selected continuously (as it happens in the optimal case), but frequently switches among all layers.

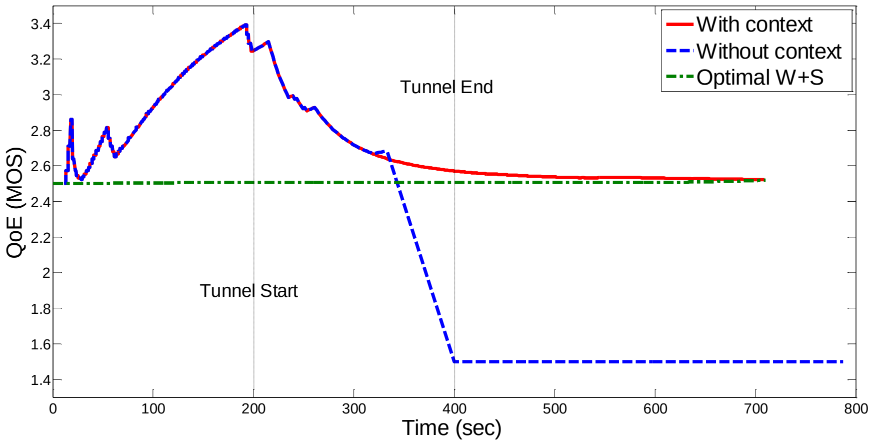
Figure 49d also reveals that even a single stalling event of a few seconds' duration has a significantly deteriorating impact on the perceived QoE, as compared to the selection of lower HAS layers. Another observation worth mentioning, is that QoE values per strategy follow the trend of layer selection: this is why the "context case" at some periods reveals higher QoE than the "optimal" case (the former requests more layer 3 segments before the outage).



(a) Buffer size evolution over time.

(b) Buffer playtime evolution over time.

(c) Selection of HAS layers.



(d) QoE evolution over time.

Figure 49: Comparison of the standard HAS strategy, the context-aware HAS strategy and the optimal solution “W+S”.

Comparing finally the enhanced HAS strategy with the optimal strategy “W+S”, we observe that the latter does a better job in selecting higher quality layers (especially layer 2 segments) up to the point of the outage start. The reason is that the optimal strategy has full awareness of the future network conditions and thus, can take more informed decisions that lead to the highest layer selection with zero stalling risk.

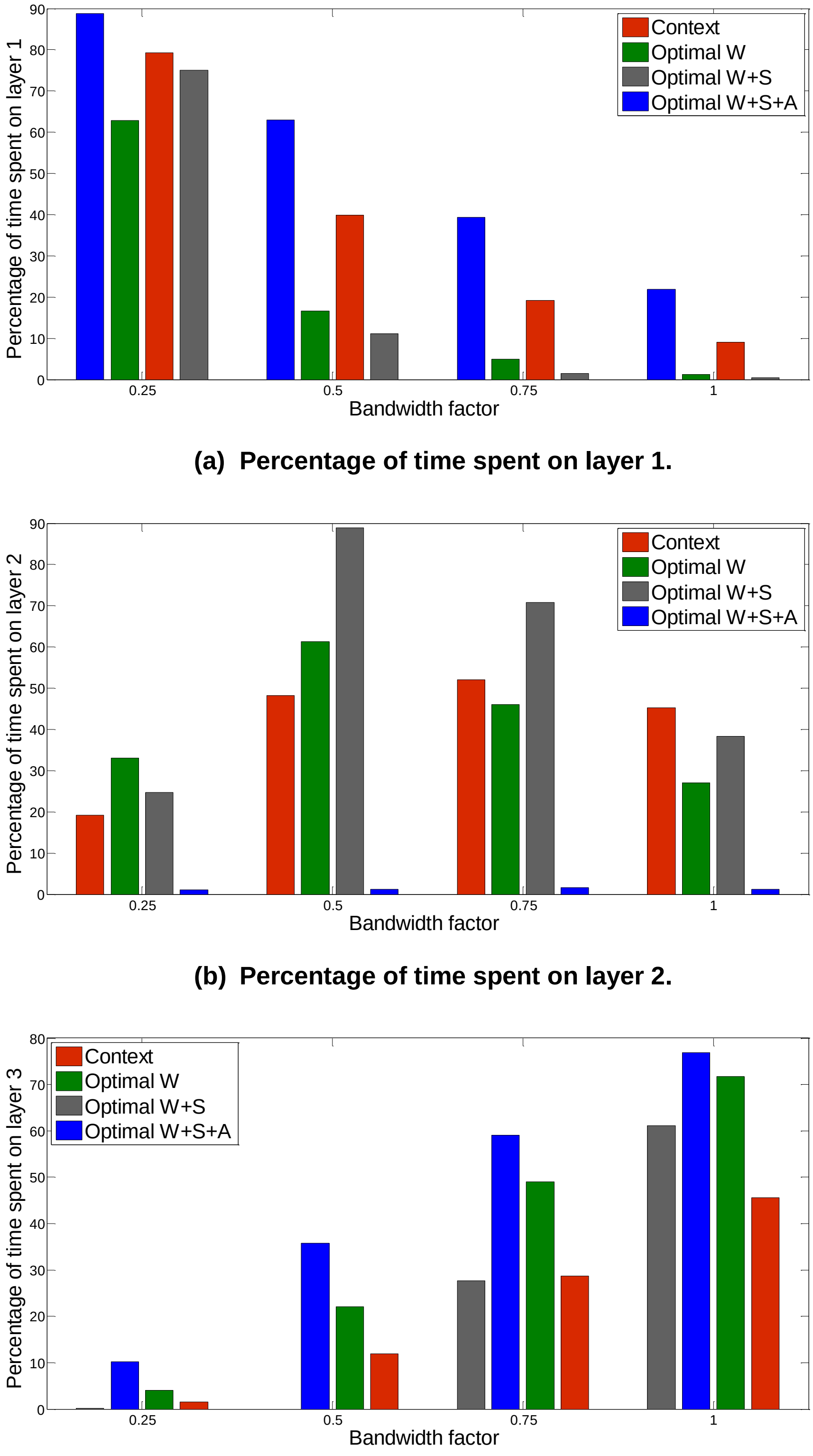
8.4.3 Comparison of different strategies

Next, we compare the behavior of the three different types of the optimal strategy (i.e., cases W / W+S / W+S+A, as described in Section 8.3.1) both among them, but also with the context-aware strategy. In Figure 50a-Figure 50e, the percentage of time spent on each of the three layers as well as the resulting number of switches and QoE are presented per strategy. All four strategies follow a similar trend as bandwidth availability increases, that is higher and higher layer 3 segments are selected, while lower and lower layer 1 segments are selected. With respect to layer 2 segments, the behavior is different when the bandwidth factor changes from 0.25 to 0.5 (increasing layer 2 selection) from when it changes from 0.5 to 1 (decreasing layer 2 selection). In terms of QoE, all strategies operate at very close MOS values, while W+S+A performs slightly better than all, in compliance with the higher layer 3 selection shown in Figure 50c.

Another interesting observation is that strategy W+S+A “avoids” layer 2 segments almost completely. The reason behind that is that layer 2 in W+S+A is mostly used as a “transition step”, used to switch to the lower layer 1 or higher layer 3, respecting the objective to keep the amplitude of two sequential layers as low as possible. Eq. (8-10) gives the same priority to staying at the same layer and to switching to a “+1” or “-1” layer. Perhaps, this is not necessarily the best action in terms of QoE, but there is no complete HAS QoE model to be able to build the perfect optimization function. However, the optimization goal of low amplitude between successive layers holds. On the contrary, strategy W+S tends to select many layer 2 segments, which is explained by its goal to minimize the switches and thus operate at a stable but safe level. We have also tested a “W+A” optimal strategy (not mentioned in Section 8.3.1), but this has been found to cause too many quality switches; so, it was not considered any further.

It is important to note that no optimal strategy is considered “better” than the other. They all represent how different optimization objectives behave under varying bandwidth conditions. However, once a validated multi-parameter QoE model for HAS becomes available in the future, the optimization problem could be revisited in order to consider not only the most decisive influence factors, but also the optimum weight per influence factor.

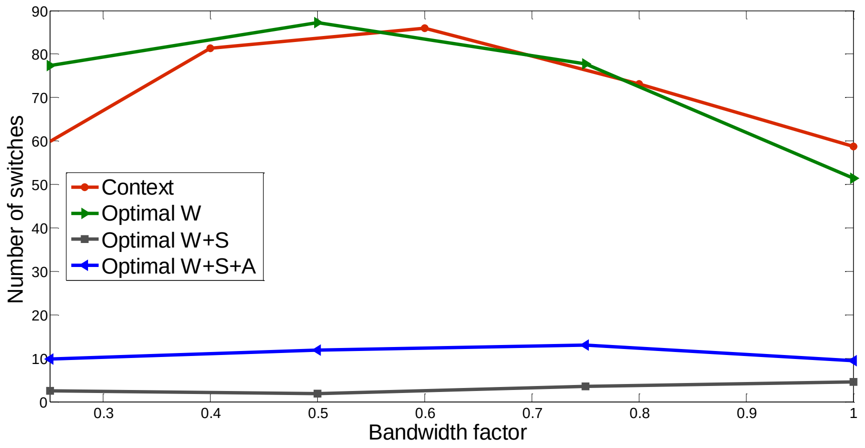
In terms of quality switches caused, which is another QoE impairment factor, the context-aware strategy and the optimal "W" strategy cause the highest number of switches, since they do not take measures to prevent them (see Figure 50d). On the contrary, the optimal W+S and optimal W+S+A strategies cause the least number of switches. Between the last two, W+S+A causes more switches, as it puts equal priority to mitigating switches and keeping the amplitude of any switches at a low level.



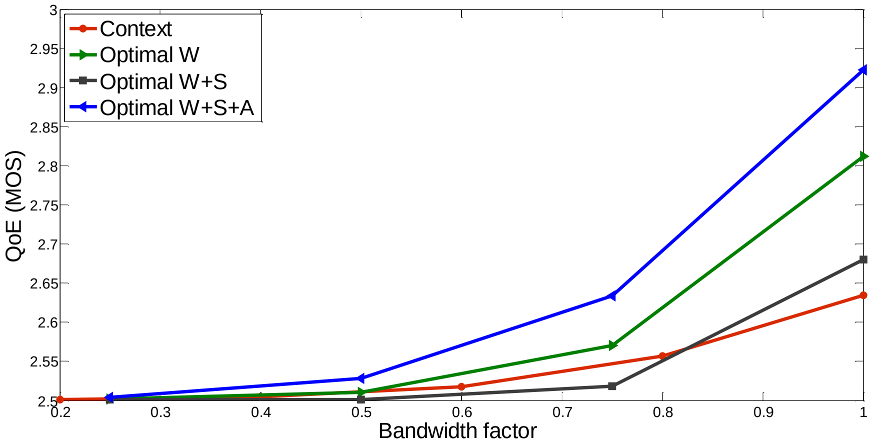
(a) Percentage of time spent on layer 1.

(b) Percentage of time spent on layer 2.

(c) Percentage of time spent on layer 3.



(d) Number of switches.



(e) QoE.

Figure 50: Simulation results for various bandwidth factors for the three optimal cases W / W+S / W+S+A and the context-aware strategy.

8.4.4 The impact of unreliability of context information

In this section we study how unreliability in the context information influences the probability of having a stalling event. In other words, we study how risky the proactive HAS strategy is to lead to a stalling, when accurate information about the outage starting point is missing or when it is impossible to have this information on time.

For the purposes of this experiment, we assume that the buffer of the user is not limited, and therefore the user will continue to download as many bits as its connectivity to the base station allows. As a consequence, the starting point of the outage plays an important role, since the further away it is from the vehicle's current location, the fuller the buffer of the client will be under normal circumstances up to that point. Thus, also the stalling probability will be lower. Overall, this study evaluates to what extent an unexpected outage is mapped to a stalling probability.

The results under this perspective are presented in Figure 51. As expected, the further away the outage, the less the stalling probability. This means, that the impact of unreliability of context information is smaller, when there is more time ahead for the user to react.

Nevertheless, even though we assumed an unlimited buffer, it might be more meaningful to conduct the same study assuming a limited buffer size of the client's application, which is a more realistic assumption. In that case, we would expect that the starting point of the outage would not play such a crucial role to the stalling probability, but the maximum size of the buffer would. Note that a normal value for an upper threshold in the number of buffered segments would be 50 segments. However, this study still provides some insights about the impact of unexpectancy regarding the outage starting point.

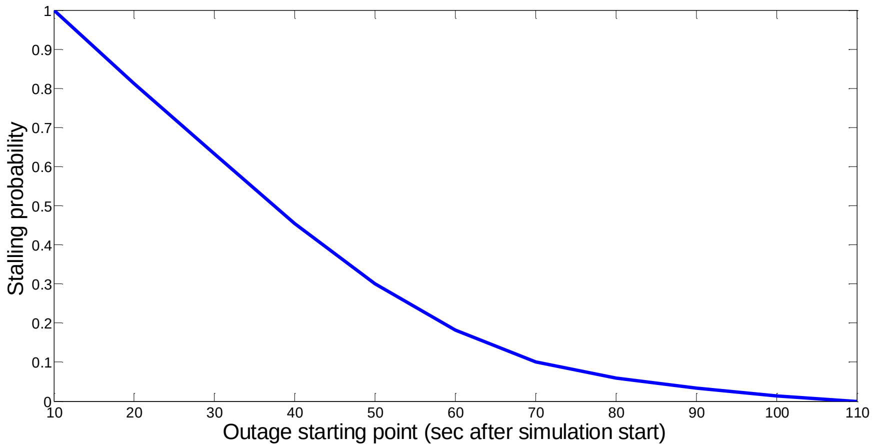
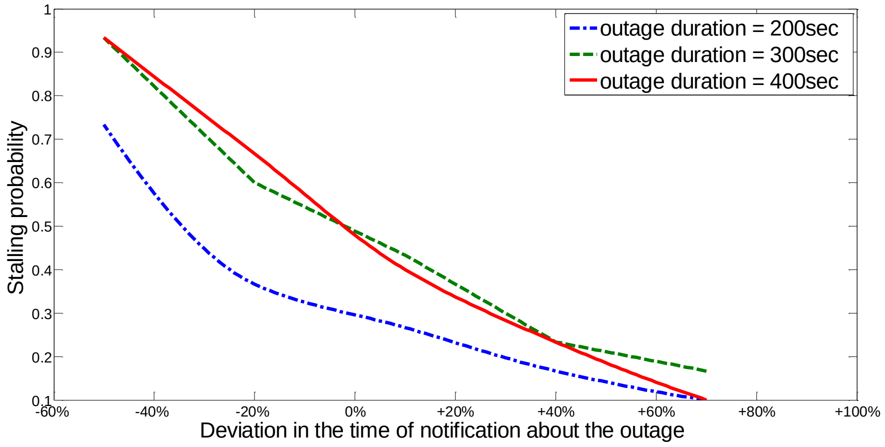


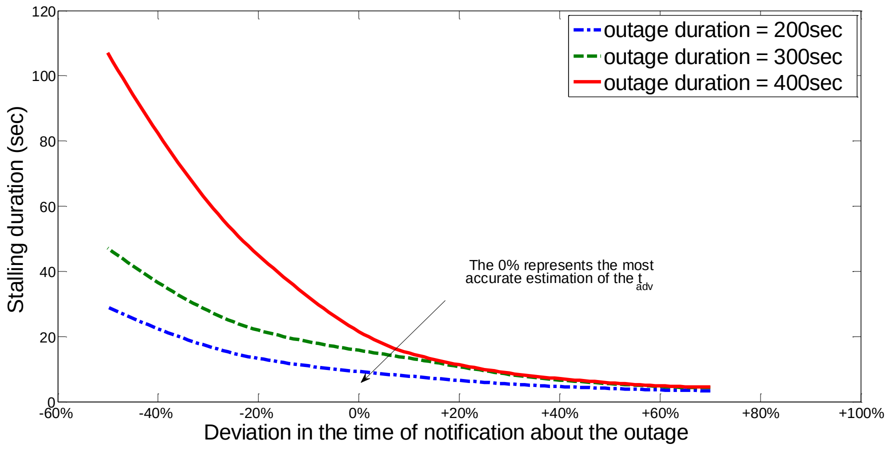
Figure 51: The impact of the outage starting point on the stalling probability.

Next, we would like to investigate what happens if the context information is not communicated to the client as 100% accurate or, similarly, if it is not communicated early enough in advance (so it is accurately communicated but with some delay). Specifically, we assume that the information about the t_{adv} deviates from its mean value, as this was estimated in Section 8.4.1. This mean value is considered to represent a “0% deviation” in the following figures. From Figure 52a and Figure 52b, which represent the stalling probability and stalling duration respectively, we draw two main conclusions. Firstly, we confirm that the mean values of t_{adv} are not enough to prevent a stalling, due to the fact that standard deviations have not been taken into account. In fact, as presented in Section 8.4.1, the standard deviations are higher for larger outage lengths and thus we observe higher stalling probabilities for the 0% values (compare the three plots per figure).

A second important conclusion, which is the emphasis of this simulation study, is that a potential uncertainty in this context information can lead to inevitable stallings. This is interpreted both in terms of stalling probabilities and stalling lengths. This emphasizes the need for accurate and timely context information, which also takes into account statistical metrics such as the standard deviation.



(a) Stalling probability.



(b) Stalling duration.

Figure 52: Stalling effects when t_{adv} deviates from its mean value.

8.5 Conclusions and future work

In this chapter, we have presented an enhanced context-aware HAS strategy, complementary to any standard HAS approach implementation. The proposed policy can successfully help a client's application be better prepared for an inevitable service outage and therefore be in the position to proactively minimize any negative impact on the viewing experience. Since HAS is considered a major trend in HTTP video streaming these days, we believe that proactive strategies such as the one proposed here will become available in real systems.

The proposed HAS strategy may run independently at the user device, relying solely on information that can be collected in a realistic environment. Therefore, it would be feasible to implement a smart "app" that runs at the user device and utilizes context and cross-layer information (map, speed, GPS, etc.). Such an app could then give the streaming user the power to prevent or minimize video stalling in light of a well-known coverage hole, such as a tunnel or a metro line.

Furthermore, the idea introduced in this study can be seen as a video stalling alert mechanism (at t_{adv}), and as such, it can be exploited in multiple ways. The proposed HAS strategy described here is just one possible method, but other approaches may be also possible: e.g., a strategy altering the buffer size of the client, or, even the implementation of a pop-up notification at the user device, warning about an inevitable video stalling and requesting for a user action.

It is also worth noting that even though this work focused on outage conditions of zero bandwidth, we could easily extend this solution to a more general problem where bandwidth may be insufficient (but not zero). Similarly, the same problem could be adjusted for cases of an imminent service disruption such as a handover, where the aforementioned HAS strategy can help prevent stallings during the disruption period (i.e., the handover period). This may become possible by exploiting handover-hinting information, a priori. In this way, the user will be better prepared for a potential interruption in his viewing experience.

It would be also interesting as future work to study a scenario of more than one mobile video streaming users using HAS, and investigate how the decisions of one user potentially affect the others. Stability and fairness issues, together with QoE analysis would be of great interest in this case. Also, in future works, it would be interesting to study HAS scenarios that rely on different context information, such as social context (e.g., Flash Crowd formation) or economic context (e.g., adjusting video consumption to a user's data plan). Furthermore, it would be valuable to make the proposed strategy more robust to QoE fluctuations.

Finally, as a general comment, we would like to point out that this work could be revisited once a standard QoE model for HAS becomes available. In that case, we could have the opportunity not only to produce a more accurate optimization problem, but also to enhance the proposed HAS strategy, focusing on the key factors that mostly influence the users' QoE.

