

A COMPREHENSIVE
GUIDE TO

5G SECURITY



EDITED BY
MADHUSANKA LIYANAGE, IJAZ AHMAD, AHMED BUXT ABRO,
ANDREI GURTOV, AND MIKA YLIANTTILA

WILEY

A Comprehensive Guide to 5G Security

A Comprehensive Guide to 5G Security

Edited by

Madhusanka Liyanage
University of Oulu, Finland

Ijaz Ahmad
University of Oulu, Finland

Ahmed Bux Abro
VMware Inc., USA

Andrei Gurtov
Linköping University, Sweden

Mika Ylianttila
University of Oulu, Finland

WILEY

This edition first published 2018
© 2018 John Wiley & Sons Ltd

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by law. Advice on how to obtain permission to reuse material from this title is available at <http://www.wiley.com/go/permissions>.

The right of Madhusanka Liyanage, Ijaz Ahmad, Ahmed Bux Abro, Andrei Gurto and Mika Ylianttila to be identified as the authors of the editorial material in this work has been asserted in accordance with law.

Registered Offices

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office

The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Limit of Liability/Disclaimer of Warranty

While the publisher and authors have used their best efforts in preparing this work, they make no representations or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives, written sales materials or promotional statements for this work. The fact that an organization, website, or product is referred to in this work as a citation and/or potential source of further information does not mean that the publisher and authors endorse the information or services the organization, website, or product may provide or recommendations it may make. This work is sold with the understanding that the publisher is not engaged in rendering professional services. The advice and strategies contained herein may not be suitable for your situation. You should consult with a specialist where appropriate. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

Library of Congress Cataloging-in-Publication Data

Names: Liyanage, Madhusanka, editor. | Ahmad, Ijaz, 1985- editor. | Abro, Ahmed Bux, editor. | Gurto, Andrei, editor. | Ylianttila, Mika, editor.

Title: A Comprehensive guide to 5G security / edited by Madhusanka Liyanage, Ijaz Ahmad, Ahmed Bux Abro, Andrei Gurto, Mika Ylianttila.

Description: Hoboken, NJ : John Wiley & Sons, 2018. | Includes index. | Identifiers: LCCN 2017040682 (print) | LCCN 2017047712 (ebook) | ISBN 9781119293088 (pdf) | ISBN 9781119293057 (epub) | ISBN 9781119293040 (cloth)

Subjects: LCSH: Mobile communication systems—Security measures. | Wireless communication systems—Security measures.

Classification: LCC TK5103.2 (ebook) | LCC TK5103.2 .C649 2018 (print) | DDC 005.8–dc23
LC record available at <https://lccn.loc.gov/2017040682>

Cover Design: Wiley

Cover Images: (Background) © cinoby/Gettyimages; (Lock overlay) © TCmake_photo/Gettyimages; (Towers) © Nikifor Todorov/Shutterstock; (Drone) © Robert Mandel/Shutterstock

Set in 10/12pt Warnock by SPi Global, Pondicherry, India

Contents

The Editors xv

About the Contributors xix

Foreword xxxi

Preface xxxii

Acknowledgements xxxix

Part I 5G Security Overview 1

1 Evolution of Cellular Systems 3

*Shahriar Shahabuddin, Sadiqur Rahaman, Faisal Rehman,
Ijaz Ahmad, and Zaheer Khan*