9. QoE-SDN APP: A RATE-GUIDED QoE-AWARE SDN-APP FOR HTTP ADAPTIVE VIDEO STREAMING

While video streaming has dominated the Internet traffic, Video Service Providers (VSPs) compete on how to assure the best QoE to their customers. HAS has become the de facto way that helps VSPs work-around potential network bottlenecks that inevitably cause stallings. However, HAS-alone cannot guarantee a seamless viewing experience, since this highly relies on the Mobile Network Operators' (MNOs) infrastructure and evolving network conditions. Software-Defined Networking (SDN) has brought new perspectives to this traditional paradigm where VSPs and MNOs are isolated, allowing the latter to open their network for more flexible, service-oriented programmability. This chapter takes advantage of recent standardization trends in SDN and proposes a programmable QoE-SDN APP, enabling network exposure feedback from MNOs to VSPs towards network-aware video segment selection and caching, in the context of HAS. A number of use cases, enabled by the QoE-SDN APP, are designed to evaluate the proposed scheme, revealing QoE benefits for VSPs and bandwidth savings for MNOs.

9.1 Introduction

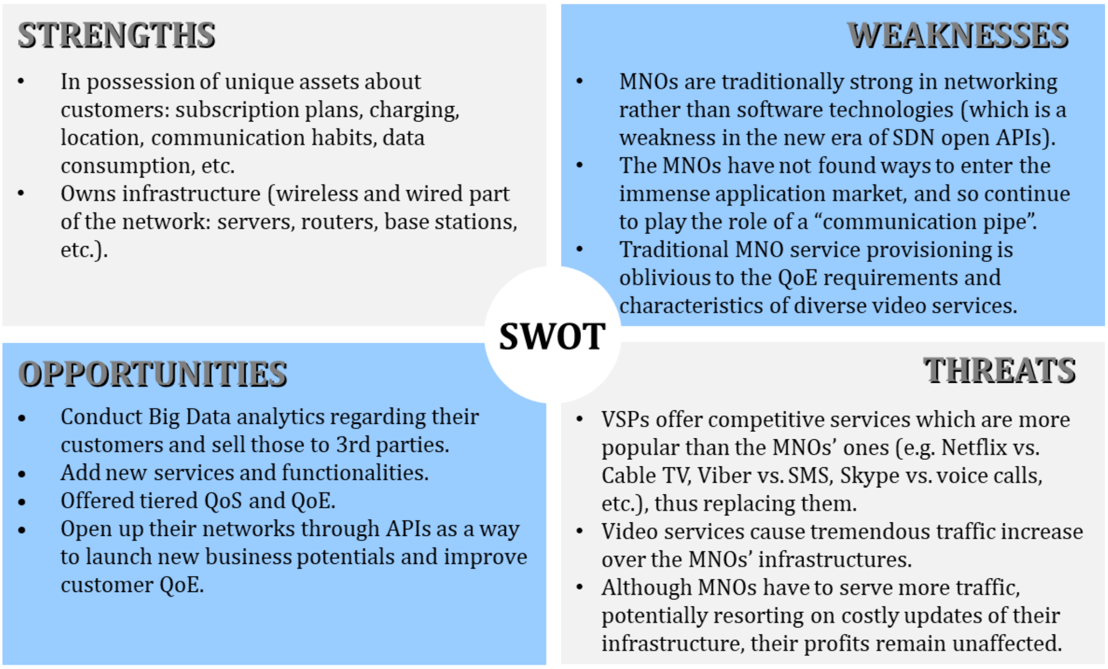
The emerging 5G networks are expected to enable a service ecosystem that facilitates new business opportunities, supporting also market players that do not necessarily own a network infrastructure, such as verticals and service/application providers. Such a 5G paradigm will scale-up further traffic volumes due to the mass adoption of content-rich multimedia applications and cloud services, introducing stringent service requirements in dense areas and on the move [84]. Alongside the launch of new 5G services including massive Internet of Things (mIoT), vehicular, and critical communications, etc., 5G networks will diversify the desired performance requirements in terms of throughput, latency, jitter, etc. This plethora of 5G services creates pressure for MNOs who cannot simply react by overprovisioning the network infrastructure, since the service race for the same set of resources is endless and the associated infrastructure cost is tremendous. Instead, to assure that the best experience is always assigned and follows a user, i.e., irrespective of location and network conditions, enhanced intelligent QoE mechanisms are needed considering the service type specifics and network conditions. Regardless of this immense potential, MNOs continue to offer only a "communication pipe", while being in search for new business models to allow them to enter the service/application provider market.

As multimedia services are dominating the mobile economy, an ever-increasing number of VSPs such as Netflix, Amazon, YouTube, etc. is expected to contribute towards a threefold grow of IP video traffic by 2021 [109]. New opportunities for video-related services still arise, especially with 5G, e.g., augmented and virtual reality video, but also outside the entertainment business with various verticals dependent on video such as e-health, security, safety, etc. Currently, VSPs offer OTT services considering the underlying infrastructure as a "black box" supporting best-effort services. HAS has appeared as a work-around solution of VSPs to confront network bottlenecks by dynamically controlling the rate at which video is offered, with the ultimate goal to avoid stalling events, which constitute the most crucial QoE degrading factor [70]. Despite the success in mitigating stallings, HAS may lead to an inevitably sub-optimal solution, since: a) quality adjustments are done re-actively after the service has already degraded, b) HAS tries to overcome a network problem without having any network control, and c) it relies on the subjective and isolated user perception regarding bandwidth availability.

The high competitiveness in the VSP market as well as the large business potential encourage service providers to find new means to offer higher QoE to their customers. The World Economic Forum recognizes that MNOs need to launch new business models, where they partner directly with various vertical markets (e.g., VSPs), in the direction of transforming their networks into more flexible, open, and customized infrastructures, as well as providing differentiation in a software-based way [110]. MNOs can therefore exploit their exclusively owned assets and capabilities, namely a) user information, b) network conditions, and c) technological options relative to their infrastructure, to create and offer additional services. Leveraging the benefit of such information and by opening their networks for collaboration, MNOs can form new business models considering network, user and service intelligence (e.g., regarding congestion and location, big data related to users, etc.) as well as open Application Programming Interfaces (APIs), enhancing the VSPs' capabilities beyond just application-level parameter control [111].

Currently, SDN [112]-[113] facilitates programmability and openness, enabling VSPs to interact with the network layer via open APIs, which allows MNOs and VSPs to build a close collaboration with a positive value for both stakeholders. In particular, the benefits of such a collaboration paradigm via the means of SDN are identified as: a) VSP customers are served with better QoE, enabled by the direct interaction among VSPs and MNOs, b) application/service-awareness allows MNOs to manage network resources more efficiently, and c) MNOs can get into the revenue loop of the APP market, offering big data and QoE-related information through their open APIs to third parties.

A SWOT analysis (Strengths, Weaknesses Opportunities, Threats) from the MNOs' perspective is provided in Figure 53, elaborating on the Weaknesses and Threats in the current Telecom status quo (where MNOs and VSPs are isolated), but also on the Strengths and Opportunities that arise from eliminating such an isolation.



STRENGTHS

- In possession of unique assets about customers: subscription plans, charging, location, communication habits, data consumption, etc.
- Owns infrastructure (wireless and wired part of the network: servers, routers, base stations, etc.).

WEAKNESSES

- MNOs are traditionally strong in networking rather than software technologies (which is a weakness in the new era of SDN open APIs).
- The MNOs have not found ways to enter the immense application market, and so continue to play the role of a "communication pipe".
- Traditional MNO service provisioning is oblivious to the QoE requirements and characteristics of diverse video services.

OPPORTUNITIES

- Conduct Big Data analytics regarding their customers and sell those to 3rd parties.
- Add new services and functionalities.
- Offered tiered QoS and QoE.
- Open up their networks through APIs as a way to launch new business potentials and improve customer QoE.

THREATS

- VSPs offer competitive services which are more popular than the MNOs' ones (e.g. Netflix vs. Cable TV, Viber vs. SMS, Skype vs. voice calls, etc.), thus replacing them.
- Video services cause tremendous traffic increase over the MNOs' infrastructures.
- Although MNOs have to serve more traffic, potentially resorting on costly updates of their infrastructure, their profits remain unaffected.

Figure 53: SWOT analysis from the MNOs' perspective.

The current work incentivizes and provides a technologically feasible realization of an MNO-VSP collaboration, where feedback from the MNO is enabled and application-awareness is enforced. A novel QoE-SDN APP is proposed, which can be flexibly programmed and customized to assure the desired QoE for verticals, VSPs and OTT providers, relying on the specifications of the SDN paradigm. The analysis considered in this study focuses on the case of video-on-demand with the objective to enhance the HAS paradigm. In particular, in our approach a feedback mechanism is facilitated from the MNO to the VSP, in order to enhance user QoE. This QoE enhancement is achieved through proactive video selection and encoding, which accounts for the user movement and the potential network conditions in the process of assigning the required video encoding rate that reduces stalling probability. We complementary explore the use of Multi-access Edge Computing (MEC) [114], which can cache HAS segments in advance based on forecasted user mobility in order to enhance QoE, while allowing MNOs to utilize the network resources more efficiently. We formulate an optimization problem with the objective of improving the user QoE. Moreover, we propose three novel use cases in the context of HAS, unlocked by the proposed framework, which incorporate mobility and rate guidance towards a better video encoding selection and a more efficient video segment caching. A set of simulations in a realistic and challenging mobile cellular environment demonstrate the added value of the proposed scheme, in terms of QoE amelioration of VSPs' customers and network resource savings for the benefit of the MNOs.

The remainder of this chapter is summarized as follows. In Section 9.2, we review the related state of the art in the areas of QoE provisioning in SDN-based environments. Section 9.3 describes the proposed QoE-SDN APP, and the supporting SDN-based architecture, including required APIs, components and operations. Then, Section 9.4 models the system and formulates an optimization problem of video encoding towards improving user QoE and presents a mobility forecasting and rate estimation logic that approach a real-time solution to this problem. Furthermore, Section 9.5 describes three novel use cases in the context of HAS that are activated by the QoE-SDN APP, presents the evaluation environment and respective QoE indicators, as well as the evaluation results. Finally, Section 9.6 describes exploitation issues and concludes this chapter.

9.2 Related work

The importance of QoE as a significant performance measure from the user's perspective with respect to an application or service, which assists MNOs and application/service providers to understand the overall quality of their services has been discussed thoroughly in this thesis. Nevertheless, SDN brings some new perspectives to QoE monitoring and management. More specifically, SDN, via the means of open APIs, can offer programmability that enables service providers to obtain QoE measures regarding the offered applications as well as the capability to interact with the network, introducing adjustments on the networking resources considering also the application requirements.

Preliminary SDN-based solutions considering QoE concentrate on the core and transport networks taking advantage of the global network view to perform dynamic traffic steering and optimal CDN selection. In [115], a jointly optimized path assignment and service utility decision for multimedia flows is performed by OpenFlow considering the resource requirements of competing services. Similarly, [116] improves the QoE of video streaming applications using an SDN controller that monitors video QoE metrics at the client side and dynamically selects delivery nodes via the means of traffic engineering. In the context of HAS, [117] investigates three different network-assisted video streaming approaches: a) Bandwidth Reservation, where optimal bandwidth slices are assigned to video flows, b) Bitrate Guidance, where optimal video bit rates are estimated centrally and then enforced to the users, and c) hybrid approaches, that combine both. Such hybrid solutions are explored in [118]-[122]. SDNDASH [118] relies on an SDN-based management and resource allocation architecture with the goal to maximize the QoE per user considering heterogeneous QoE requirements. Each user's adaptation logic is then based on a combination of optimal bit rate recommendations and buffer levels. As an extension to this work, [119] proposes a more scalable architecture, called SDNHAS, which estimates optimal QoE policies for groups of users and requests a bandwidth constraint slice allocation, while providing encoding recommendations to HAS players. Furthermore, [120] proposes a network application controller, called Service Manager, which oversees video traffic and fairly allocates network resources among competing HAS flows, while enforcing QoS guarantees. A target bit rate is assigned to each client, which can be used as a reference in their adaptation logic regarding the maximum encoding they should request. Then, [121] considers caching, and proposes an SDN-based Adaptive Bit Rate (SABR) architecture, where video users are informed regarding each cache's content as well as get a short-term prediction of the bottleneck bandwidth to reach each cache, so that their adaptation decisions are better. In parallel, OpenFlow guides routing between clients and selected caches. Finally, [122] proposes an OpenFlow-assisted QoE Fairness Framework (QFF), with the objective to fairly optimize QoE among HAS clients with heterogeneous device requirements, expressed via bitrate-to-QoE utility functions. Our QoE-SDN APP adopts joint network and application programmability via the means of open APIs, but in contrary to all previous approaches, we concentrate our efforts on mobile networks, which require a higher flexibility due to constantly evolving network dynamics. Moreover, our approach guides HAS-related decisions considering also longer-term forecasted information regarding user mobility and network load. A point-by-point comparison with aforementioned SDN-based HAS solutions is presented at Table 17.

Table 17: Comparison of SDN-based HAS solutions.

Solution	Approach	Network	Prediction	HAS strategy	Asset	Weakness	SDN add-on
A. Bentaleb et al. [118]-[119]	Hybrid	Fixed	No	Upper bounded bit rate recommendation and buffer level	Optimized QoE per user, User heterogeneity support	A new user communication interface is required	Internal and external SDN-based resource management components
J. W. Kleinrouweler et al. [120]	Hybrid	Fixed	No	Target bit rate pushed to each user	Explicit adaptation assistance with fairness criteria	Users have to cooperate with the Service Manager	HAS-aware Service Manager
D. Bhat et al. [121]	Hybrid	Fixed	Short-term prediction (ARIMA)	User assisted with information about cache location and link bandwidth	Video segment decision remains at the user's control (scalable)	Overhead due to both bandwidth and cache occupancy monitoring	SABR module
P. Georgopoulos et al. [122]	Hybrid	Fixed	No	Optimum bit rates that ensure fairness pushed to users	Optimized QoE, Heterogeneity support, Fairness	Utility functions need to be pre-calculated and stored for all video content at each resolution	Orchestrating OpenFlow module
QoE-SDN APP	Bitrate Guidance	Mobile	Longer-term (cluster based)	Rate-guided, prediction-based	Network exposure feedback enabled, No change needed at HAS clients	Assumes VSP-MNO collaboration	QoE-SDN APP

For RANs, the notion of flexibility and programmability goes beyond the standard processes of routing and forwarding, due to mobility, load and radio conditions and, hence, the role of SDN is crucial for auguring QoE. One of the earliest proposals for softwarizing the access network (and not just the core network) has been elaborated in [123], where the “SoftRAN” vision is described. The SoftRAN architecture describes a software-defined controller that abstracts physical base stations, while it conducts radio access mechanisms such as load and interference management in a logically centralized manner. Other examples in the direction of “Software-Defined Mobile Networks (SDMN)” are described in [124], where the technical- and business-added value of such schemes is thoroughly analyzed. Furthermore, a flexible 5G RAN architecture based on software-defined control is proposed in [125], where a QoE/QoS mapping and monitoring function dictates the way in which the radio or core networks are (re)configured with respect to the decomposition and allocation of Virtual Network Functions (VNFs). However, the use of SDN in these proposals focuses on MNOs’ efficient resource management considering the requirements of the application but not actively interacting with third parties (e.g., VSPs), nor leveraging the capabilities of VSPs.

Assuring a desired QoE in mobile networks may also involve admission control and policy provision, where new connections will be restricted, or existing ones will be handed over, based on QoE criteria. Such mechanisms are explored considering femtocell networks in [126], where a “QoS/QoE mapper” creates a statistical profile of relevant QoS metrics (e.g., bandwidth availability) and maps this to user satisfaction, defining a QoE-based admission control policy. Moreover, in the context of HAS, [127] describes a novel mobile edge function for transcoding video segments on-the-fly, in the case that this requirement is triggered by a QoE assessor, while [128] introduces an SDN-enabled resource allocation mechanism, called UFair, to fairly orchestrate resources among competing HAS flows.

The adoption of SDN logic in a network can also serve the purposes of application awareness and data analytics. For instance, [129] envisions an architecture relying on a “Video Quality Application”, which queries information regarding video content, client information, and network data in order to help the operators better understand their network (e.g., congestion points) through QoE analytics. QoE analytics may also result in a user recommendation engine, as proposed in the case of the “u-map” system [130], where user collected subjective and objective quality metrics are uploaded in the u-map server, followed by feedback to the users regarding the performance of provided services in a specific region. In this study, we build-up on our previous work in [131], introducing a QoE-SDN APP that allows VSPs to program and control the desired QoE with the assistance of the MNO.

A collaboration model between OTT parties and ISPs is also described in [132], but from a revenue perspective, thus, proving the concept, viability and mutual benefit of such collaboration paradigms. Also, [133], explores the MEC paradigm, proposing a reference architecture for orchestration and management, where Channel State Information (CSI) is sampled to enforce service-level management.

The proposed QoE-SDN APP allows MNOs to dynamically provide network capability exposure feedback to the corresponding VSP based on mobility and rate forecasting mechanisms, proactively guiding in this way the video segment distribution towards particular edge caches as well as the video segment encoding, in order to avoid stalling events.

9.3 VSP-MNO collaboration architecture: The QoE-SDN APP

9.3.1 VSP-MNO collaboration possibilities

Multiple use cases can be envisioned depending on the level and type of interactions between the MNO and VSP parties. In Table 18 we describe all possible interaction types, as foreseen by the insightful position paper [134] and complement them with the description of concrete QoE management possibilities. As it can be seen in this Table, various and different management decisions can be taken based on the use case, e.g., either a) by the VSP provider (e.g., change the resolution of an HTTP Adaptive Stream - “Application self-optimization” use case), or b) by the MNO (e.g., priority in scheduling - “Application controls network” use case), etc.

Table 18: QoE management possibilities in the VSP-MNO collaboration paradigm.

Interaction type/Use case	Direction	QoE management possibilities
Application self-optimization	Information: MNO to VSP Control: within VSP	- Application tuning, e.g., live encoding of a video on the video server, based on information about the current or predicted status of the network (i.e., encoding will no longer be network-agnostic). Then QoE control will be possible through application means (e.g., change the video resolution / encoding or affect user application decisions).
Network self-optimization	Information: VSP to MNO Control: within MNO	- Update of the network infrastructure and anchor points based on traffic requirements imposed by VSPs. - Higher efficiency in resource usage (e.g., cell planning).
Application controls network	Information: MNO to VSP Control: VSP to MNO	- The VSP instructs the MNO about the handling of specific users’ flows (e.g., because a user is “premium”, or a flow requires attention). Then any network engineering mechanisms based on the different QoE requirements will be triggered. For instance: a) At access network: Admission control / Mobility management / etc. b) At core network: Change the bearer or policy of certain flows / Select S-/PDN-GW / Packet marking / Flow manipulation / etc.
Network controls application	Information: VSP to MNO Control: MNO to VSP	- Network asks application to virtualize a critical function / a server / a cache / etc. at a specific problematic location of the infrastructure.
Mediation	Information: MNO to VSP and VSP to MNO Control: VSP to MNO and MNO to VSP	- Joint optimization of network and application by an intermediate central intelligent QoE manager.
Offline info sharing	Information: MNO to VSP	- Any potential use cases enabled by data analytics.

As a characteristic example of a QoE management cycle in the VSP-MNO collaboration paradigm, we describe the scenario where an MNO exposes information regarding its assets and current state to the VSP, so that the latter can impose more informed decisions that will be actualized by the MNO (i.e., “Application controls network” use case). In order to describe the logic behind this paradigm, we assume the use of the three components of the QoE management cycle described in Section 3.3, i.e., the QoE-Controller, QoE-Monitor and QoE-Manager.

Key in the “Application controls network” scenario is the QoE-Controller, which installs “monitoring rules” at selected elements in the infrastructure network. The goal of these rules is to collect particular input parameters that can be used by the QoE-Monitor to estimate QoE for a specific service and/or user. The QoE-Monitor provides the parameters for such rules and makes QoE estimations. Last, the QoE-Manager is responsible for controlling the network in an elastic, QoE-driven way.

The overall scenario operates as follows (the flow of this procedure is depicted in Figure 54):

1. The QoE-Monitor, which is programmed by the VSP, periodically and/or on demand requests the collection of specific KPIs from the MNO, through the QoE-Controller.
2. These requests are translated by the QoE-Controller into plausible rules/requirements for the MNO and passed down to the network infrastructure.
3. The MNO collects the respective data by appropriate network elements and reports them back to the QoE-Controller.
4. This information is sent back to the QoE-Monitor, where it is translated to QoE “language” (i.e., a MOS score or a quality metric), via a VSP-programmed QoE assessment logic.
5. Based on the current use case, and if QoE is below a threshold (reactive case) or if an imminent problem is identified (proactive case), network (and/or application) QoE management mechanisms are triggered by the VSP.
6. These decisions are actualized by the QoE-Manager. In the “Application controls network” use case, the actualization is done by the MNO via appropriate instructions.

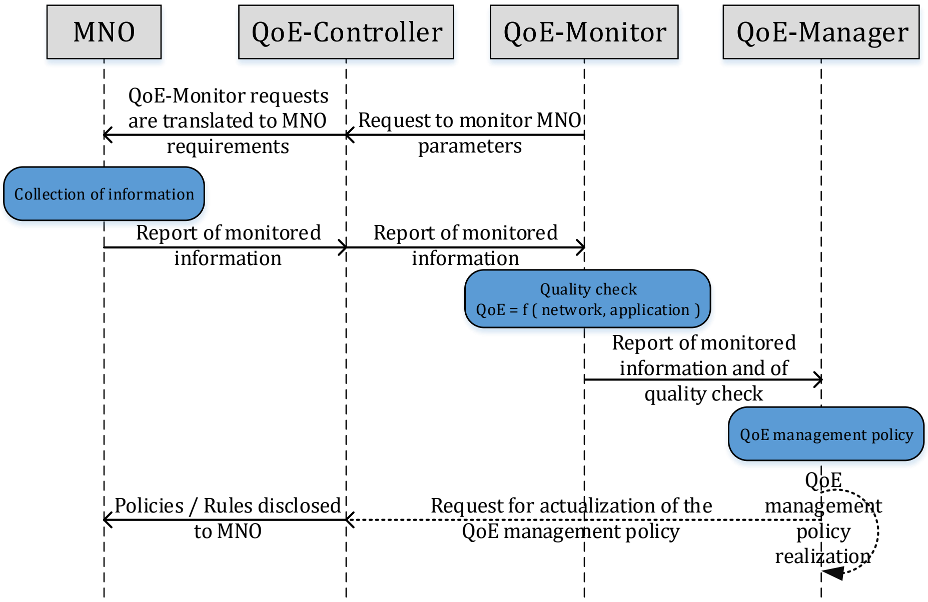

MNO
QoE-Controller
QoE-Monitor
QoE-Manager
QoE-Monitor requests are translated to MNO requirements
Request to monitor MNO parameters
Collection of information
Report of monitored information
Report of monitored information
Quality check
QoE = f (network, application)
Report of monitored information and of quality check
QoE management policy
QoE management policy realization
Policies / Rules disclosed to MNO
Request for actualization of the QoE management policy

Figure 54: Abstract QoE management flow cycle in a VSP-MNO collaboration paradigm.

Having described the abstract QoE management logic in a VSP-MNO collaboration paradigm, next we propose a concrete SDN-based architecture that actualizes this paradigm in a realistic way.

9.3.2 QoE-SDN APP functions and architecture

The QoE-SDN APP relies on the SDN architecture [112]-[113] allowing the SDN controller to maintain a corresponding APP template. Such template offers VSPs the opportunity to program their QoE requirements and QoE assessment logic once subscribed. VSPs can then use the QoE-SDN APP to enhance their video segment encoding and distribution procedures by getting network feedback exposed by the MNOs. The VSP can contact the MNO to request the setup of the QoE-SDN APP via conventional 3GPP management system means, i.e., through the Network Exposure Function (NEF) [135] and an open API, such as GSMA OneAPI [136]. The NEF provides authentication and secure access for VSPs, charging, as well as the means for requesting the QoE-SDN APP. Once a VSP QoE-SDN APP request is authorized, the network management system installs the corresponding SDN-related functions within the SDN controller and within the corresponding Network Elements (NE), e.g., eNBs, via the Coordinator function. The Coordinator is contacted through the conventional Itf-N interface and Element Manager.

The basic functions of the QoE-SDN APP within the SDN controller are the following:

- **VSP QoE Control Agent** is a function that allows VSPs to collaborate with the underlying MNO’s infrastructure and resides within the SDN controller. It facilitates the communication and control between the two parties, i.e., providing feedback to the VSP regarding required encoding rates, and control capabilities related to the data plane within the MNO infrastructure, here, in the context of HAS. The QoE control agent uses a relative global view of the underlying network, i.e., a relative RAN Information Base (RIB), considering the abstracted resources allocated to the particular VSP via a Virtualizer component.
- **QoE Assessment Logic** is the core of the QoE-SDN APP, which can be programmed by the VSP according to the application characteristics and requirements. In particular, the VSP can provide the QoE estimation model and associated parameters, the desired monitoring metrics as well as the policy for retrieving such metrics, e.g., monitoring periodicity, etc. These QoE estimation models, which are different per service type, are programmed by the VSPs, therefore, they can be easily updatable and manageable, as they constitute proprietary (VSP-owned) or standardized (recommended by standardization bodies) software functions.

  The QoE assessment logic is responsible for: a) determining the QoE per application using the MOS scale or appropriate application-specific KPIs, e.g., stalling events in case of video streaming, b) instructing the Data Plane Control Function to introduce alternations into the allocated network resources with the purpose of maximizing the perceived QoE, and c) determining guidance decisions for the VSP regarding the encoding rate and caching strategy that should be adopted considering future user mobility and network load. The QoE assessment logic relies on feedback collected by the Data Plane Control Function from NE agents or from the MNO management system.

  Another significant process of the QoE assessment logic is user mobility forecasting that determines future user positions considering the current location, duration of a session and gravity points, i.e., areas with higher user concentration. Based on such forecasted users’ locations, and with the assistance of a rate estimation function, the traffic load can be determined at particular RAN points with respect to time, which can be used to guide the encoding rate of VSP content and the video segment distribution, considering also potential re-configurations of the network resources.

- **Policer** defines the policy applied to the allocated resources of the VSP and corresponding QoE-SDN APP.

The Data Plane Control Function operates on the allocated resources carrying out all QoE-SDN APP processes related with data acquisition, video segment distribution and potentially network resources' programmability. The data acquisition process takes place periodically or optionally, on-demand, and can also adjust the input type of collected QoE data including its nature, i.e., real-time measurements or statistics, which are retrieved via agents of specified NEs located in the RAN and in the core network that can capture service-related parameters. These agents can be dynamically configured considering topology changes, e.g., upon a user movement. The QoE-SDN APP functionalities within each NE include a NE VSP QoE Control Agent and Policer, which are responsible for carrying out QoE monitoring and policy processes on the allocated resources, i.e., relative NE RIB, within the NE. In this way, each NE VSP QoE Control Agent "represents" the VSP tenant over this NE. An overview of the QoE-SDN APP architecture is illustrated in Figure 55.

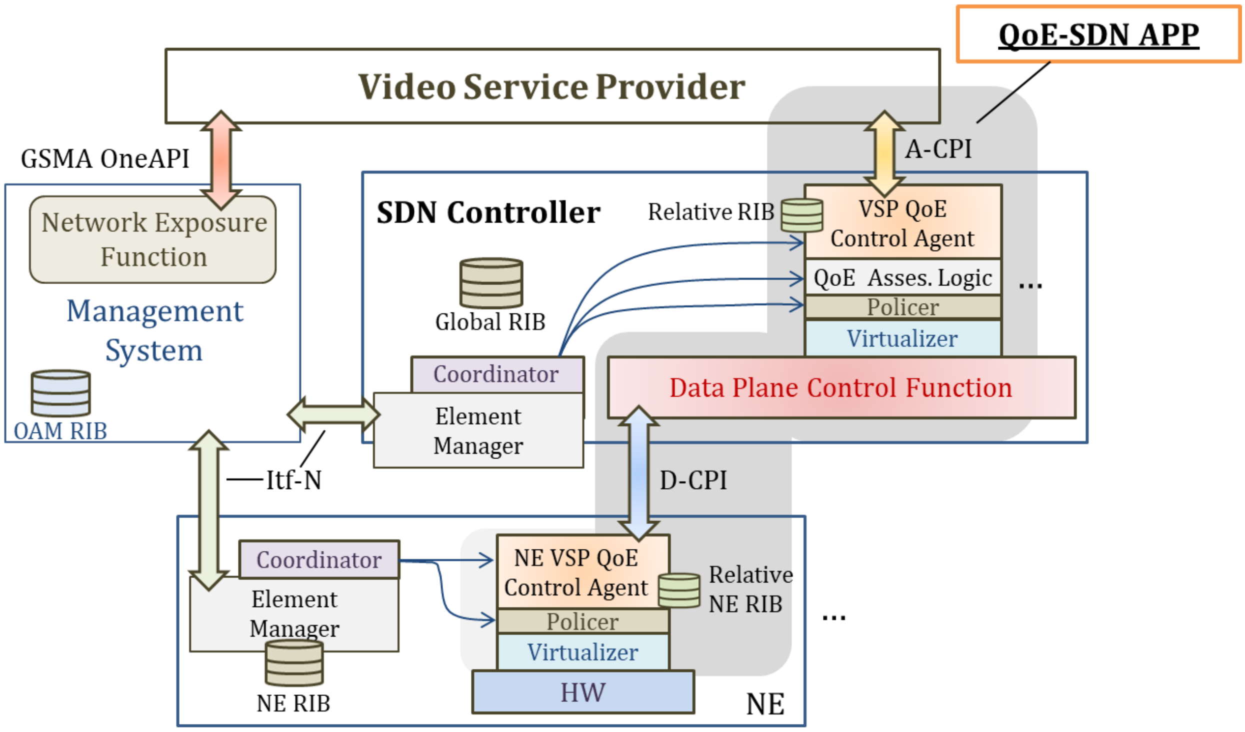


Figure 55: QoE-SDN APP functions and architecture.

Moreover, the SDN controller can communicate with the 3GPP network management system in order to collect the conventional network monitoring information such as interference, load and other KPIs, which can be stored in a RIB creating a global network view.

The Application-Controller Plane Interface (A-CPI) can facilitate programmability for the VSPs in order to program the QoE assessment logic, while the Data-Controller Plane Interface (D-CPI) offers the interaction means between the SDN controller and the corresponding NE of the MNO, carrying out QoE monitoring as well as resource and policy re-configuration instructions.

Related to the discussion of Section 9.3.1, we may map the QoE assessment logic to the QoE-Monitor and QoE-Manager components, while the Data Plane Control Function resembles the functionality of the QoE-Controller. However, the borders of each component are not strictly defined.

9.4 System model and problem formulation

9.4.1 Generic problem formulation

The system under study, where the QoE-SDN APP will be integrated, is considered an Orthogonal Frequency-Division Multiple Access (OFDMA) cellular network (e.g., LTE) that consists of a ring topology of tri-sector eNBs. Each eNB serves the mobile UEs that are located within its coverage, while handovers between eNBs are enforced as UEs move. Each eNB is co-located with a MEC server, used for caching of video segments.

Initially, we formulate the per-user segment selection strategy of HAS logic as a Knapsack optimization problem, using the optimization problem of Section 8.3.1 as a basis (the notation slightly changes). We consider a video split into $s = 1..S$ video segments, while each segment is available in $l = 1..L$ quality layers. Moreover, there are $u = 1..U$ mobile users in the system and $m = 1..M$ eNBs (and equal MEC platforms). In this Knapsack problem, the value which quantifies the level of importance associated with each decision is the quality layer. The higher the index of the quality layer, the more valuable the solution. On the other hand, the cost of each decision is the size of the video segment needed to transfer to satisfy it. The basic parameters of this problem are represented as:

- v_{sl} = the value associated with segment s of quality l (here: quality is the quality layer index).
- c_{sl} = the cost associated with segment s of quality l (here: the size of segment s of quality l).
- $V(t)$ = the total data downloaded until moment t.
- R_u = the achieved data rate per user u with respect to the eNB where the user is attached (in bps).
- D_k = the deadline of segment k, meaning that segment k needs to be downloaded by that moment, otherwise a stalling will occur.

In order to estimate $V(t)$, the information about the R_u is required, so:

$$V(t) = R_u * t \tag{9-1}$$

Moreover, the deadline of segment k can be found as follows:

$$D_k = T_0 + k\tau, \forall\, k = 1..S \tag{9-2}$$

where T_0 is the video start-up delay (initial delay), and τ is the segment duration. The unknown optimization variable in this problem is x_{musl}, which represents the selection of a segment with index number s of quality l that is destined for user u from the eNB/MEC m. It is a binary variable, namely a segment with index number s of quality l is either selected or not. Using the above notation, the optimization problem of segment selection is formulated as follows:

$$maximize \sum_{m=1}^{M}\sum_{u=1}^{U}\sum_{s=1}^{S}\sum_{l=1}^{L} v_{sl}x_{musl} \tag{9-3}$$

$subject\ to$:

$$x_{musl} \in \{0,1\} \tag{9-4}$$

$$\sum_{l=1}^{L}\sum_{m=1}^{M} x_{musl} = 1, \ \ \forall\, u = 1..U, \forall\, s = 1..S \tag{9-5}$$

$$\sum_{m=1}^{M}\sum_{u=1}^{U}\sum_{s=1}^{k}\sum_{l=1}^{L} c_{sl}x_{musl} \leq V(D_k),\ \ \forall\, k = 1..S \tag{9-6}$$

Equation (9-3) expresses the optimization goal of maximizing the quality layers of the segments selected, as those will bring higher video bit rates to the users. In terms of the constraints imposed, equation (9-4) expresses the binary nature of the unknown variable x_{musl}, while equation (9-5) mandates that each user can request a segment at only one quality layer and from only one MEC platform. Finally, the last constraint (9-6) expresses the requirement that all segments need to be downloaded before their deadline (on the right-hand side of (9-6) $V(D_k)$ expresses the maximum amount of data that can be downloaded until the deadline of k, so as to prevent a stalling). This optimization problem restricts the existence of any stalling events, due to constraint (9-6). Therefore, if a stalling event is inevitable, then the optimization problem will be infeasible, namely it will not be solved by an optimizer such as GUROBI.

All parameters in this problem are available when the proposed architecture is used. Specifically:

- The QoE-SDN APP logic in the proposed architecture ensures the exposure of feedback information about the expected rate per user, R_u, which is in turn used to estimate the downloaded data per user, $V(t)$ using equation (9-1). The same information about the R_u is additionally used to estimate the initial delay per user (T_0 in equation (9-2)).
- The rest of the input is known even in the state of the art case, namely information about the video parameters v_{sl} and c_{sl} are provided by standard HAS protocols.

Nevertheless, solving this optimization problem requires a priori perfect knowledge of R_u for all users, and for the whole duration of the video streaming session (namely until all segments S are downloaded), which is impossible in real networks. Also, each user's attached eNB, or equivalently each user's serving MEC, need to be known a priori, for the purposes of caching the appropriate segment (based on R_u) to the appropriate location.

What is more, in cases where stallings are inevitable, a solution to this problem will be infeasible; therefore, it makes sense to propose novel algorithms that reduce stalling probability (see use cases in Section 9.5). Finally, it is an NP-hard problem, not complying to the real-time constraints that network operation mandates, especially when scalability is an issue (e.g., many users in the system).

Next, a mobility prediction and rate estimation function are proposed, which manage to estimate R_u in a real-time basis per user. Then, in Section 9.5 some novel use cases are proposed that solve the segment selection and segment caching problem described above in a realistic and real-time fashion, namely using information that can be realistically acquired using the proposed architecture. These use cases also serve as a proof of concept and demonstrator of the potential of the proposed architecture.

9.4.2 Mobility prediction function and rate adaptation heuristics

As commented above, this section provides a solution to the per-user segment selection and segment caching problem. In particular, a mobility prediction solution and a rate adaptation algorithm are provided for the segment-to-quality layer and segment-to-eNB/MEC mapping problems. For this purpose, we have implemented a mobility prediction algorithm based on the Self Similar Least-Action Walk (SLAW) mobility model [137], taking advantage of the "clusters" introduced by this model, as described next.

The SLAW mobility model is a realistic mobility pattern based on empirical studies of real-life human-walk traces. One main property of SLAW is the existence of gravity points or “clusters”, namely of popular points where users tend to accumulate with certain probability (“self-similar waypoints”). This mobility model provides a realistic outlook in terms of network traffic per square meter, as compared to random mobility models. A real-life example of the behavior of the SLAW model is that users outside of a mall (i.e., a gravity point) would tend to go inside this mall.

A SLAW mobility pattern is characterized by multiple parameters in terms of mobility trace generation, which are: 1) the duration of trace generation, 2) the size of the mobility area, 3) the number of visit-able waypoints, 4) the minimum and maximum pause time of the mobile users and a levy exponent for pause time (parameter “beta”), 5) a “hurst parameter” determining the degree of self-similarity of waypoints, 6) a clustering range, and 7) a parameter “alpha” that determines the probability of selecting the next waypoint using the Least-Action Trip Planning (LATP) algorithm. Figure 56 presents an example of a produced SLAW mobility pattern using MATLAB. Overall, SLAW creates challenging network conditions, since many users tend to be accumulated close to each other, which means that one eNB will be asked to serve an un-proportionally large amount of traffic (as compared to less realistic random mobility patterns).

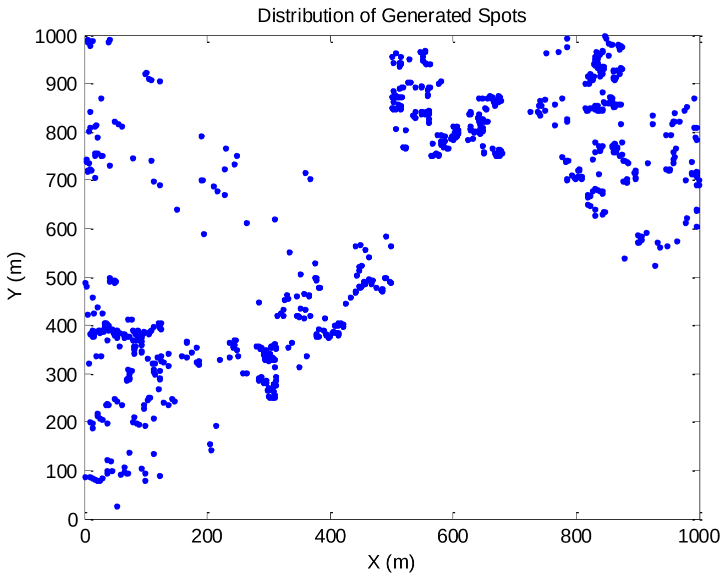


Figure 56: SLAW mobility model snapshot.

In the proposed QoE-SDN APP, mobility prediction is introduced to guide QoE control decisions at the VSP and network layer. The mobility prediction algorithm adopted is based on SLAW’s inherent characteristics and it runs per user, relying on information that the MNO has at its disposal, i.e., the popularity of visited locations and the user current positions. As far as the popularity of visited locations is concerned, this is available from statistics kept at the MNO regarding previously visited locations of all UEs. Regarding the UE current positions, these are already known by the MNO. Such information can be fed to the QoE assessment logic via the SDN controller, which communicates with the network management system via the Itf-N interface.