1.1 Introduction 3

1.2 Early Development 4

1.3 First Generation Cellular Systems 6

 1.3.1 Advanced Mobile Phone Service 7

 1.3.2 Security in 1G 7

1.4 Second Generation Cellular Systems 8

 1.4.1 Global System for Mobile Communications 8

 1.4.2 GSM Network Architecture 9

 1.4.3 Code Division Multiple Access 10

 1.4.4 Security in 2G 10

 1.4.5 Security in GSM 11

 1.4.6 Security in IS-95 14

1.5 Third Generation Cellular Systems 15

 1.5.1 CDMA 2000 15

 1.5.2 UMTS WCDMA 15

 1.5.3 UMTS Network Architecture 16

 1.5.4 HSPA 17

 1.5.5 Security in 3G 17

 1.5.6 Security in CDMA2000 17

 1.5.7 Security in UMTS 18

1.6 Cellular Systems beyond 3G 20

 1.6.1 HSPA+ 20

 1.6.2 Mobile WiMAX 20

| | | |
|----------|-----------------------------------------------------------------------------------------|-----------|
| 1.6.3 | LTE | 21 |
| 1.6.4 | LTE Network Architecture | 21 |
| 1.7 | Fourth Generation Cellular Systems | 22 |
| 1.7.1 | Key Technologies of 4G | 23 |
| 1.7.2 | Network Architecture | 24 |
| 1.7.3 | Beyond 3G and 4G Cellular Systems Security | 25 |
| 1.7.4 | LTE Security Model | 26 |
| 1.7.5 | Security in WiMAX | 27 |
| 1.8 | Conclusion | 27 |
| | References | 28 |
| 2 | 5G Mobile Networks: Requirements, Enabling Technologies, and Research Activities | 31 |
| | <i>Van-Giang Nguyen, Anna Brunstrom, Karl-Johan Grinnemo, and Javid Taheri</i> | |
| 2.1 | Introduction | 31 |
| 2.1.1 | What is 5G? | 31 |
| 2.1.2 | Typical Use Cases | 32 |
| 2.2 | 5G Requirements | 33 |
| 2.2.1 | High Data Rate and Ultra Low Latency | 34 |
| 2.2.2 | Massive Connectivity and Seamless Mobility | 35 |
| 2.2.3 | Reliability and High Availability | 35 |
| 2.2.4 | Flexibility and Programmability | 36 |
| 2.2.5 | Energy, Cost and Spectrum Efficiency | 36 |
| 2.2.6 | Security and Privacy | 36 |
| 2.3 | 5G Enabling Technologies | 37 |
| 2.3.1 | 5G Radio Access Network | 38 |
| 2.3.2 | 5G Mobile Core Network | 44 |
| 2.3.3 | 5G End-to-End System | 46 |
| 2.4 | 5G Standardization Activities | 48 |
| 2.4.1 | ITU Activities | 48 |
| 2.4.2 | 3GPP Activities | 49 |
| 2.4.3 | ETSI Activities | 50 |
| 2.4.4 | IEEE Activities | 51 |
| 2.4.5 | IETF Activities | 52 |
| 2.5 | 5G Research Communities | 52 |
| 2.5.1 | European 5G Related Activities | 52 |
| 2.5.2 | Asian 5G Related Activities | 53 |
| 2.5.3 | American 5G Related Activities | 54 |
| 2.6 | Conclusion | 55 |
| 2.7 | Acknowledgement | 55 |
| | References | 55 |
| 3 | Mobile Networks Security Landscape | 59 |
| | <i>Ahmed Bux Abro</i> | |
| 3.1 | Introduction | 59 |