In detail, the mobility prediction algorithm uses as input the set w of visit-able waypoints and the set c of clusters, with the objective to find the next visited cluster per user, based on the user’s current position p. All clusters that a user can potentially visit are sorted by popularity, with the logic that more waypoints will be accumulated in the most popular clusters. Each user is going to visit a total of v clusters, subject to the trace generation duration. Then for each user, the algorithm estimates the distances $d(p, v)$ between the user's position p and the center of each yet unvisited cluster, c_k, ordering them in increasing distance from p. The predicted next movement will be towards the cluster center at the smallest distance out of this list, while the exact position for the next prediction interval will be a function of the user's velocity and direction. Therefore, the main concept of this mobility prediction algorithm is that users from one cluster will tend to travel towards the closest most popular cluster.

The operation of the SLAW-based mobility prediction is illustrated in Algorithm 1:

```
Algorithm 1: SLAW-based mobility prediction
- Set of all waypoints based on SLAW pattern: $w = \{w_i, i = 1..W\}$
- Set of all clusters $c = \{c_k, k = 1..C\}$, where $c_k = \{w_m, \dots, w_l\}$, so that
  $d(w_i, w_j) < clustering\ range$ for all $w_{i,j} \in c_k$
- Set of visited clusters: $v'$
- Starting user waypoint: $s \in v'$
- Present user waypoint: $p = (x_p, y_p) \in v'$
- $p \leftarrow s$
- Identify to which cluster $c_k$ the waypoint $p$ belongs, $v' \leftarrow c_k$
- Set $cluster\_ratio$ (percentage of clusters to visit), $velocity$,
  $prediction\_interval$
for $cluster = 1{:}C$
    - Calculate each cluster's popularity as the number of waypoints per
      cluster over the total clusters available: $P = \frac{|c_k|}{C}$
    - Order the first $\frac{C}{cluster\_ratio}$ number of clusters in descending popularity →
      set $v$ of clusters to visit
    - Calculate cluster centers $\bar{c}$: $\overline{c_k} = mean(w_m, \dots, w_l)$
end for
while $v$ is not empty do
    - Calculate distances from $p$ to the center $\bar{c}$ of all unvisited clusters $v$:
      $d(p, v) = \|p - \bar{c}\|^2$, for all $v \neq v'$
    - Order clusters in increasing distance omitting the one with the least
      distance (which is the current cluster)
    - The next movement prediction is towards cluster $c_k$ which is the first
      element of the previous vector
    - Future predicted position: $(x_f, y_f) = (x_p + velocity *$
      $prediction\_interval * cos\varphi, y_p + velocity * prediction\_interval *$
      $sin\varphi)$ where $\varphi = \tan^{-1}\frac{\overline{y_{c_k}} - y_p}{\overline{x_{c_k}} - x_p}$
    - $v' \leftarrow v' \cup c_k$
    - $v \leftarrow v - \{c_k\}$
end while
```

Based on the user mobility prediction we then estimate the corresponding data rate in order to identify and proactively handle congestion conditions in the RAN, considering bandwidth conditions on a cluster-basis, as elaborated in Algorithm 2. Algorithm 2 uses as input the mobility prediction estimations, which reveal the set of clusters that each user can potentially visit during a pre-defined future time window. Based on such information, it can approximate the rate for each user as the mean data rate of the cluster that it will reach. In this way, when a user moves from a low-congested to a higher-congested cluster, the estimated data rate will be conservative (ensuring no stalling events), i.e., it may be predicted lower compared to what each user would subjectively perceive, since this prediction will be based on the mean data rate of the to-be-visited cluster.

A similar idea may be found in [138], where a mobility-prediction-aware bandwidth reservation scheme is proposed. This scheme predicts when a user will perform handovers along his movement path, while a rate estimation scheme calculates the available bandwidth along this path in order to drive call admission control with QoS guarantees for ongoing calls.

Algorithm 2: Congestion-aware proactive rate estimation

```
- Set of future predicted positions per user $f = (x_f, y_f)$
- Set of cluster centers $\bar{c}$
for $user = 1: all$
   for $step = current\_tti: current\_tti + prediction\_interval$
          -     Read the next predicted position of the UE: $f = (x_f, y_f)$
          -     Find cluster $k$ closest to this position: $arg_k \min\{\|f - \bar{c_k}\|^2\}$
          -     Identify other users belonging to the same cluster $k$
          -     Estimate the mean data rate from all users in the cluster, $r$,
        during the latest second
          -     The predicted rate for this user for this step is equal to $r$
   end for
end for
```

Such rate forecasting estimates can then help the QoE assessment logic to guide VSPs to take proactive service provisioning decisions, as will be shown by the evaluation use cases in the next section. The MNO, in turn, is aware of the achieved data rate per user, as each user positively or negatively acknowledges the scheduled packets per TTI to the serving eNB.

9.5 Simulation setup and evaluation analysis

9.5.1 Simulation setup

The performance evaluation is carried out using the Vienna simulator, a 3GPP-compliant LTE system-level simulator [139], which inherently supports physical and MAC layer stacks (channel models, fast fading, scheduling, etc.). Moreover, various traffic types are supported in this simulator, namely VoIP, file download, web browsing, and video streaming (but not HAS). We have significantly extended this simulator implementing the proposed QoE-SDN APP introducing the QoE assessment logic that contains the mobility prediction and rate estimation algorithms, as well as the corresponding SDN programmability functionalities for providing feedback to VSPs regarding the HAS encoding rate and segment distribution. For the purposes of the simulations, the complete end-to-end HAS logic (i.e., video file encodings at different rates, streaming logic, user HAS strategies, user buffers with a maximum buffer size and a minimum playout threshold, etc.) is adopted considering also caching logic within eNBs that represent MEC platforms, while the user distribution and mobility are implemented using the SLAW model. In parallel, QoE-related measurements and KPI estimations are implemented, as well as the use cases presented next. For ensuring fairness, Proportional Fair scheduling is used. The simulation specific parameters regarding SLAW, network, and application parameters are summarized in Table 19.

Table 19: Basic simulation parameters for QoE-SDN APP use cases.

Parameter	Value
SLAW parameters (their meaning is explained in [137])	
Number of waypoints	1000
Hurst parameter	0.75
Alpha, Beta	3, 1
Pause time	0 sec
Clustering range	50 m
Trace generation time	Set to simulation time of 1 min
Maximum area size	Set to simulation area

User speed	1.38 m/sec
Network parameters	
Bandwidth available	20 MHz
Radio scheduler	Proportional fair
Network geometry	1 cell with 3 sectors
Inter-eNodeB distance	500 m
Number of mobile users	24 users
Initial user positions	SLAW-based
Prediction interval	4 sec
Traffic distribution	FTP: 10%, HTTP: 10%, VoIP: 10%, Video streaming: 70%
Application parameters	
Max buffer size	64 sec
Min buffer playout threshold	2.5 sec
Segment duration	2 sec
Available video bit rates (representations)	235, 375, 560, 750, 1050, 1750, 2350, 3000, 3850, 4300 kbps
First segment selection	At lowest quality layer
QoE model	$MOS = 3.5 \cdot e^{-(0.15 \cdot L + 0.19) \cdot N} + 1.5$
Video utility model	$VQ_{720p} = -4.85 \cdot {Br_V}^{-0.647} + 1.011$
Simulation parameters	
Number of SLAW topologies tested per use case	4 randomly created SLAW topologies

We concentrate our evaluation on HAS, considering both user and network KPIs. The former include QoE-related metrics that the user perceives, while the latter focus on overall network performance metrics.

User perspective: For the users' experience, we use QoE insights extracted via subjective experiments, which have led to the identification of the following main KPIs affecting the video delivery quality [70]:

- **Stalling events**, as elaborated in previous sections, refer to the interruption of video playback that occurs when the playout buffer runs out, and they are the most significant QoE degradation factor. According to the IQX hypothesis, and for the case of YouTube, the relationship between QoE and QoS is:

$$MOS = 3.5 \cdot e^{-(0.15 \cdot L + 0.19) \cdot N} + 1.5 \tag{9-7}$$

  where N and L are the number and duration of stalling events, as discussed thoroughly in Section 5.2.3.2.

- **Video characteristics** that shape QoE concentrate on the resolution and video bit rate, i.e., a higher resolution and video bit rate result in more satisfied users. A video utility model can be used to represent the video quality, using as input the video resolution and mean bit rate [140]. For the cases of 720p videos, the video utility function is as follows:

$$VQ_{720p} = -4.85 \cdot Br_V{}^{-0.647} + 1.011 \tag{9-8}$$

where Br_V is the video bit rate experienced by the user. Video utility takes values between 0 and 1, where 1 represents the highest quality. Moreover, the percentage of time at each quality layer that the user spent while watching a video is another meaningful KPI, strongly correlated to the resulting video bit rate. Nevertheless, the impact of unexpected stallings is much more severe than a controlled bandwidth reduction on the video bit rate [141]; therefore, stallings are the main QoE performance KPI we judge in the following evaluation subsection.

Network perspective: The average system throughput is a generic quality indicator typically not sufficient to accurately capture the video streaming experience from a network perspective. For instance, considering the following two extreme cases where: a) all users are served with a medium-quality layer, versus b) half users are served with a high-quality layer and half with a low-quality layer, both cases lead to the same average experienced throughput. However, the QoE among users significantly differs. Hence, a useful complementary KPI is fairness in the achieved QoE values (i.e., MOS), which can be estimated using Jain's index as follows, when there are U users in the system:

$$QoE\ fairness = \frac{(\sum_{u=1}^{U} MOS_u)^2}{U \cdot \sum_{u=1}^{U} MOS_u{}^2} \tag{9-9}$$

Another interesting KPI from the network perspective is the amount of network resources (i.e., bandwidth) consumed to achieve a mean QoE performance. If, for instance, specific techniques provide the same QoE level with others but at a lower bandwidth cost, these would be preferred from the MNO as more efficient.

9.5.2 Use cases enabled by the SDN QoE-APP

For our evaluation analysis, we adopted the following three use cases, considering first the HAS segment selection enforcement problem, then the segment encoding and placement (i.e., caching), and finally the proactive segment selection and placement. These use cases fall into the Bitrate Guidance category [117].

In line with the proposed architecture (Section 9.3.2), the communication flow that realizes the proposed use cases, once the QoE-SDN APP is setup by the VSP, is as follows: (1) The QoE assessment logic requests a periodic estimate of the data rates and positions of users of interest, i.e., VSP customers. An MNO can facilitate this requirement by the Data Plane Control Function via the D-CPI interface. (2) The MNO installs monitoring rules to any involved eNBs in order to collect and provide, in response, this information back to the QoE assessment logic (namely, eNBs serve as Network Elements). (3) The QoE assessment logic then predicts the data rate that each monitored HAS user is expected to achieve, based on per-cluster rate forecasting and mobility prediction (using Algorithms 1 and 2). (4) Finally, the QoE assessment logic enforces the segment selection of each user (use case 1), the segment encoding and placement (use case 2), or the proactive segment selection and placement (use case 3) and passes this information to the VSP side by the QoE control agent via the A-CPI interface. In more detail:

1. *Use case 1: Segment selection enforcement*

This use case demonstrates the potential of assisting users in their HAS segment selection decisions through the QoE-SDN APP. The information exposed by the MNOs to the VSPs is meant to help users take better decisions reflecting how the user perceived rate is expected to evolve. Such a procedure can be useful in cases of unexpected or rapid congestion, i.e., when the conventional segment selection decisions might prove detrimental and lead to stalling events. The QoE assessment logic collects the desired KPIs periodically and forecasts the expected rate based on the estimated per-cluster rate and mobility prediction (using Algorithms 1 and 2). Such estimated data rate is then used to guide the VSPs either by directly replacing the segment selection of particular users if required, or by indirectly limiting their available selection options (in the case video streaming is about to begin and the manifest file is prepared). Therefore, the suggested segment selection enforcement that takes place serially per user overrides the user's selection and delivers a safer segment alternative. Hence, the goal of this scheme is: a) to reduce stallings by proactively decreasing the quality layer that a user has individually selected based on his current perception of the network, if rate was overestimated, or b) to maximize the quality layer selection if rate was underestimated.

Referring to the basic optimization problem of Section 9.4.1, the first use case provides a real-time solution to the selection of x_{musl} segments. This estimation is based on prediction-based values about the data rates R_u per UE, which are made periodically available per prediction interval.

2. Use case 2: Segment encoding and placement

This use case considers the network-aware encoding and potential distribution of segments to MECs based on expected network conditions within each eNB coverage area. HAS traditionally requires the encoding of the video content at multiple bitrates (quality layers), which are pre-defined. Since the content is encoded in a network-agnostic way, it does not flexibly represent the current network conditions and load, nor does it allow for differentiation among different cells with different conditions (or even for differentiation in the same cell with timely variable congestion profiles).

Especially in cases of live video streaming, this implies a large waste of backhaul resources: During live video, the video segments are periodically encoded at the pre-defined available quality layers after they have been recorded, and then delivered to the MECs close to the users. Caching video streams at all available quality layers results in unnecessary backhaul resource waste, since some layers may never be requested due to the specific network profile of the area each cache serves (e.g., high bit rate representations will not be requested by users in a congested area).

Based on this observation, the novel opportunity arises to propose the flexible encoding of video segments that better reflect the network resources that vary in time and place. An example is the encoding and placement of very low video quality layers to high congested cells, and of higher quality layers to cells with light traffic. The benefits of such a scheme are twofold: a) backhaul resources will be saved, as only appropriate video representations will be periodically cached, and b) by limiting the available representations based on network congestion prediction, users will be indirectly led to take HAS decisions closer to reality, potentially reducing stalling events and increasing QoE.

To enable such segment encoding and placement, the QoE assessment logic should estimate the user expected rate at particular locations using Algorithms 1 and 2, and communicate it to the corresponding VSP, so that eventually the VSP will self-configure the content encoding based on this forecasting. Such segment encoding and placement decisions will be valid for a next interval, and then the entire process will be repeated. Therefore, in cases of live video streaming, such a procedure could lead to significant backhaul bandwidth savings, as the non-placed, redundant quality layers are actually savings for the backhaul MNO resource usage.

Referring to the basic optimization problem of Section 9.4.1, the second use case selects the x_{musl} from a subset of available quality layers, therefore $l = L_{m1}..L_{mL}$, where $L_{m1}..L_{mL}$ are the discrete quality layers cached in MEC m. This subset of available quality layers is determined by the R_u of all users.

3. Use case 3: Proactive segment selection and placement

In contrary to use case 2 that periodically performs a massive caching of video segments for all users, the third use case proactively enforces the caching of pre-recorded (i.e., offline) video segments in advance destined for a user, i.e., before the user requests a segment. The rational is to proactively cache appropriate segment encodings (based on rate estimation) in appropriate edge cloud platforms or MEC locations. This is done considering user's mobility prediction, thus avoiding the backhaul delay, that would be imposed when delivering a segment upon request instead of proactively bringing it close to the edge, while regulating congestion on backhaul links. In other words, the logic of this scheme is to proactively surpass the backhaul delay that is inevitably imposed when transferring a segment on demand from its original location to the network edge, thus makes users less prone to stallings. Such proactive segment placement relies on the QoE assessment logic that provides the user expected rate at particular locations using Algorithms 1 and 2. The appropriate segment is placed on the MEC server closer to the user, considering the user mobility prediction with respect to a predefined prediction window, and will be offered to the user replacing the original segment selection that may lead to stalling events.

Referring to the basic optimization problem of Section 9.4.1, the third use case makes the user segment selections x_{musl} and caches this content at appropriate MEC locations, diminishing backhaul delays related to the segment request.

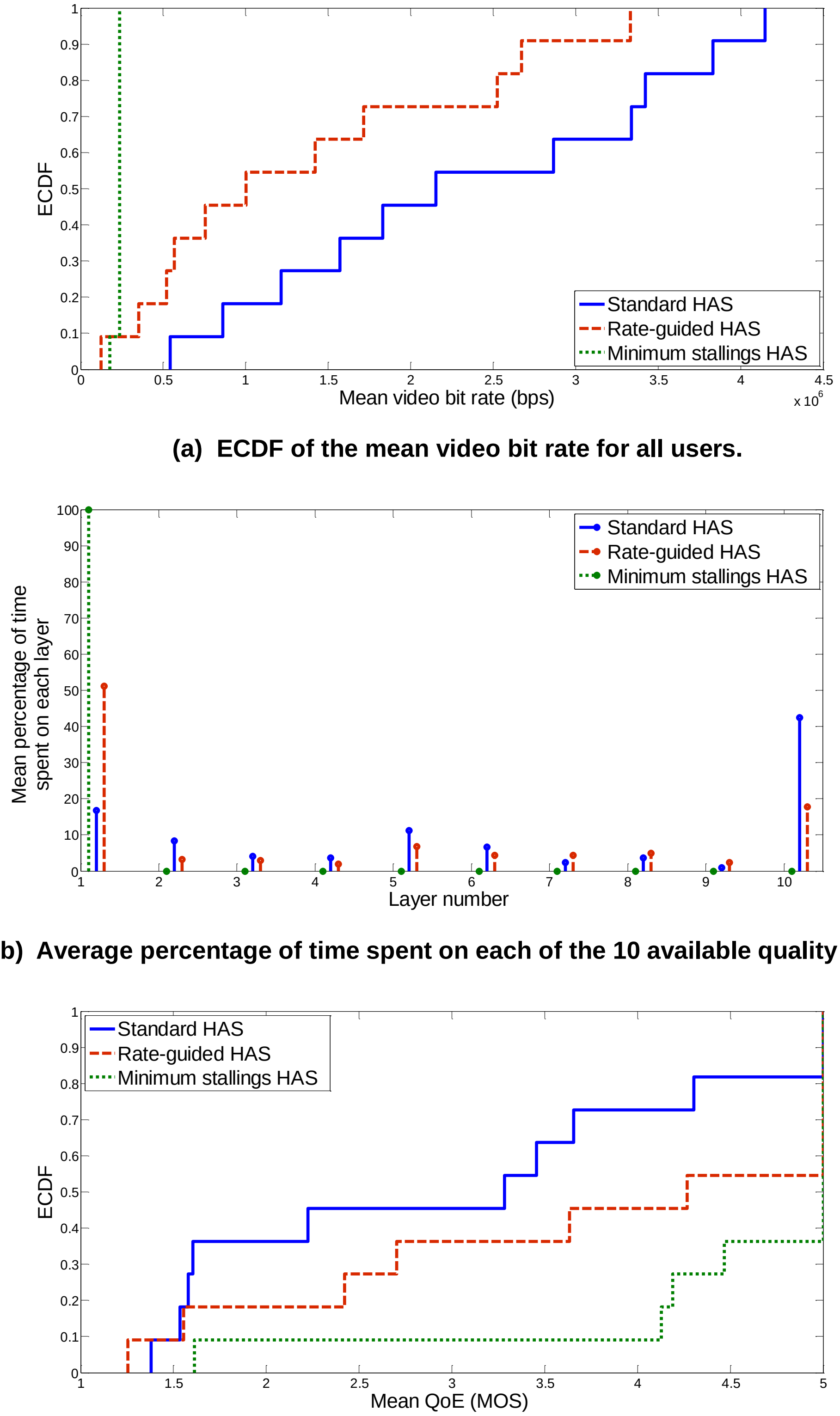
9.5.3 Evaluation results

Simulations were conducted comparing the aforementioned use cases with a standard, i.e., state of the art, version of HAS and with a conservative HAS variation that introduces minimum stalling events. The evaluation process was performed for each use case separately considering measurements in terms of various meaningful KPIs, i.e., mean video bit rate, mean quality layer downloaded, mean QoE, QoE fairness, mean video utility, mean stalling probability, mean stalling duration, and average stallings per user. For the second use case, we also measure the average number of active layers and the bandwidth savings estimate.

1. Evaluation analysis of use case 1

Segment selection enforcement considers three different HAS variations: a) the *standard HAS*, where always 10 representations are available per segment (this is the baseline strategy), b) the *rate-guided HAS*, where the segment selection of each user is guided by the QoE-SDN APP providing feedback to the VSP based on mobility- and cluster-based rate estimations, and finally c) the *minimum stallings HAS*, where only the lowest bit rate is requested (here 235 kbps per segment), leading to the least number of stalling events at the cost of very low video bit rate. The latter case represents a benchmark in terms of stalling events taking into account the specifics of the simulation environment.

The evaluation results are illustrated in Figure 57 and in Table 20. Figure 57 presents: a) the ECDF for mean video bit rate in the system, b) the average time spent viewing the video on each quality layer, and c) the ECDF for mean user QoE in the MOS scale, while Table 20 includes the mean values for various significant KPIs, such as mean MOS, stalling probability, video utility, fairness, etc.



(a) ECDF of the mean video bit rate for all users.

(b) Average percentage of time spent on each of the 10 available quality layers.

(c) ECDF of MOS for all users.

Figure 57: QoE-SDN APP - use case 1 evaluation results.

As shown in Figure 57a, the experienced mean video bit rate per user in the system is higher for the standard case, followed by the rate-guided HAS (with the QoE-SDN APP) and the minimum stallings HAS. This is due to the fact that the standard HAS case allows users to select segments with a higher quality layer in contrast with the proposed rate-guided HAS, which takes a more conservative approach, guiding users to select segments with a lower quality, as shown in Figure 57b. However, the proposed rate-guided HAS as well as the minimum stallings HAS, allow more segments (i.e., more playtime) to be buffered, preparing the video player better for imminent congestion and worse channel conditions. Therefore, such higher quality layer selection for standard HAS, is the result of overestimated subjective bandwidth calculations that mislead users to request segments with a higher quality layer, and thus, eventually experience stalling events. This effect is illustrated in Figure 57c, where the QoE model of Eq. (9-7) gives an estimation of the MOS as a function of the number and duration of stalling events, showing the benefits in terms of QoE for the proposed rate-guided HAS. Since stalling is the most important QoE shaping factor, such an improvement is highly desirable for the users (and therefore, the VSPs). Finally, even though the minimum stallings HAS leads to the lowest stalling rate (and thus, phenomenally higher QoE), it is not an acceptable solution, since it completely ignores the adaptation logic of HAS providing no video utility improvements even in low congestion scenarios.

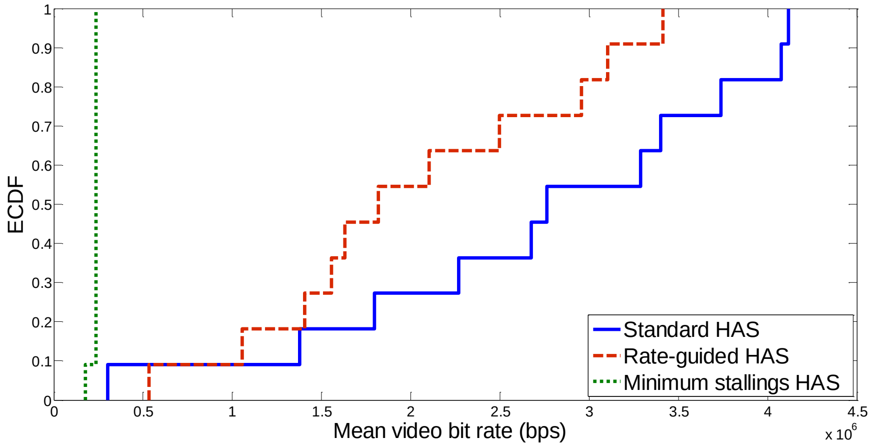
Table 20: KPI estimations - QoE-SDN APP use case 1.

HAS logic	Mean video bit rate (bps)	Mean quality layer downloaded	Mean QoE (MOS)	QoE fairness	Mean video utility	Mean stalling probability	Mean stalling duration (sec)	Average stallings per user
Standard	2.47E+06	7.65	2.98	0.78	0.95	0.63	11.52	1.52
Rate-guided	1.48E+06	4.10	3.64	0.83	0.91	0.41	16.33	0.61
Min stallings	2.31E+05	1	4.46	0.93	0.85	0.15	13.76	0.21

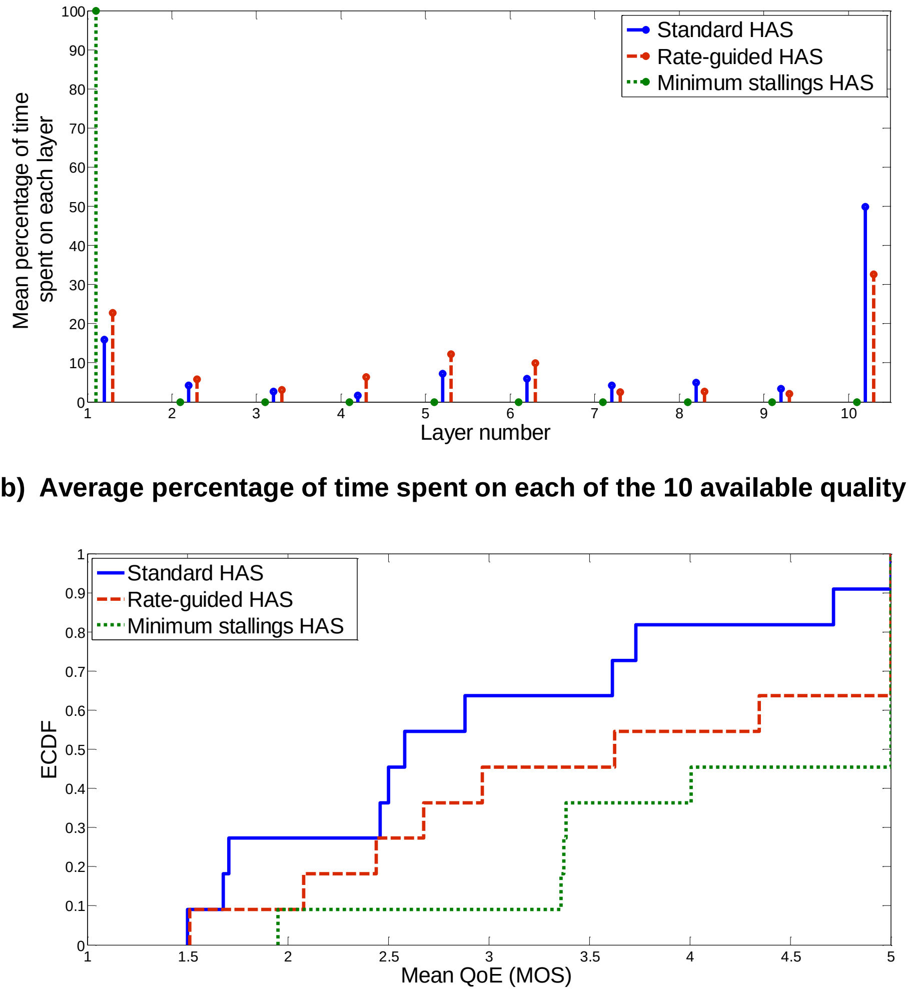
2. *Evaluation analysis of use case 2*

Segment encoding and placement demonstrates high benefits in terms of QoE, preserving a high video bit rate, while it can save backhaul capacity. As before, three HAS variations are considered: a) the *standard HAS*, where all 10 quality layers are encoded and cached, b) the *rate-guided HAS*, where the cached amount and video bit rate of the quality layers are driven by per-cluster rate estimation, and c) the *minimum stallings HAS*. Figure 58 and Table 21 illustrate the compared evaluation results.

Similarly to use case 1, the standard HAS provides the highest bit rate, since segments with higher quality layers are selected (Figure 58a and Figure 58b) at the cost of QoE, since MOS is tightly connected to stalling events (Figure 58c). For the same reasons, the minimum stallings HAS assures a better MOS since it always selects segments with the lowest quality layer, which however impacts significantly the user-experienced video bit rates and is not a viable adaptive video streaming logic.



(a) ECDF of the mean video bit rate for all users.



(b) Average percentage of time spent on each of the 10 available quality layers.

(c) ECDF of MOS for all users.

Figure 58: QoE-SDN APP - use case 2 evaluation results.

It is also observed, that the proposed rate-guided HAS not only provides a fair trade-off between video bit rates and MOS, but it can also result in significant backhaul capacity savings. Specifically, as presented in Table 21 (columns "Average number of active layers" and "Bandwidth savings"), on average only 2.24 quality layers instead of all 10 quality layers need to be cached, leading to significant bandwidth savings. What is more, such bandwidth savings are combined with higher QoE scores for the rate-guided case, as the users are indirectly prevented from a plethora of stalling-prone segment selections.

Table 21: KPI estimations - QoE-SDN APP use case 2.

HAS logic	Mean video bit rate (bps)	Mean quality layer downloaded	Mean QoE (MOS)	QoE fairness	Mean video utility	Mean stalling probability	Mean stalling duration (sec)	Average stallings per user
Standard	2.71E+06	7.86	2.94	0.81	0.95	0.65	10.82	1.31
Rate-guided	2.01E+06	6.08	3.60	0.84	0.94	0.44	10.65	0.73
Min stallings	229660	1	4.18	0.90	0.84	0.24	8.06	0.43
HAS logic	**Average number of active layers**				**Bandwidth savings (bps)**			

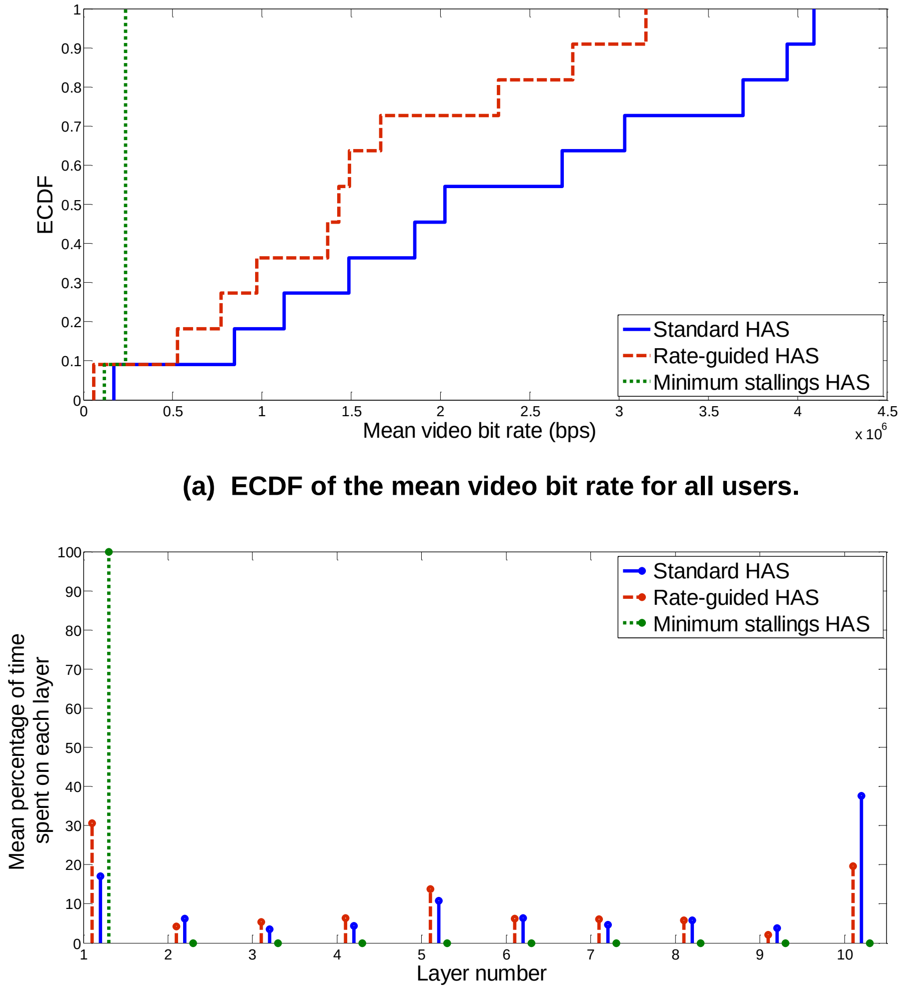
Standard	10	-
Rate-guided	2.24	1.43E+07
Min stallings	1	-

3. *Evaluation analysis of use case 3*

Proactive segment selection and placement studies the impact of proactive HAS segment caching. For the purposes of evaluation, we introduce a simplistic backhaul delay, which depends on the size of the transmitted segment, as:

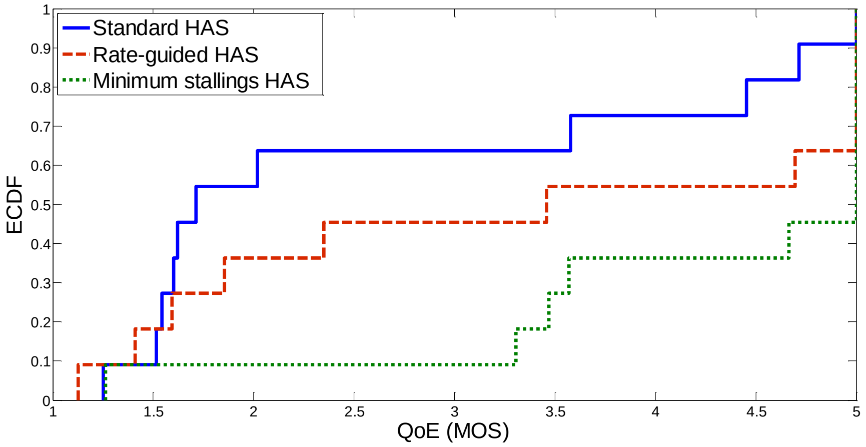
$$Backhaul\ delay = \frac{Segment\ size}{Backhaul\ rate} \tag{9-10}$$

in order to demonstrate the impact of the backhaul. The backhaul rate is set to 10Mbps (i.e., the achieved backhaul rate per user on average), so that the access network connectivity is not backhaul restricted (actually, the mean video bit rate is much less, as shown in Table 22). As before, three HAS variations are considered, namely: a) the *standard HAS*, where there is no proactive caching, b) the *rate-guided HAS*, where the rate estimation is used to enforce the VSP segment selection, with the mobility prediction guiding the *proactive caching* of these selected segments to the appropriate MEC locations, and c) the *minimum stallings HAS*, that caches the lowest segment quality layers only. The results obtained are showed in Figure 59 and Table 22.



(a) ECDF of the mean video bit rate for all users.

(b) Average percentage of time spent on each of the 10 available quality layers.



(c) ECDF of MOS for all users.

Figure 59: QoE-SDN APP - use case 3 evaluation results.

Similarly to the previous simulations, the proposed rate-guided HAS provides a better balance in the achieved video bit rate and MOS compared to the standard and the minimum stallings HAS strategies. It is also observed that stalling events are less likely to occur when proactive caching is used. The reason for that, additionally to the benefits of the rate-guided segment selection process, is that this scheme reduces the backhaul delay that is required to fetch a video segment upon request; therefore, the user has more chances of downloading this segment early enough, i.e., before the segment's deadline.

Table 22: KPI estimations - QoE-SDN APP use case 3.

HAS logic	Mean video bit rate (bps)	Mean quality layer downloaded	Mean QoE (MOS)	QoE fairness	Mean video utility	Mean stalling probability	Mean stalling duration (sec)	Average stallings per user
Standard	2.27E+06	7.01	2.63	0.76	0.92	0.71	10.77	1.69
Rate-guided	1.50E+06	4.96	3.31	0.80	0.87	0.49	11.95	0.84
Min stallings	2.24E+05	1	4.20	0.90	0.82	0.20	16.82	0.22

9.6 Conclusions

In this chapter, we have introduced a programmable QoE-SDN APP, based on the openness and flexibility provided by the SDN paradigm. This QoE-SDN APP can serve the customers of VSPs, improving their QoE by reducing the occurrence of the highly undesirable stalling events. Focusing on HAS applications, and by running a mobility forecasting and rate estimation function within the MNO's domain, the proposed scheme manages to significantly improve the QoE of video streaming users. This improvement has been highlighted and quantified through the proposal and evaluation of use cases for video segment encoding, selection and placement that are "unlocked" by the proposed architecture. These techniques take advantage of network feedback information exposed by the MNO related to the positions and data rates of mobile users, in order to trade off stalling events with video bit rates, since the former have a much stronger QoE impact. Based on the simulations conducted, the rate-guided HAS strategies enforced by the QoE-SDN APP also ensure fairness among users, in parallel to improving QoE.

Apart from the technical novelty of the proposed scheme, added business value is expected. Specifically, the introduction of the QoE-SDN APP has an impact not only on the reputation of various service providers, but also on the revenues of the MNOs, stemming from bandwidth savings and from direct financial benefits through API exposure to service providers. The activation of the QoE-SDN APP can be on-demand, rather than being an "always-on" function and can be programmed according to particular service needs. For instance, some VSPs already differentiate their customers, based on their subscription type, to gold or standard users; in this case, the QoE-SDN APP can be triggered only for the former type of users. Similarly, the QoE-SDN APP may be designed as an add-on feature, which customers can activate on-demand, and for a limited amount of time, i.e., in the form of time-bounded purchased tokens or pay-as-you-go schemes. When any of these schemes is recognized, then the QoE-SDN APP and the accompanying QoE management cycle will automatically instantiate the essential monitoring and control actions within the MNO that will boost the customer QoE.

The need to improve the end-users' experience together with the emergence of technologies such as SDN, MEC and personalized network slicing, which enable such improvements through service/application and user/OTT differentiation, pose a challenge to net neutrality principles. The QoE-SDN APP offers a differentiated and enhanced experience to the users of VSPs that choose to adopt it, in a broad sense. However, it raises none net neutrality concerns, since in the context of the HAS use cases, the QoE-SDN APP does not require any special traffic treatment to different traffic flows by the MNO, such as prioritization against other traffic classes; it just enables QoE assessment and network exposure feedback mechanism to VSPs that helps them better handle video streaming.

Future work involves the real implementation of the proposed QoE-SDN APP on an SDN testbed, to showcase the applicability of this scheme for real HAS services and devices. Moreover, even though this study has concentrated on HAS, the benefits for other types of services and verticals remain to be investigated.

10. QoE TOWARDS 5G

5G is rapidly moving from vision to reality and there is already some consensus regarding the technical requirements of 5G. According to the NGMN alliance, "*5G is an end-to-end ecosystem to enable a fully mobile and connected society. It empowers value creation towards customers and partners, through existing and emerging use cases, delivered with consistent experience, and enabled by sustainable business models*" [84]. This definition incorporates the importance that 5G gives on the user experience (i.e., "consistent experience"), referring to the importance of seamless service delivery. Service delivery in 5G should also account for the sheer diversity of existing and emerging use cases as well as for the vast variety of demands per service type. These requirements would only be addressed by a shift from system-centric to user-centric architectures. Given this background, Section 10.1 identifies key QoE requirements that need to be integrated in the 5G ecosystem and highlights the importance of network designs that have the user at their epicenter. Then, Section 10.2 envisions the emergence of "experience packages" towards a 5G ecosystem that is flexible and dynamic in terms of user experience.