| | | |
|----------|---------------------------------------------------------------------------------------------------------|-----------|
| 3.2 | Mobile Networks Security Landscape | 59 |
| 3.2.1 | Security Threats and Protection for 1G | 61 |
| 3.2.2 | Security Threats and Protection for 2G | 62 |
| 3.2.3 | Security Threats and Protection for 3G | 63 |
| 3.2.4 | Security Threats and Protection for 4G | 63 |
| 3.2.5 | Security Threats and Protection for 5G | 66 |
| 3.3 | Mobile Security Lifecycle Functions | 70 |
| 3.3.1 | Secure Device Management | 71 |
| 3.3.2 | Mobile OS and App Patch Management | 71 |
| 3.3.3 | Security Threat Analysis and Assessment | 72 |
| 3.3.4 | Security Monitoring | 72 |
| 3.4 | Conclusion | 73 |
| | References | 73 |
| 4 | Design Principles for 5G Security | 75 |
| | <i>Ijaz Ahmad, Madhusanka Liyanage, Shahriar Shahabuddin, Mika Ylianttila, and Andrei Gurkov</i> | |
| 4.1 | Introduction | 75 |
| 4.2 | Overviews of Security Recommendations and Challenges | 76 |
| 4.2.1 | Security Recommendations by ITU-T | 77 |
| 4.2.2 | Security Threats and Recommendations by NGMN | 78 |
| 4.2.3 | Other Security Challenges | 79 |
| 4.3 | Novel Technologies for 5G Security | 81 |
| 4.3.1 | 5G Security Leveraging NFV | 82 |
| 4.3.2 | Network Security Leveraging SDN | 83 |
| 4.3.3 | Security Challenges in SDN | 84 |
| 4.3.4 | Security Solutions for SDN | 86 |
| 4.4 | Security in SDN-based Mobile Networks | 88 |
| 4.4.1 | Data Link Security | 88 |
| 4.4.2 | Control Channels Security | 89 |
| 4.4.3 | Traffic Monitoring | 91 |
| 4.4.4 | Access Control | 91 |
| 4.4.5 | Network Resilience | 91 |
| 4.4.6 | Security Systems and Firewalls | 92 |
| 4.4.7 | Network Security Automation | 92 |
| 4.5 | Conclusions and Future Directions | 94 |
| 4.6 | Acknowledgement | 95 |
| | References | 95 |
| 5 | Cyber Security Business Models in 5G | 99 |
| | <i>Julius Francis Gomes, Marika Iivari, Petri Ahokangas, Lauri Isotalo, Bengt Sahlin, and Jan Melén</i> | |
| 5.1 | Introduction | 99 |
| 5.2 | The Context of Cyber Security Businesses | 100 |
| 5.2.1 | Types of Cyber Threat | 101 |
| 5.2.2 | The Cost of Cyber-Attacks | 102 |

| | | |
|-------|------------------------------------------------------|-----|
| 5.3 | The Business Model Approach | 103 |
| 5.3.1 | The 4C Typology of the ICT Business Model | 104 |
| 5.3.2 | Business Models in the Context of Cyber Preparedness | 105 |
| 5.4 | The Business Case of Cyber Security in the Era of 5G | 106 |
| 5.4.1 | The Users and Issues of Cyber Security in 5G | 108 |
| 5.4.2 | Scenarios for 5G Security Provisioning | 109 |
| 5.4.3 | Delivering Cyber Security in 5G | 110 |
| 5.5 | Business Model Options in 5G Cyber Security | 112 |
| 5.6 | Acknowledgement | 114 |
| | References | 114 |

Part II 5G Network Security 117

6 Physical Layer Security 119

*Simone Soderi, Lorenzo Mucchi, Matti Hämäläinen, Alessandro Piva,
and Jari Linatti*

| | | |
|-------|--------------------------------------------------------------|-----|
| 6.1 | Introduction | 119 |
| 6.1.1 | Physical Layer Security in 5G Networks | 120 |
| 6.1.2 | Related Work | 121 |
| 6.1.3 | Motivation | 121 |
| 6.2 | WBPLSec System Model | 123 |
| 6.2.1 | Transmitter | 124 |
| 6.2.2 | Jamming Receiver | 126 |
| 6.2.3 | Secrecy Metrics | 126 |
| 6.2.4 | Secrecy Capacity of WBPLSec | 128 |
| 6.2.5 | Secrecy Capacity of iJAM | 129 |
| 6.3 | Outage Probability of Secrecy Capacity of a Jamming Receiver | 131 |
| 6.3.1 | Simulation Scenario for Secrecy Capacity | 134 |
| 6.4 | WBPLSec Applied to 5G networks | 136 |
| 6.5 | Conclusions | 138 |
| | References | 139 |

7 5G-WLAN Security 143

*Satish Anamalamudi, Abdur Rashid Sangi, Mohammed Alkatheiri,
Fahad T. Bin Muhaya, and Chang Liu*

| | | |
|-------|------------------------------------------------------------------------|-----|
| 7.1 | Chapter Overview | 143 |
| 7.2 | Introduction to WiFi-5G Networks Interoperability | 143 |
| 7.2.1 | WiFi (Wireless Local Area Network) | 143 |
| 7.2.2 | Interoperability of WiFi with 5G Networks | 144 |
| 7.2.3 | WiFi Security | 144 |
| 7.3 | Overview of Network Architecture for WiFi-5G Networks Interoperability | 146 |
| 7.3.1 | MAC Layer | 147 |
| 7.3.2 | Network Layer | 147 |
| 7.3.3 | Transport Layer | 148 |
| 7.3.4 | Application Layer | 149 |