10.1 QoE requirements in the 5G ecosystem

In this section, we sketch how the user experience should look like in the 5G ecosystem by describing its desired QoE requirements. The main objective is to draw and emphasize the necessity of these requirements as the only way to provide an excellent and solid user experience, as expected by the next generation of cellular networks. It is crucial, that these attributes are identified early enough, so as to push towards the design of more user-centric networks, which will enable these requirements using current or emerging technologies. With this objective, we identify that the user experience in 5G ecosystems should have the next characteristics:

Consistency

The requirement for "consistency" has been clearly identified in the vision of NGMN. It refers to the uninterrupted, seamless and invariable (or with as low variance as possible), but still excellent quality of the offered service. Consistency spans across many dimensions such as time, space, infrastructure, operator/vendor, end-device and application. Therefore, a 5G user should expect to receive a continuous service, without disruptions, and with limited fluctuations. Some of the main obstacles in achieving this requirement in a mobile environment are the uncontrollable and unstable channel conditions, the competition over the scarce spectrum resources and the high heterogeneity and density of these networks, causing constant handovers and unpredictable interference levels. To overcome these challenges, traditional network management decisions have to be revisited and transformed to smarter, QoE-aware mechanisms, as the ones proposed in the current thesis. Such mechanisms will then be able to account for the impact of various QoS-based parameters on the user experienced quality, and drive network operations accordingly. Note, that consistency has been the motivation for the work conducted in Chapter 7.

Transparency

Transparency refers to the requirement of the network to "hide" its complexity and efforts on delivering excellent and seamless quality to its customers. Best experience should always follow the user, while he/she is spectrum and system agnostic. This means, that although the user is considered to be the epicenter of a 5G network, his/her implicit input or intervention in any network or service management decisions should be avoided. For instance, even though providing a solid experience is a clearly subjective issue, the user is not expected to be actively involved in QoE measuring and monitoring procedures; this should be done automatically by the network either by exploiting passive feedback from the user's application or device or by using network probes and DPI techniques. This does not mean, however, that interactivity between the network and the user should be avoided; on the contrary, such interactions should be present but limited to value-adding occasions, i.e., when it is meaningful and somehow expected by the user. For instance, the operator may ask for explicit user feedback regarding the experienced quality, similar to how Skype does after a user has completed a video call.

Resource and energy efficient QoE-awareness

Adding QoE awareness and, in turn, implementing QoE-aware service and network management, will inevitably introduce more complexity in the network. For instance, periodic QoE probing and monitoring will have to be implemented in both edge and core network nodes, increasing their battery consumption. Similarly, extra control signaling overhead will be imposed in the access network, which may cause a resource-insufficiency problem. Since enabling QoE in the network is translated to such resource and energy costs, we need to make sure that the energy and resource costs per "QoE unit" are maintained to a reasonable, minimum level. Opportunities to control these ratios may stem from the science of human perception (Psychophysics), amongst others.

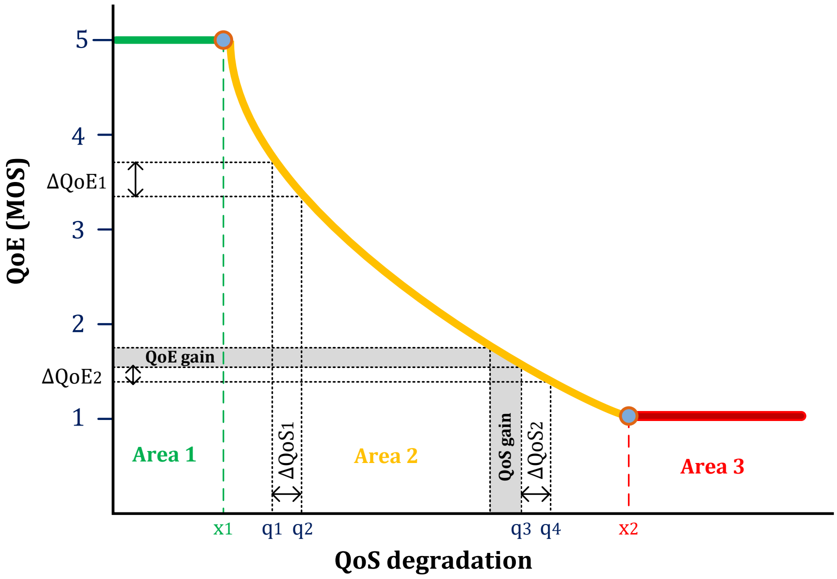


Figure 60: QoE-driven resource scheduling sketch.

For instance, a simple idea towards a more efficient resource scheduling in terms of QoE is given in Figure 60. This idea exploits the "area 2" of sinking QoE of the IQX hypothesis [20]. Assume that $User1$ has been scheduled in a way that he/she achieves a QoS value equal to q_1 (e.g., data rate). In parallel, another $User2$ achieves q_3, where q_1 is better than q_3 (e.g., $User1$ has lower packet loss ratio or higher throughput), because, for instance, he enjoys better channel conditions. At a later timeframe, and assuming that the network has become more congested, the radio scheduler has to reduce the allocated resources, say by ΔQoS. The question is how to perform that in the most efficient way. Widely used schedulers such as the Proportional Fair scheduler, account only for QoS metrics, i.e., how to maximize the total system throughput considering also some fairness among competing requests. Therefore, these schedulers see no difference between a) shifting $User1$ from q_1→q_2 (equals to a ΔQoS_1 degradation) or b) shifting $User2$ from q_3→q_4 (equals to a $\Delta QoS_2 = \Delta QoS_1$ degradation), since the resulting QoS degradations in both cases are equal. In the QoE domain, though, this is no longer a valid assumption. As observed in Figure 60, it is much preferable to shift $User2$ and leave $User1$ unaffected, achieving in this way an overall system QoE gain (QoE gain -

shaded area) which directly translates to more satisfied users on average. This problem may be also solved in reverse, namely, for a presumed fixed ΔQoE decrease, how to save the most resources in the QoS domain (QoS gain - shaded area).

User personalization and service differentiation

Services provided over a 5G network should be tailored to specific users or user profiles. The key to achieving this is through enabling QoE personalization inside the network. Netflix already distinguishes among gold/silver/bronze users, based on their subscription profiles, and configures the offered quality accordingly. However, explicitly paying for a subscription profile and, thus, receiving correlate quality is just one possibility of enabling QoE personalization. In fact, users may be equally (or more) interested in other aspects such as low battery consumption, increased privacy, or low charges, and in general, they may have different willingness-to-pay profiles. This claim is enforced by recent subjective studies [142], where it has been shown that some users are willing to compromise the quality they receive in order to spend less money, while others are willing to pay more and more to receive "virtually" better quality (a "placebo effect"). Therefore, we may envision 5G as a system that tries to comprehend its subscribers' expectations and differentiate the offered services accordingly. This will be further argued in the next subsection.

Except for differentiation on a per-user basis, a differentiation per service and application type is expected. QoE is tightly dependent on the type of application, and different QoE requirements are needed for different applications. Therefore, it is required that the 5G systems are flexible when serving diverse applications, tailoring their quality monitoring and provisioning techniques a) according to the different influence factors per application, b) according to the different impact and tolerability that the same QoS parameters have on different applications, and c) according to the applications' adaptability to varying network conditions.

To achieve this degree of personalization, the provisioned QoE in 5G networks should account for the context of each communication session, which, as discussed in Section 8.1.2, is a very challenging task. A result of such personalization might be that, for instance, high demanding users (e.g., business users) are prioritized over users who would not perceive or care about some extra delays during their communication sessions. Reaching such context-awareness may enable a more meaningful network and service management and, thus, become a powerful tool of 5G networks.

10.2 The concept of "experience packages" in the 5G era

Handling QoE as a MOS value is not sufficient to meet the requirements of 5G communication networks. In this direction, the concept of "experience package" emerges. Experience packages may be configured and delivered in a way that fine-grained differentiation is achieved, respecting the user, application and communication context.

QoE should be provisioned in a way that users create the impression that the best experience "always follows them", regardless a) the application they use, b) the communication context in which they are currently involved, and c) the current network conditions. What is more, users are not uniform in their QoE expectations or requirements. Their satisfaction with a service is the result of their psychology, cognitive and psychophysical characteristics, and current state. Nevertheless, so far, mobile cellular networks do not allow a fine-grained differentiation in the offered experience, since:

- All users are represented by an "average user".

- Most applications are treated as best effort.
- The context of communication (e.g., task urgency, environment, billing, etc.) are not actually taken into consideration in service provisioning.

Nevertheless, a rational development in the current communication paradigm as well as a key for reaching 5G requirements is the support of user personalization, application differentiation and quality adjustment based on the current communication scenario, as also discussed before. User personalization is currently performed just by distinguishing subscription profiles and configuring the offered quality accordingly. However, telecom operators do not really engage in such a per-user differentiation on a monetary basis, at least for the time being.

However, if we move one step further, we can consider QoE not only as a single MOS value, but more generally as an "experience package". For instance, we could easily claim that a user is not only interested in receiving the best quality, but may be equally (or more) interested in communicating in an energy-efficient way through his/her device, in minimizing the charges imposed when using a service, in being prioritized for a specific task with respect to another, in being served in the securest way possible, or in combinations of the previous. Similarly, other dimensions can be integrated into an experience package.

Capitalizing on this observation, we envision future architectures, where the network builds "experience packages" based on actual communication scenarios and user profiles. We assume that such profiles are built based on the users' communication habits, preferences, mobility patterns, physical environment, used equipment, etc., while a group of users will fit a specific profile. These profiles can be built offline, based on accumulated information about the users, while they will be updated in case of different or unusual user behavior. Once such a "pool of profiles" is deduced and becomes available at the operator's side, then configuring (i.e., tuning) the offered "experience package" during any communication session is a 4-step process:

- **Step 1:** Match a user to a suitable profile based on demographics, preferences, subscription types or other factors that are meaningful and measurable.
- **Step 2:** Derive the application unique characteristics and requirements, such as tolerability and adaptability to various network conditions.
- **Step 3:** Deduce the context of the current communication session (based on insights from the past, sensed environment, device info, etc.).
- **Step 4:** Finally, build an "experience package" in real-time upon service delivery, which will remain valid throughout the user's session. The package building decision is inevitably a compromise between the requirements derived from steps 1-3 (i.e., demand) and the actual capabilities of the network in terms of momentary available resources (i.e., supply).

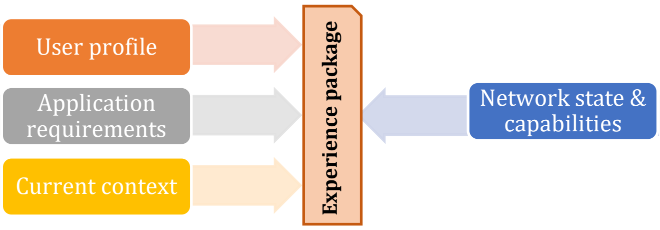


Figure 61: The process of building an experience package.

Therefore, as depicted in Figure 61, a network's decision about how to build an experience package is a function of a) the user profile, b) the specific application requirements, c) the current context, and on the other side, d) the current network state and capabilities. The latter includes any restrictions or limitations in the network, such as resource availability, current load, energy constraints, operator policies, etc. The experience packages may be built at a central node within the operator's core network, while the 3GPP PCRF can be considered as a potential host of this procedure, as it is already responsible for the creation of rules and policies for each subscriber [143]. Similarly, the Policy and Charging Enforcement Function (PCEF) and the Access Network Discovery and Selection Function (ANDSF) can be considered as candidates for the enforcement of these packages to the users.

An experience package does not refer to entities sent to the users. It is a reference of how service provisioning should be adjusted per user's session in order to provide special treatment. In that sense, it reminds of the QoS-centric "bearer" concept of LTE [144], while it embraces much broader QoE-centric intelligence with respect to the user, application and context. As a result, a user will perceive a much more personalized and friendly experience, tailored to his/her specific session needs.

As an example, we consider experience packages as a weighted sum of the following three dimensions: $\{QoE, price, energy\ cost\}$, while this sum is subject to the actual network capabilities, i.e., $\sum_{i=1}^{3} weight_i * dimension_i \leq network\_specific\_value$.

In Figure 62, we abstractly present five experience packages that derive from matching these three weighted dimensions to the actual network capabilities (note that, for simplicity, context and application awareness are ignored). We can observe various possible configurations: Delivering experience package "1" puts more emphasis (weight) on QoE, so it represents a situation where a user is mostly interested in quality, regardless of the price- or energy-to-pay. Similarly, packages "2" and "3" represent users who care more about charges and energy costs, respectively, at the expense of quality. Finally, packages "4" and "5" imply some trade-offs between all dimensions, subject to the current network conditions. In real-life, package "2" could lead to a reduction of the video quality layers in a HAS session in order to reduce data consumption and subsequent charges, while package "3" could handover a user's uplink to the closest access-point (e.g., femto- instead of macro-base station) to reduce the device's transmitted power.

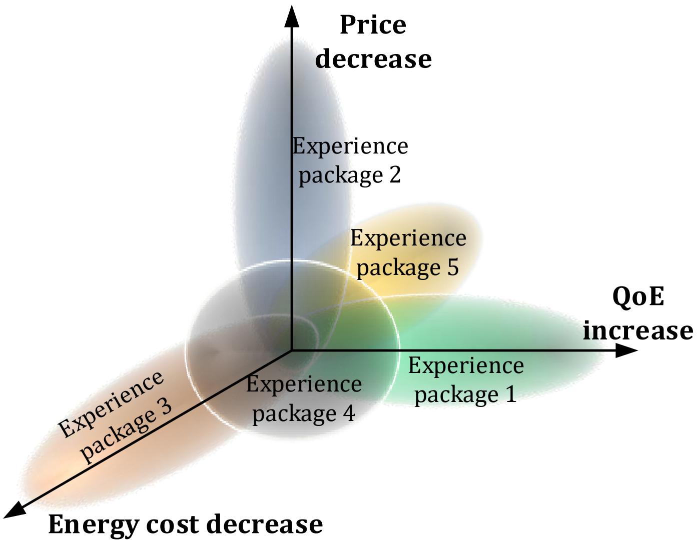


Figure 62: Possible configurations of experience packages.

To enable the creation and empowerment of experience packages in a network, various enablers can be called forth. First, a QoE management entity is required that implements the functions of QoE collection, monitor and management (as described in Chapter 3), that help evaluate and control the users' QoE. In terms of QoE evaluation, traditionally used models such as the E-model need to be confirmed or revisited (e.g., regarding the model's default values and permitted ranges), while new QoE estimation models for the emerging services towards the 5G era, such as immersive and 360° video, need to be created.

Apart from this, network slicing appears to play a key role [145]. The concept of slicing has been proposed towards 5G as a way to handle the plethora of verticals that are integrated into this new ecosystem. Slicing per vertical can be seen, however, as a coarse-grained solution. We can however envision a more fine-grained slicing concept, that is slicing per user, application and context. This would imply that two neighbor users might be served by different softwarized-eNBs or softwarized-EPCs, on the basis of successfully delivering each one's experience packages. These decisions will be driven by the creation of a fine-grained slice per user flow.

Resource allocation mechanisms also need to evolve in order to serve the experience packages' provisioning. Thinking "out of the box", we may envision "elastic resources" as a superset of spectrum, processing power, memory, energy, etc., namely of any dimensions that build an experience package. Then, through the virtualization of elastic resources in combination with an abstraction of the wireless medium, flexible experience package provisioning could be enabled. However, technical feasibility and challenges need to be addressed first.

11. CONCLUSIONS AND FUTURE WORK

The introduction of QoE intelligence and QoE-aware capabilities in mobile cellular networks (e.g., LTE/LTE-A) changes the network management approach. Future network management implements a QoE management cycle, where a) QoE-related intelligence is gathered, b) QoE modeling and monitoring reveals the user satisfaction level or warns about imminent problems, and finally, c) a QoE control procedure triggers proactive or reactive actions to appropriate network elements and functions.

This thesis has dealt with the challenges arising from the need to integrate QoE intelligence in a mobile cellular network, which mainly concern the real-time evaluation of QoE, the improvement of existing network mechanisms, and the proposal of new QoE-inspired algorithms, stemming from the inherent characteristics of QoE and the non-linear impact of conventional QoS parameters on the user perception. Under this perspective, the main achievements and results of this thesis, accompanied by potential future research directions, are the following:

- A conceptual framework, which enables QoE provisioning in mobile cellular networks. This framework is analyzed with respect to its design, its building blocks and their interactions, while practical challenges regarding its adoption by network operators are discussed. An evaluation of a simple QoE-based admission controller serves as proof of concept and proof of potential of such a scheme. Nevertheless, the exact integration of the proposed QoE provisioning cycle lies in the capabilities and interests of the stakeholders, who would need to adopt this framework and customize it according to specific requirements and needs.
    - An interesting future direction related to this study would be the integration of this QoE provisioning framework in upcoming 5G networks. Since the 5G architecture is currently undergoing its early implementation phase, such a QoE management cycle could become an integral part of this architecture, by specifying exact access and core network components involved, as well as functionalities and signaling required. For instance, the integration of QoE logic in the Policy Control Function (PCF) of the 5G system architecture could be investigated as an option, as according to [146] it supports unified policy framework to govern network behavior and provides policy rules to Control Plane function(s) to enforce them.
- The classification of QoE models and, mainly, the identification and evaluation of parametric QoE models and KPIs that are appealing for the purposes of real-time QoE evaluation of widely used services (i.e., VoIP, online video, video streaming, web browsing, Skype, IPTV and file download services). The input parameters of these models allow their collection from various network elements, making this study a handy tutorial towards practical QoE assessment in a network, and towards understanding the impact of network decisions on the user’s perception.
    - QoE modeling alone is a huge research area, which even non-networking experts are thoroughly exploring. Therefore, future research on QoE provisioning in communication networks needs to closely follow the advances in the QoE modeling area, keeping an eye not only on newly standardized models, but also on insights from new subjective experiments concerning quality estimation and KPIs for new resource-hungry immersive services, such as the 360° video, the 3D video and the Virtual Reality (VR) content.
- A network management scheme driven by QoE measurements in order to control the operational mode of mobile users in LTE-A networks with D2D support. The

signaling required to support this QoE-driven D2D management is proposed with respect to 3GPP standards. Simulation results for a specific network deployment demonstrate benefits for both operators and users, exhibiting an average user QoE improvement of up to 35%, and a parallel cell throughput increase of up to 18%. Such a QoE-driven scheme may become the enabler for introducing D2D into the market, by allowing operators to qualify for justified and acceptable user charges, when provisioning this new technology.

  - The current study has focused on VoIP communication between D2D users; however, in the more general context of proximity services, scenarios related to Vehicle-to-Vehicle (V2V) communications, Machine-to-Machine (M2M) communications, commercial D2D discovery, or even public safety could be considered. The main challenge in this direction is the proper definition of the QoE concept under these different scenarios (e.g., QoE may be related to the user comfort for the case of M2M applications) followed by the integration of QoE centricity in their operation.

- A new, QoE-inspired radio scheduling logic that accounts for the impact of network throughput fluctuations on QoE. This proposal stems from recent studies that characterize “consistency” as a key QoE influence factor, neglected beforehand, though. Evaluating and comparing conventional radio schedulers, we first reach to the conclusion that fairness inherently favors consistency. Moving one step further, we propose an inherently consistent scheduler, which further improves users’ QoE by moderating throughput fluctuations, and by achieving higher minimum MOS values compared to conventional schedulers. Overall, this study shows the necessity to re-consider existing network mechanisms with the objective of providing consistency, proposing in parallel this novel research direction for future works.

  - This study has exploited some early insights regarding the impact of consistency on user QoE for interactive applications. Future work involves following any potential new subjective experiments in this field that might reveal updated QoE models, as well as studying the impact of consistency when other applications or services are considered (e.g., immersive services). Finally, there is still room to design more sophisticated fluctuations-aware schedulers than the one proposed in this thesis, which optimize the decision-making process, considering in parallel real-time constraints.