| | | |
|----------|-----------------------------------------------------------------------------------|-----|
| 7.4 | 5G-WiFi Security Challenges | 150 |
| 7.4.1 | WIFI-5G Security Challenges with Respect to a Large Number of Device Connectivity | 151 |
| 7.4.2 | Security Challenges in 5G Networks and WiFi | 151 |
| 7.5 | Security Consideration for Architectural Design of WiFi-5G Networks | 156 |
| 7.5.1 | User and Device Identity Confidentiality | 156 |
| 7.5.2 | Integrity | 156 |
| 7.5.3 | Mutual Authentication and Key Management | 157 |
| 7.6 | LiFi Networks | 158 |
| 7.7 | Introduction to LiFi-5G Networks Interoperability | 159 |
| 7.8 | 5G-LiFi Security Challenges | 160 |
| 7.8.1 | LIFI-5G Security Challenges with Respect to a Large Number of Device Connectivity | 160 |
| 7.8.2 | Security Challenges in 5G Networks and LiFi | 160 |
| 7.9 | Security Consideration for Architectural Design of LiFi-5G Networks | 160 |
| 7.10 | Conclusion and Future Work | 161 |
| | References | 161 |
| 8 | Safety of 5G Network Physical Infrastructures | 165 |
| | <i>Rui Travanca and João André</i> | |
| 8.1 | Introduction | 165 |
| 8.2 | Historical Development | 168 |
| 8.2.1 | Typology | 168 |
| 8.2.2 | Codes | 170 |
| 8.2.3 | Outlook | 170 |
| 8.3 | Structural Design Philosophy | 171 |
| 8.3.1 | Basis | 171 |
| 8.3.2 | Actions | 174 |
| 8.3.3 | Structural Analysis | 179 |
| 8.3.4 | Steel Design Verifications | 180 |
| 8.4 | Survey of Problems | 181 |
| 8.4.1 | General | 181 |
| 8.4.2 | Design Failures | 182 |
| 8.4.3 | Maintenance Failures | 183 |
| 8.4.4 | Vandalism or Terrorism Failures | 186 |
| 8.5 | Opportunities and Recommendations | 188 |
| 8.6 | Acknowledgement | 190 |
| | References | 191 |
| 9 | Customer Edge Switching: A Security Framework for 5G | 195 |
| | <i>Hammad Kabir, Raimo Kantola, and Jesus Llorente Santos</i> | |
| 9.1 | Introduction | 195 |
| 9.2 | State-of-the-art in Mobile Networks Security | 197 |
| 9.2.1 | Mobile Network Challenges and Principles of Security Framework | 200 |
| 9.2.2 | Trust Domains and Trust Processing | 202 |

| | | |
|-------|--------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 9.3 | CES Security Framework | 203 |
| 9.3.1 | DNS to Initiate Communication | 205 |
| 9.3.2 | CETP Policy-based Communication | 206 |
| 9.3.3 | Policy Architecture | 209 |
| 9.3.4 | CES Security Mechanisms | 209 |
| 9.3.5 | Realm Gateway | 210 |
| 9.3.6 | RGW Security Mechanisms | 212 |
| 9.4 | Evaluation of CES Security | 213 |
| 9.4.1 | Evaluating the CETP Policy-based Communication | 214 |
| 9.4.2 | Evaluation of RGW Security | 217 |
| 9.5 | Deployment in 5G Networks | 222 |
| 9.5.1 | Use Case 1: Mobile Broadband | 224 |
| 9.5.2 | Use Case 2: Corporate Gateway | 225 |
| 9.5.3 | Use Case 3: National CERT Centric Trust Domain | 226 |
| 9.5.4 | Use Case 4: Industrial Internet for Road Traffic and Transport | 227 |
| 9.6 | Conclusion | 228 |
| | References | 230 |
| 10 | Software Defined Security Monitoring in 5G Networks | 231 |
| | <i>Madhusanka Liyanage, Ijaz Ahmad, Jude Okwuibe, Edgardo Montes de Oca, Hoang Long Mai, Oscar López Perez, and Mikel Uriarte Itzazelaia</i> | |
| 10.1 | Introduction | 231 |
| 10.2 | Existing Monitoring Techniques | 232 |
| 10.3 | Limitations of Current Monitoring Techniques | 233 |
| 10.4 | Use of Monitoring in 5G | 234 |
| 10.5 | Software-Defined Monitoring Architecture | 235 |
| 10.6 | Expected Advantages of Software Defined Monitoring | 238 |
| 10.7 | Expected Challenges in Software Defined Monitoring | 240 |
| 10.8 | Conclusion | 242 |
| | References | 243 |