- A proactive context-aware HAS strategy complementary to any standard HAS approach implementation, which, if activated on time, helps prevent stallings in light of bandwidth-challenging situations. For the purposes of this study, the video quality degradation problem is formulated analytically, followed by a thorough evaluation in terms of HAS-related KPIs both for the optimal case, and for a real-time context-aware implementation (e.g., per-layer percentage of video playout time, stalling occurrence, etc.). This study reveals the potential of using context-awareness and cross-layer information to serve conventional networking mechanisms, such as HAS, and the impact of such approaches on the user QoE.

  - Future work of interest concerns the study of parallel multiple mobile video HAS users, who are competing for the same pool of spectrum resources. This scenario includes, first of all, the study of how mobile users behave under these competing conditions (e.g., a greedy HAS behavior might be revealed), as well as the study of the network’s behavior in terms of stability (or consistency), fairness and QoE. Assuming that an unfair, unstable or greedy behavior is observed, novel centralized HAS management schemes could be proposed to improve this phenomenon, elaborating, however, on the trade-offs between such centralized or semi-centralized approaches and current fully distributed HAS strategies.

- A programmable QoE-SDN APP, and its accompanying SDN-based architecture that promote and enable a technologically feasible realization of a collaboration paradigm between service providers and mobile network operators. This QoE-SDN APP can serve video service customers, improving their QoE by reducing the occurrence of the highly undesirable stalling events. Focusing on HAS applications, the potential of this architecture is highlighted through the proposal and evaluation of three use cases unlocked by this architecture. In this paradigm, feedback about the network throughput is provided to the VSP in order to redefine encoding, caching, and per-user video segment selection in a network-gnostic, QoE-smarter way. This study, therefore, incentivizes a futuristic (but probably inevitable) networking paradigm, where service providers and network operators interact, for the mutual interest of both parties.
  - Future work regarding this study involves the real implementation of the proposed QoE-SDN APP on an SDN testbed, to demonstrate the proof of concept, applicability and measurable benefit of this scheme for real HAS services and devices. Moreover, even though this study has concentrated on HAS, the benefits for other types of services and verticals (other than mobile broadband) remain to be investigated. Finally, the design of the QoE-SDN APP needs to be constantly updated following the latest trends and specifications of SDN.
- Identification of the essential attributes that can shape QoE-centric networks towards the 5G era, and introduction of the “experience package” concept. Experience packages can lead to more personalized service provisioning to users, considering not only technical parameters, but also the user profile and the context of the communication.
  - The discussion about experience packages has been left to a conceptual level. Therefore, future work involves the investigation of practical realization requirements, constraints and benefits from this concept.

It is worth mentioning, that an aspect not discussed in this thesis, but lately identified as another key technology for QoE improvement of network services is Network Functions Virtualization (NFV) [147]. Whereas SDN refers to the decoupling of the control and data plane, allowing a network to be configured centrally in a software-based (i.e., flexible) way, NFV, on the other hand, enables the implementation of network functions as software, which can then run on generic hardware, and can be moved or instantiated in various locations in the network on demand. To achieve that, NFV configures the available network-, storage- and processing- resources based on policies from a central orchestration and management system. To leverage the advantages of this emerging technology, a QoE orchestrator may be envisioned for network/service management as an integral part of the ETSI Management and Orchestration (MANO) architecture, with the objective to introduce user-centricity in this module.

As a general comment, the research conducted in this thesis has focused on the integration of QoE to research topics that are currently under intense research interest from academia and industry, such as D2D, HAS, radio scheduling and SDN. However, this is just a subset of potential solutions that may be proposed, when QoE intelligence is integrated into the real-time operation of a future network. Nevertheless, this thesis provides valuable insights and useful findings in this direction, further encouraging research in the area of QoE characterization and provisioning in mobile cellular networks.

ABBREVIATIONS – ACRONYMS

3GPP	3rd Generation Partnership Project
A-CPI	Application-Controller Plane Interface
ACR	Absolute Category Rating
ACR-HR	Absolute Category Rating with Hidden Reference
ADT	Average Download Throughput
ANDSF	Access Network Discovery and Selection Function
API	Application Programming Interface
ARCU	Application, Resource, Context and User
ARIMA	Auto Regressive Integrated Moving Average
BET	Blind Equal Throughput
BLER	Block Error Ratio
BR	Bit Rate
CDF	Cumulative Distribution Function
CDN	Content Distribution Network
CEM	Customer Experience Management
CF	Correcting Factor
CQI	Channel Quality Indicator
CSI	Channel State Information
D2D	Device-to-Device
D-CPI	Data-Controller Plane Interface
DL	Downlink
DPI	Deep Packet Inspection
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double-Stimulus Impairment Scale
EADT	Effective Average Download Throughput
ELA	Experience Level Agreement
eNB	evolved Node B
EPC	Evolved Packet Core
ETSI	European Telecommunications Standards Institute
E-UTRAN	Evolved UMTS Terrestrial Radio Access Network
FDD	Frequency Division Duplex
FR	Full Reference
FR	Frame Rate

FTP	File Transfer Protocol
GBR	Guaranteed Bit Rate
GoP	Group of Pictures
GoS	Grade of Service
GPS	Global Positioning System
GSM	Global System for Mobile Communications
HAS	HTTP Adaptive Streaming
HeNB	Home evolved Node B
HTTP(S)	Hypertext Transfer Protocol (Secure)
IETF	Internet Engineering Task Force
IP	Internet Protocol
IPTV	Internet Protocol Television
IQX	Exponential Interdependency of Quality of Experience and Quality of Service
ISP	Internet Service Provider
ITU	International Telecommunication Union
ITU-D	ITU Telecommunication Development Sector
ITU-T	ITU Telecommunication Standardization Sector
KPI	Key Performance Indicator
KQI	Key Quality Indicator
LATP	Least-Action Trip Planning
LQO	Listening Quality Opinion
LTD	Low-Throughput Duration
LTE	Long Term Evolution
LTE-A	Long Term Evolution - Advanced
M2M	Machine-to-Machine
MANO	Management and Orchestration
MCP	Major Configuration Parameter
MCS	Modulation and Coding Scheme
MEC	Multi-access Edge Computing
mIoT	massive Internet of Things
MNO	Mobile Network Operator
MOS	Mean Opinion Score
MPEG	Moving Picture Experts Group
MPQM	Moving Pictures Quality Metric

MSE	Mean Square Error
MT	Maximum Throughput
NACK	Negative Acknowledgment
NE	Network Element
NEF	Network Exposure Function
NGMN	Next Generation Mobile Networks
NGN	Next Generation Network
NR	No Reference
OFDMA	Orthogonal Frequency-Division Multiple Access
OTT	Over-The-Top
P2P	Peer-to-Peer
PC	Pair Comparison
PCEF	Policy and Charging Enforcement Function
PCRF	Policy and Charging Rules Function
PDA	Personal Digital Assistant
PDCCH / PUCCH	Physical Downlink/Uplink Control Channels
PDN-GW	Packet Data Network Gateway
PDSCH / PUSCH	Physical Downlink/Uplink Shared Channels
PER	Packet Error Rate
PESQ	Perceptual Evaluation of Speech Quality
PF	Proportional Fair
PLC	Packet Loss Concealment
PMI	Precoding Matrix Indicator
POLQA	Perceptual Objective Listening Quality Analysis
PSNR	Peak Signal to Noise Ratio
PSQA	Pseudo-Subjective Quality Assessment
PSQM	Perceptual Speech Quality Measure
QFF	QoE Fairness Framework
QoE	Quality of Experience
QoR	Quality of Resilience
QoS	Quality of Service
QQVGA	Quarter Quarter VGA
QVGA	Quarter VGA

RAN	Radio Access Network
RB	Resource Block
RF	Resource Fair
RI	Rank Indicator
RIB	RAN Information Base
RNN	Random Neural Network
RR	Reduced Reference
RSS	Received Signal Strength
RTCP-XR	Real-Time Control Protocol-Extensive Report
RTP	Real-time Transport Protocol
SABR	SDN-based Adaptive Bit Rate
SBR	Send Bitrate
SDMN	Software-Defined Mobile Networks
SDN	Software-Defined Networking
S-GW	Serving Gateway
SINR	Signal to Interference plus Noise Ratio
SIR	Signal to Interference Ratio
SLA	Service Level Agreement
SLAW	Self Similar Least-Action Walk
SLTD	Selective Low-Throughput Duration
SNR	Signal to Noise Ratio
SSCQE	Single Stimulus Continuous Quality Evaluation
SSIM	Structural Similarity Index
SWOT	Strengths, Weaknesses Opportunities, Threats
TCP	Transmission Control Protocol
TELR	Talker Echo Loudness Rating
TJ	Throughput Jitter
TTI	Transmission Time Interval
UDP	User Datagram Protocol
UE	User Equipment
UL	Uplink
UMTS	Universal Mobile Telecommunications System
UPN	User Provided Networking
UX	User Experience

V2V	Vehicle-to-Vehicle
VAD	Voice Activity Detection
VGA	Video Graphics Array
VNF	Virtual Network Function
VoD	Video on Demand
VoIP	Voice over IP
VQM	Video Quality Metric
VQMT	Video Quality Measurement Tool
VR	Virtual Reality
VSP	Video Service Provider
WFL	Weber-Fechner Law

REFERENCES

[1] Patricia Seybold Group, “Customers.com - Quality of experience benchmark - What kind of branded customer experience does your e-business deliver?,” Report, 2000.

[2] A. van Moorsel, “Metrics for the Internet age: Quality of Experience and Quality of Business,” HP Report, 2001.

[3] Nokia, “Quality of Experience (QoE) of mobile services: Can it be measured and improved?,” White Paper, 2004.

[4] Recommendation ITU-T P.10/G.100 (2006) - Amendment 2 (07/08), Vocabulary for performance and quality of service - New definitions for inclusion in Recommendation ITU-T P.10/G.100.

[5] ETSI TR 102 643 v1.0.2, Human Factors (HF); Quality of Experience (QoE) requirements for real-time communication services, 2010.

[6] European Network on Quality of Experience in Multimedia Systems and Services (QUALINET), “Definitions of Quality of Experience (QoE) and related concepts,” White Paper, 2012.

[7] Dagstuhl Seminar on Demarcating User Experience, “User experience white paper - Bringing clarity to the concept of user experience,” White Paper, 2010.

[8] Recommendation ITU-T E.800 (2008), Terms and definitions related to quality of service and network performance including dependability.

[9] ITU-T Study Group 2 (2005), Teletraffic Engineering Handbook.

[10] E. Crawley, R. Nair, B. Rajagopalan, and H. Sandick, “RFC 2386: A framework for QoS-based routing in the Internet,” IETF Technical Report, 1998.

[11] G. Ghinea and J. P. Thomas, “Quality of perception: User quality of service in multimedia presentations,” IEEE Trans. Multimed., vol. 7, no. 4, pp. 786–789, Aug. 2005.

[12] Open Learning World.Com, Course on Quality of Design and Quality of Conformance.

[13] J. Zhang and N. Ansari, “On assuring end-to-end QoE in next generation networks: Challenges and a possible solution,” IEEE Commun. Mag., vol. 49, no. 7, pp. 185–191, July 2011.

[14] ACE project - Advancing the Customer Experience, Online: https://www.schatz.cc/qoe/ace.

[15] L. Skorin-Kapov and M. Varela, “A multi-dimensional view of QoE: the ARCU model,” International Convention MIPRO, pp. 662–666, 2012.

[16] P. Brooks and B. Hestnes, “User measures of quality of experience: Why being objective and quantitative is important,” IEEE Network, vol. 24, no. 2, pp. 8–13, Mar. 2010.

[17] D. Soldani, M. Li, and R. Cuny (Eds.), “QoS and QoE management in UMTS cellular systems,” Wiley, 2006.

[18] R. Schatz, S. Schwarzmann, T. Zinner, O. Dobrijevic, E. Liotou, P. Pocta, S. Barakovic, J. Barakovic Husic, and L. Skorin-Kapov, “QoE Management for future networks,” Autonomous Control for a Reliable Internet of Services: Methods, Models, Approaches, Techniques, Algorithms and Tools, Springer International Publishing, Editors: I. Ganchev, R. van der Mei, J.L. van den Berg, to appear.

[19] P. Reichl, B. Tuffin, and R. Schatz, “Logarithmic laws in service quality perception: Where microeconomics meets psychophysics and quality of experience,” Telecommun. Syst., vol. 52, no. 2, pp. 587–600, Feb. 2013.

[20] M. Fiedler, T. Hoßfeld, and P. Tran-Gia, “A generic quantitative relationship between quality of experience and quality of service,” IEEE Network, vol. 24, no. 2, pp. 36–41, Mar. 2010.

[21] S. Khorsandroo, R. M. Noor, and S. Khorsandroo, “A generic quantitative relationship to assess interdependency of QoE and QoS,” KSII Transactions on Internet and Information Systems, vol. 7, no. 2, pp. 327–346, Feb. 2013.

[22] D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, “The need for QoE-driven interference management in femtocell-overlaid cellular networks,” Mobile and Ubiquitous Systems: Computing, Networking, and Services, vol. 131., I. Stojmenovic, Z. Cheng, and S. Guo, Eds.: Springer International Publishing, pp. 588–601, 2014.

[23] G. Gómez, J. Lorca, R. García, and Q. Pérez, “Towards a QoE-driven resource control in LTE and LTE-A networks,” J. Comput. Networks Commun., vol. 2013, article ID 505910, pp. 1–15, 2013.

[24] P. Reichl, P. Maille, P. Zwickl, and A. Sackl, “On the fixpoint problem of QoE-based charging,” International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS), pp. 235–242, 2012.

[25] P. Maille, P. Reichl, and B. Tuffin, “Economics of quality of experience,” Telecommunications Economics, vol. 7216, A.M. Hadjiantonis and B. Stiller Eds., Springer-Verlag, pp. 158–166, 2012.

[26] R. Stankiewicz, P. Cholda, and A. Jajszczyk, “QoX: What is it really?,” IEEE Commun. Mag., vol. 49, no. 4, pp. 148–158, Apr. 2011.

[27] FCC website, Online: http://www.fcc.gov/openinternet.

[28] A. Cuadra-Sanchez, M. Cutanda-Rodriguez, I. Perez-Mateos, A. Aurelius, K. Brunnström, J.-P. Laulajainen, M. Varela, and J. De Vergara, “A global customer experience management architecture,” Future Network & Mobile Summit (FutureNetw), pp. 1–8, 2012.

[29] S. Möller, W.-Y. Chan, N. Cote, T. H. Falk, A. Raake, and M. Wältermann, “Speech quality estimation: Models and trends,” IEEE Signal Process. Mag., vol. 28, no. 6, pp. 18–28, Nov. 2011.

[30] J. Seppänen, M. Varela, and A. Sgora, “An autonomous QoE-driven network management framework,” J. Vis. Commun. Image Represent., vol. 25, no. 3, pp. 565–577, Apr. 2014.

[31] S. Baraković and L. Skorin-Kapov, “Survey and challenges of QoE management issues in wireless networks,” J. Comput. Networks Commun., vol. 2013, article ID 165146, pp. 1–28, 2013.

[32] A. Takahashi, D. Hands, and V. Barriac, “Standardization activities in the ITU for a QoE assessment of IPTV,” IEEE Commun. Mag, vol. 46, no. 2, pp. 78–84, Feb. 2008.

[33] S. Jelassi, G. Rubino, H. Melvin, H. Youssef, and G. Pujolle, “Quality of experience of VoIP service: A survey of assessment approaches and open issues,” IEEE Commun. Surveys Tuts., vol. 14, no. 2, pp. 491–513, Jan. 2012.

[34] R. Schatz, T. Hoßfeld, L. Janowski, and S. Egger, “From packets to people: Quality of experience as a new measurement challenge,” Data Traffic Monitoring and Analysis, vol. 7754, E. Biersack, C. Callegari, and M. Matijasevic, Eds. Heidelberg: Springer Berlin, pp. 219–263, 2013.

[35] P. Rengaraju, C.-H. Lung, F. Yu, and A. Srinivasan, “On QoE-Monitoring and E2E service assurance in 4G wireless networks,” IEEE Wireless Commun., vol. 19, no. 4, pp. 89–96, Aug. 2012.

[36] G. Piro, L.A. Grieco, G. Boggia, F. Capozzi, and P. Camarda, “Simulating LTE cellular systems: An open source framework,” IEEE Trans. Veh. Technol., vol. 60, no. 2, pp. 498–513, Feb. 2011.

[37] Recommendation ITU-T P.800 (1998), Methods for subjective determination of transmission quality.

[38] K.-T. Chen, C.-C. Tu, and W.-C. Xiao, “OneClick: A framework for measuring network quality of experience,” IEEE International Conference on Computer Communications (INFOCOM), pp. 702–710, 2009.

[39] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, “Best practices for QoE crowdtesting: QoE assessment with crowdsourcing,” IEEE Trans. Multimed., vol. 16, no. 2, pp. 541–558, Feb. 2014.

[40] Recommendation ITU-T G.1011 (2010), Reference guide to quality of experience assessment methodologies.

[41] Y. Chen, K. Wu, and Q. Zhang, “From QoS to QoE: A tutorial on video quality assessment,” IEEE Commun. Surveys Tuts., vol. 17, no. 2, pp. 1126–1165, Jan. 2015.

[42] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, “Objective video quality assessment methods: A classification, review, and performance comparison,” IEEE Trans. Broadcast., vol. 57, no. 2, pp. 165–182, Jun. 2011.

[43] P. Juluri, V. Tamarapalli, and D. Medhi, “Measurement of quality of experience of video-on-demand services: A survey,” IEEE Commun. Surveys Tuts., vol. 18, no. 1, pp. 401–418, Jan. 2016.

[44] M. Alreshoodi and J. Woods, “Survey on QoE\QoS correlation models for multimedia services,” International Journal of Distributed and Parallel Systems (IJDPS), vol.4, no.3, May 2013.

[45] Recommendation ITU-T P.862 (2001), Perceptual evaluation of speech quality (PESQ).

[46] Recommendation ITU-T P.564 (2007), Conformance testing for voice over IP transmission quality assessment models.

[47] D. De Vera, P. Rodriguez-Bocca, and G. Rubino, “Automatic quality of experience measuring on video delivering networks,” ACM SIGMETRICS Perform. Eval. Rev., vol. 36, no. 2, pp. 79–82, Aug. 2008.

[48] P. Charonyktakis, M. Plakia, I. Tsamardinos, and M. Papadopouli, “On user-centric modular QoE prediction for VoIP based on machine-learning algorithms,” IEEE Trans. Mob. Comput., vol. 15, no. 6, pp. 1443–1456, Jun. 2016.

[49] A. Khan, L. Sun, E. Ifeachor, J.-O. Fajardo, F. Liberal, and H. Koumaras, “Video quality prediction models based on video content dynamics for H.264 video over UMTS networks,” Int. J. Digit. Multimed. Broadcast., article ID 608138, pp. 1–17, 2010.

[50] M. Volk, J. Sterle, U. Sedlar, and A. Kos, “An approach to modeling and control of QoE in next generation networks,” IEEE Commun. Mag., vol. 48, no. 8, pp. 126–135, Aug. 2010.

[51] J. Hosek, P. Vajsar, L. Nagy, M. Ries, O. Galinina, S. Andreev, Y. Koucheryavy, Z. Sulc, P. Hais, and R. Penizek, "Predicting user QoE satisfaction in current mobile networks," IEEE International Conference on Communications (ICC), pp. 1088–1093, 2014.

[52] Recommendation ITU-T P.800.1 (2003), Mean Opinion Score (MOS) terminology.

[53] Y. Wang, "Survey of objective video quality measurements," Technical Report, 2006.

[54] S. Möller, "Quality Engineering - Qualität Kommunikationstechnischer Systeme," Springer, 2010.

[55] Recommendation ITU-T G.107 (2014), The E-model: A computational model for use in transmission planning.

[56] Recommendation ITU-T G.108 (1999), Application of the E-model: A planning guide.

[57] R. G. Cole and J. H. Rosenbluth, "Voice over IP performance monitoring," ACM SIGCOMM Comput. Commun. Rev., vol. 31, no. 2, pp. 9–24, Apr. 2001.

[58] Recommendation ITU-T G.1070 (2012), Opinion model for video-telephony applications.

[59] J. Joskowicz, R. Sotelo, and J. C. Lopez Arado, "Comparison of parametric models for video quality estimation: Towards a general model," IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), pp. 1–7, 2012.

[60] T. Liu, N. Narvekar, B. Wang, R. Ding, D. Zou, G. Cash, S. Bhagavathy, and J. Bloom, "Real-time video quality monitoring," EURASIP J. Adv. Signal Process., vol. 2011, no. 1, p. 122, Dec. 2011.

[61] S. Thakolsri, S. Khan, E. Steinbach, and W. Kellerer, "QoE-driven cross-layer optimization for high speed downlink packet access," Journal of Communications, vol. 4, no. 9, pp. 669–680, Oct. 2009.

[62] P. Ameigeiras, J. J. Ramos-Munoz, J. Navarro-Ortiz, P. Mogensen, and J. M. Lopez-Soler, "QoE oriented cross-layer design of a resource allocation algorithm in beyond 3G systems," Computer Communications, vol. 33, no. 5, pp. 571–582, Mar. 2010.

[63] A. Khan, L. Sun, E. Jammeh, and E. Ifeachor, "Quality of experience driven adaptation scheme for video applications over wireless networks," IET Communications, vol. 4, no. 11, pp. 1337–1347, July 2010.

[64] A. Khan, L. Sun, and E. Ifeachor, "Content clustering based video quality prediction model for MPEG4 video streaming over wireless networks," IEEE International Conference on Communications (ICC), pp. 1–5, 2009.

[65] T. Hoßfeld, R. Schatz, E. W. Biersack, and L. Plissonneau, "Internet video delivery in YouTube: From traffic measurements to quality of experience," Data Traffic Monitoring and Analysis, E. Biersack, C. Callegari, and M. Matijasevic, Eds. Springer Berlin Heidelberg, pp. 264–301, 2013.

[66] P. Casas, R. Schatz, and T. Hoßfeld, "Monitoring YouTube QoE: Is your mobile network delivering the right experience to your customers?," IEEE Wireless Communications and Networking Conference (WCNC), pp. 1609–1614, 2013.

[67] P. Szilagyi and C. Vulkan, "Network side lightweight and scalable YouTube QoE estimation," IEEE International Conference on Communications (ICC), pp. 3100–3106, 2015.

[68] T. Hoßfeld, M. Seufert, C. Sieber, and T. Zinner, "Assessing effect sizes of influence factors towards a QoE model for HTTP adaptive streaming," IEEE International Workshop on Quality of Multimedia Experience (QoMEX), pp. 111–116, 2014.

[69] M. Seufert, T. Hoßfeld, and C. Sieber, "Impact of intermediate layer on quality of experience of HTTP adaptive streaming," International Conference on Network and Service Management (CNSM), pp. 256–260, 2015.

[70] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," IEEE Commun. Surveys Tuts., vol. 17, no. 1, pp. 469–492, Jan. 2015.

[71] F. Wamser, S. Deschner, T. Zinner, and P. Tran-Gia , "Investigation of different approaches for QoE oriented scheduling in OFDMA networks," Mobile Networks and Management, Eds. Springer, vol. 125, pp. 172–187, 2013.

[72] Y. Wang and X. Lin, "User-provided networking for QoE provisioning in mobile networks," IEEE Wireless Commun., vol. 22, no. 4, pp. 26–33, Aug. 2015.

[73] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!," IEEE International Workshop on Quality of Multimedia Experience (QoMEX), pp. 131–136, 2011.

[74] K. Doppler, M. Rinne, C. Wijting, C. B. Ribeiro, and K. Hugl, "Device-to-Device communication as an underlay to LTE-advanced networks," IEEE Commun. Mag., vol. 47, no. 12, pp. 42–49, Dec. 2009.

[75] A. Asadi, Q. Wang, and V. Mancuso, "A survey on Device-to-Device communication in cellular networks," IEEE Commun. Surv. Tutorials, vol. 16, no. 4, pp. 1801–1819, Apr. 2014.

[76] C. Quadros, E. Cerqueira, A. Neto, A. Pescapé, A. Riker, R. Immich, and M. Curado, "A quality of experience handover system for heterogeneous multimedia wireless networks," International Conference on Computing, Networking and Communications (ICNC), pp. 1064–1068, 2013.

[77] M. Varela and J.-P. Laulajainen, "QoE-driven mobility management - Integrating the users' quality perception into network - Level decision making," IEEE International Workshop on Quality of Multimedia Experience (QoMEX), pp. 19–24, 2011.

[78] H. A. Tran, A. Mellouk, S. Hoceini, and B. Augustin, "Global state-dependent QoE based routing," IEEE International Conference on Communications (ICC), pp. 131–135, 2012.

[79] L. Xie, C. Hu, W. Wu, and Z. Shi, "QoE-aware power allocation algorithm in multiuser OFDM systems," International Conference on Mobile Ad-hoc and Sensor Networks (MSN), pp. 418–422, 2011.

[80] P. Reichl, B. Tuffin, and R. Schatz, "Economics of logarithmic quality-of-experience in communication networks," Telecommunications Internet and Media Techno Economics (CTTE), pp. 1–8, 2010.

[81] M. Yang, S. Lim, H. Park, and N. H. Park, "Solving the data overload: Device-to-Device bearer control architecture for cellular data offloading," IEEE Veh. Technol. Mag., vol. 8, no. 1, pp. 31–39, Mar. 2013.

[82] L. Lei, Z. Zhong, C. Lin, and X. Shen, "Operator controlled Device-to-Device communications in LTE-advanced networks," IEEE Wireless Commun., vol. 19, no. 3, pp. 96–104, June 2012.

[83] 3GPP TS 36.300, Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN), Rel. 9, 2013.

[84] NGMN Alliance, 5G White Paper, version 1.0, March 2015.

[85] P. Casas, A. Sackl, R. Schatz, L. Janowski, J. Turk, and R. Irmer, "On the quest for new KPIs in mobile networks: The impact of throughput fluctuations on QoE," IEEE International Conference on Communication Workshop (ICCW), pp. 1705–1710, 2015.

[86] T. Hoßfeld, M. Seufert, C. Sieber, T. Zinner, and P. Tran-Gia, "Identifying QoE optimal adaptation of HTTP adaptive streaming based on subjective studies," Computer Networks, vol. 81, pp. 320–332, Apr. 2015.

[87] G. Tian and Y. Liu, "Towards agile and smooth video adaptation in dynamic HTTP streaming," International Conference on Emerging Networking Experiments and Technologies (CoNEXT), pp. 109–120, 2012.

[88] E. Jammeh, I. Mkwawa, A. Khan, M. Goudarzi, L. Sun, and E. Ifeachor, "Quality of Experience (QoE) driven adaptation scheme for voice/video over IP," Telecommun. Syst., vol. 49, no. 1, pp. 99–111, Jan. 2012.

[89] C. Chen, X. Zhu, G. de Veciana, A. C. Bovik, and R. W. Heath, "Rate adaptation and admission control for video transmission with subjective quality constraints," IEEE J. Sel. Top. Signal Process., vol. 9, no. 1, pp. 22–36, Feb. 2015.

[90] S. Thakolsri, W. Kellerer, and E. Steinbach, "QoE-based cross-layer optimization of wireless video with unperceivable temporal video quality fluctuation," IEEE International Conference on Communications (ICC), pp. 1–6, 2011.

[91] A. Sackl, P. Casas, R. Schatz, L. Janowski, and R. Irmer, "Quantifying the impact of network bandwidth fluctuations and outages on web QoE," IEEE International Workshop on Quality of Multimedia Experience (QoMEX), pp. 1–6, 2015.

[92] 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 Commun. Surveys Tuts., vol. 15, no. 2, pp. 678–700, Jan. 2013.

[93] M. Taranetz, T. Blazek, T. Kropfreiter, M. K. Müller, S. Schwarz, and M. Rupp, "Runtime precoding: Enabling multipoint transmission in LTE-Advanced system-level simulations," IEEE Access, vol. 3, pp. 725–736, Jun. 2015.

[94] S. Schwarz, C. Mehlführer, and M. Rupp, "Low complexity approximate maximum throughput scheduling for LTE," Asilomar Conference on Signals, Systems and Computers, pp. 1563–1569, 2010.

[95] Ericsson Mobility Report: Mobile World Congress Edition, Feb. 2015.

[96] C. Verikoukis, L. Alonso, and T. Giamalis, "Cross-layer optimization for wireless systems: A European research key challenge," IEEE Commun. Mag., vol. 43, no. 7, pp. 1–3, July 2005.

[97] G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles, "Towards a better understanding of context and context-awareness," Handheld and Ubiquitous Computing, vol. 1707, HW. Gellersen, Eds., Heidelberg: Springer, Berlin, pp. 304–307, 1999.

[98] A. Sarma, S. Chakraborty, and S. Nandi, “Context aware handover management: Sustaining QoS and QoE in a public IEEE 802.11e hotspot,” IEEE Trans. on Network and Service Management, vol. 11, no. 4, pp. 530–543, Dec. 2014.

[99] Y. Zhu, I. Heynderickx, and J. A. Redi, “Understanding the role of social context and user factors in video quality of experience,” Comput. Human Behav., vol. 49, pp. 412–426, Aug. 2015.

[100] K. Mitra, A. Zaslavsky, and C. Ahlund, “Context-aware QoE modelling, measurement, and prediction in mobile computing systems,” IEEE Trans. Mob. Comput., vol. 14, no. 5, pp. 920–936, May 2015.

[101] T. Hoßfeld, L. Skorin-Kapov, Y. Haddad, P. Pocta, V. Siris, A. Zgank, and H. Melvin, “Can context monitoring improve QoE? A case study of video flash crowds in the internet of services,” IFIP/IEEE International Symposium on Integrated Network Management (IM), pp. 1274–1277, 2015.

[102] H. Riiser, T. Endestad, P. Vigmostad, C. Griwodz, and P. Halvorsen, “Video streaming using a location-based bandwidth-lookup service for bitrate planning,” ACM Trans. Multimed. Comput. Commun. Appl., vol. 8, no. 3, pp. 1–19, July 2012.

[103] J. Hao, R. Zimmermann, and H. Ma, “Gtube: Geo-predictive video streaming over HTTP in mobile environments,” ACM Multimedia Systems Conference (MMSys), pp. 259–270, 2014.

[104] V. Ramamurthi, O. Oyman, and J. Foerster, “Using link awareness for HTTP adaptive streaming over changing wireless conditions,” International Conference on Computing, Networking and Communications (ICNC), pp. 727–731, 2015.

[105] S. Sadr and S. Valentin, “Anticipatory buffer control and resource allocation for wireless video streaming,” arXiv:1304.3056, Apr. 2013.

[106] S. Mekki and S. Valentin, “Anticipatory quality adaptation for mobile streaming: Fluent video by channel prediction,” IEEE International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), pp. 1–3, 2015.

[107] S. Mekki, T. Karagkioules, and S. Valentin, “HTTP adaptive streaming with indoors-outdoors detection in mobile networks,” arXiv:1705.08809, May 2017.

[108] C. Müller, S. Lederer, and C. Timmerer, “An evaluation of dynamic adaptive streaming over HTTP in vehicular environments,” Workshop on Mobile Video (MoVid), pp. 37–42, 2012.

[109] Cisco Visual Networking Index: Forecast and Methodology, 2016–2021, Sept. 2017.

[110] World Economic Forum, Digital Transformation Initiative Telecommunications Industry, White Paper, 2017.

[111] GSMA, “The Mobile Economy,” Report, 2017.

[112] ONF SDN, SDN Architecture issue 1.0 - Open Networking Foundation, 2014.

[113] ONF SDN, SDN Architecture issue 1.1 - Open Networking Foundation, 2016.

[114] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella “On multi-access edge computing: A survey of the emerging 5G network edge architecture & orchestration,” IEEE Commun. Surveys Tuts., vol. 19, no. 3, pp. 1657–1681, May 2017.

[115] A. Kassler, L. Skorin-Kapov, O. Dobrijevic, M. Matijasevic, and P. Dely, “Towards QoE-driven multimedia service negotiation and path optimization with software defined networking,” International Conference on Software, Telecommunications and Computer Networks (SoftCOM), pp. 1–5, 2012.

[116] H. Nam, K. H. Kim, J. Y. Kim, and H. Schulzrinne, “Towards QoE-aware video streaming using SDN,” IEEE Global Communications Conference (GLOBECOM), pp. 1317–1322, 2014.

[117] G. Cofano, L. De Cicco, T. Zinner, A. Nguyen-Ngoc, P. Tran-Gia, and S. Mascolo, “Design and experimental evaluation of network-assisted strategies for HTTP adaptive streaming,” International Conference on Multimedia Systems (MMSys), article no. 3, pp. 1–12, 2016.

[118] A. Bentaleb, A. C. Begen, and R. Zimmermann, “SDNDASH: Improving QoE of HTTP adaptive streaming using software defined networking,” ACM on Multimedia Conference (MM), pp. 1296–1305, 2016.

[119] A. Bentaleb, A. C. Begen, R. Zimmermann, and S. Harous, “SDNHAS: An SDN-enabled architecture to optimize QoE in HTTP Adaptive Streaming,” IEEE Trans. Multimed., vol. 19, no. 10, pp. 2136–2151, Oct. 2017.

[120] J. W. Kleinrouweler, S. Cabrero, and P. Cesar, “Delivering stable high-quality video: An SDN architecture with DASH assisting network elements,” International Conference on Multimedia Systems (MMSys), article no. 4, pp. 1–10, 2016.

[121] D. Bhat, A. Rizk, M. Zink, and R. Steinmetz, “Network assisted content distribution for adaptive bitrate video streaming,” ACM on Multimedia Systems Conference (MMSys), pp. 62–75, 2017.

[122] P. Georgopoulos, Y. Elkhatib, M. Broadbent, M. Mu, and N. Race, “Towards network-wide QoE fairness using openflow-assisted adaptive video streaming,” ACM SIGCOMM Workshop on Future human-centric multimedia networking (FhMN), pp. 15–20, 2013.

[123] A. Gudipati, D. Perry, L. E. Li, and S. Katti, “SoftRAN: Software defined radio access network,” ACM SIGCOMM Workshop on Hot topics in Software Defined Networking (HotSDN), pp. 25–30, 2013.

[124] T. Chen, M. Matinmikko, X. Chen, X. Zhou, and P. Ahokangas, “Software defined mobile networks: Concept, survey, and research directions,” IEEE Commun. Mag., vol. 53, no. 11, pp. 126–133, Nov. 2015.

[125] M. Gramaglia, I. Digon, V. Friderikos, D. von Hugo, C. Mannweiler, M. A. Puente, K. Samdanis, and B. Sayadi, “Flexible connectivity and QoE/QoS management for 5G Networks: The 5G NORMA view,” IEEE International Conference on Communications Workshops (ICC), pp. 373–379, 2016.

[126] T. Taleb and A. Ksentini, “QoS/QoE predictions-based admission control for femto communications,” IEEE International Conference on Communications (ICC), pp. 5146–5150, 2012.

[127] S. Dutta, T. Taleb, P. A. Frangoudis, and A. Ksentini, “On-the-fly QoE-aware transcoding in the mobile edge,” IEEE Global Communications Conference (GLOBECOM), pp. 1–6, 2016.

[128] M. Mu, M. Broadbent, A. Farshad, N. Hart, D. Hutchison, Q. Ni, and N. Race, “A scalable user fairness model for adaptive video streaming over SDN-assisted future networks,” IEEE J. Sel. Areas Commun., vol. 34, no. 8, pp. 2168–2184, Aug. 2016.

[129] S. Ramakrishnan, and X. Zhu, “An SDN based approach to measuring and optimizing ABR video quality of experience,” Cisco Systems, Technical Paper, 2014.

[130] M. Katsarakis, G. Fortetsanakis, P. Charonyktakis, A. Kostopoulos, and M. Papadopouli, “On user-centric tools for QoE-based recommendation and real-time analysis of large-scale markets,” IEEE Commun. Mag., vol. 52, no. 9, pp. 37–43, Sep. 2014.

[131] E. Liotou, G. Tseliou, K. Samdanis, D. Tsolkas, F. Adelantado, and C. Verikoukis, “An SDN QoE-service for dynamically enhancing the performance of OTT applications,” IEEE International Workshop on Quality of Multimedia Experience (QoMEX), pp. 1–2, 2015.

[132] A. Ahmad, A. Floris, and L. Atzori “QoE-centric service delivery: A collaborative approach among OTTs and ISPs,” Computer Networks, vol. 110, pp. 168–179, Dec. 2016.

[133] S. Peng, J. O. Fajardo, P. S. Khodashenas, B. Blanco, F. Liberal, C. Ruiz, C. Turyagyenda, M. Wilson, and S. Vadgama, “QoE-oriented mobile edge service management leveraging SDN and NFV,” Mob. Inf. Syst., vol. 2017, pp. 1–14, Jan. 2017.

[134] M. Jarschel, “Chances and challenges of SDN-enabled QoE management,” European Conference on Networks and Communications (EuCNC), 2015.

[135] 3GPP TR 23.708, Architecture enhancement for service capability exposure, Rel.13, 2015.

[136] GSMA OneAPI, Online: www.gsma.com/oneapi.

[137] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, “SLAW: Self-similar least-action human walk,” IEEE/ACM Trans. Netw., vol. 20, no. 2, pp. 515–529, Apr. 2012.

[138] A. Nadembega, A. Hafid, and T. Taleb, “Mobility-Prediction-Aware bandwidth reservation scheme for mobile networks,” IEEE Trans. Veh. Technol., vol. 64, no. 6, pp. 2561–2576, Jun. 2015.

[139] J. C. Ikuno, M. Wrulich, and M. Rupp, “System level simulation of LTE networks,” IEEE Vehicular Technology Conference (VTC-Spring), pp. 1–5, 2010.

[140] A. Farshad, P. Georgopoulos, M. Broadbent, M. Mu, and N. Race, “Leveraging SDN to provide an in-network QoE measurement framework,” IEEE Conference on Computer Communications Workshops (INFOCOM), pp. 239–244, 2015.

[141] T. Zinner, T. Hoßfeld, T. N. Minash, and M. Fiedler, “Controlled vs. uncontrolled degradations of QoE: The provisioning-delivery hysteresis in case of video,” EuroITV Workshop: Quality of Experience for Multimedia Content Sharing, 2010.

[142] A. Sackl, P. Zwickl, and P. Reichl, “The trouble with choice: An empirical study to investigate the influence of charging strategies and content selection on QoE,” International Conference on Network and Service Management (CNSM), pp. 298–303, 2013.

[143] 3GPP TS 23.203, Policy and charging control architecture, Rel. 14, 2017.

[144] H. Ekstrom, “QoS Control in the 3GPP evolved packet system,” IEEE Commun. Mag., vol. 47, no. 2, pp. 76–83, Feb. 2009.

[145] M. Richart, J. Baliosian, J. Serrat, and J.-Luis Gorricho, “Resource slicing in virtual wireless networks: A survey,” IEEE Trans. on Network and Service Management, vol. 13, no. 3, pp. 462–476, Sept. 2016.

[146] 3GPP TS 23.501, System architecture for the 5G system, Rel. 15, 2017.

[147] ETSI GS NFV 002 v1.1.1, Network Functions Virtualisation (NFV); Architectural Framework, 2013.