Part III 5G Device and User Security 245

| | | |
|--------|-----------------------------------------------------|-----|
| 11 | IoT Security | 247 |
| | <i>Mehrnoosh Monshizadeh and Vikramajeet Khatri</i> | |
| 11.1 | Introduction | 247 |
| 11.2 | Related Work | 248 |
| 11.3 | Literature Overview and Research Motivation | 249 |
| 11.3.1 | IoT Devices, Services and Attacks on Them | 250 |
| 11.3.2 | Research Motivation | 253 |
| 11.4 | Distributed Security Platform | 254 |
| 11.4.1 | Robot Data Classification | 254 |
| 11.4.2 | Robot Attack Classification | 255 |
| 11.4.3 | Robot Security Platform | 256 |

| | | |
|-----------|--------------------------------------------------------------------------------------------------------------------------------------|------------|
| 11.5 | Mobile Cloud Robot Security Scenarios | 259 |
| 11.5.1 | Robot with SIMcard | 259 |
| 11.5.2 | SIMless Robot | 260 |
| 11.5.3 | Robot Attack | 263 |
| 11.5.4 | Robot Communication | 263 |
| 11.6 | Conclusion | 263 |
| | References | 265 |
| 12 | User Privacy, Identity and Trust in 5G | 267 |
| | <i>Tanesh Kumar, Madhusanka Liyanage, Ijaz Ahmad, An Braeken, and Mika Ylianttila</i> | |
| 12.1 | Introduction | 267 |
| 12.2 | Background | 268 |
| 12.3 | User Privacy | 269 |
| 12.3.1 | Data Privacy | 269 |
| 12.3.2 | Location Privacy | 271 |
| 12.3.3 | Identity Privacy | 272 |
| 12.4 | Identity Management | 273 |
| 12.5 | Trust Models | 274 |
| 12.6 | Discussion | 277 |
| 12.7 | Conclusion | 278 |
| | References | 279 |
| 13 | 5G Positioning: Security and Privacy Aspects | 281 |
| | <i>Elena Simona Lohan, Anette Alén-Savikko, Liang Chen, Kimmo Järvinen, Helena Leppäkoski, Heidi Kuusniemi, and Päivi Korpisaari</i> | |
| 13.1 | Introduction | 281 |
| 13.2 | Outdoor versus Indoor Positioning Technologies | 283 |
| 13.3 | Passive versus Active Positioning | 283 |
| 13.4 | Brief Overview of 5G Positioning Mechanisms | 285 |
| 13.5 | Survey of Security Threats and Privacy Issues in 5G Positioning | 291 |
| 13.5.1 | Security Threats in 5G Positioning | 291 |
| 13.6 | Main Privacy Concerns | 294 |
| 13.7 | Passive versus Active Positioning Concepts | 295 |
| 13.8 | Physical-Layer Based Security Enhancements Mechanisms for Positioning in 5G | 296 |
| 13.8.1 | Reliability Monitoring and Outlier Detection Mechanisms | 296 |
| 13.8.2 | Detection, Location and Estimation of Interference Signals | 297 |
| 13.8.3 | Backup Systems | 298 |
| 13.9 | Enhancing Trustworthiness | 299 |
| 13.10 | Cryptographic Techniques for Security and Privacy of Positioning | 299 |
| 13.10.1 | Cryptographic Authentication in Positioning | 300 |
| 13.10.2 | Cryptographic Distance-Bounding | 301 |
| 13.10.3 | Cryptographic Techniques for Privacy-Preserving Location-based Services | 303 |

| | | |
|---------|------------------------------------------------------------------------------------|-----|
| 13.11 | Legislation on User Location Privacy in 5G | 304 |
| 13.11.1 | EU Policy and Legal Framework | 304 |
| 13.11.2 | Legal Aspects Related to the Processing of Location Data | 306 |
| 13.11.3 | Privacy Protection by Design and Default | 306 |
| 13.11.4 | Security Protection | 307 |
| 13.11.5 | A Closer Look at the e-Privacy Directive | 307 |
| 13.11.6 | Summary of EU Legal Instruments | 308 |
| 13.11.7 | International Issues | 308 |
| 13.11.8 | Challenges and Future Scenarios in Legal Frameworks and Policy | 309 |
| 13.12 | Landscape of the European and International Projects related to Secure Positioning | 311 |
| | References | 312 |

Part IV 5G Cloud and Virtual Network Security 321

| | | |
|-----------|---------------------------------------------------------|-----|
| 14 | Mobile Virtual Network Operators (MVNO) Security | 323 |
| | <i>Mehrnoosh Monshizadeh and Vikramajeet Khatri</i> | |
| 14.1 | Introduction | 323 |
| 14.2 | Related Work | 324 |
| 14.3 | Cloudification of the Network Operators | 325 |
| 14.4 | MVNO Security | 326 |
| 14.4.1 | Data Security in TaaS | 327 |
| 14.4.2 | Hypervisor and VM Security in TaaS | 328 |
| 14.4.3 | Application Security in TaaS | 333 |
| 14.4.4 | Summary | 334 |
| 14.4.5 | MVNO Security Benchmark | 337 |
| 14.5 | TaaS Deployment Security | 338 |
| 14.5.1 | IaaS | 338 |
| 14.5.2 | PaaS | 340 |
| 14.5.3 | SaaS | 340 |
| 14.6 | Future Directions | 340 |
| 14.7 | Conclusion | 341 |
| | References | 342 |
| 15 | NFV and NFV-based Security Services | 347 |
| | <i>Wenjing Chu</i> | |
| 15.1 | Introduction | 347 |
| 15.2 | 5G, NFV and Security | 347 |
| 15.3 | A Brief Introduction to NFV | 348 |
| 15.4 | NFV, SDN, and a Telco Cloud | 351 |
| 15.5 | Common NFV Drivers | 353 |
| 15.5.1 | Technology Curve | 353 |
| 15.5.2 | Opportunity Cost and Competitive Landscape | 353 |
| 15.5.3 | Horizontal Network Slicing | 354 |
| 15.5.4 | Multi-Tenancy | 354 |

| | | |
|--------|---------------------------------------------------------------------------|-----|
| 15.5.5 | Rapid Service Delivery | 354 |
| 15.5.6 | XaaS Models | 354 |
| 15.5.7 | One Cloud | 355 |
| 15.6 | NFV Security: Challenges and Opportunities | 355 |
| 15.6.1 | VNF Security Lifecycle and Trust | 355 |
| 15.6.2 | VNF Security in Operation | 358 |
| 15.6.3 | Multi-Tenancy and XaaS | 359 |
| 15.6.4 | OPNFV and Openstack: Open Source Projects for NFV | 360 |
| 15.7 | NFV-based Security Services | 364 |
| 15.7.1 | NFV-based Network Security | 365 |
| 15.7.2 | Policy-based Security Services | 366 |
| 15.7.3 | Machine Learning for NFV-based Security Services | 369 |
| 15.8 | Conclusions | 370 |
| | References | 370 |
| 16 | Cloud and MEC Security | 373 |
| | <i>Jude Okwuibe, Madhusanka Liyanage, Ijaz Ahmad, and Mika Ylianttila</i> | |
| 16.1 | Introduction | 373 |
| 16.2 | Cloud Computing in 5G Networks | 374 |
| 16.2.1 | Overview and History of Cloud Computing | 375 |
| 16.2.2 | Cloud Computing Architecture | 376 |
| 16.2.3 | Cloud Deployment Models | 377 |
| 16.2.4 | Cloud Service Models | 378 |
| 16.2.5 | 5G Cloud Computing Architecture | 379 |
| 16.2.6 | Use Cases/Scenarios of Cloud Computing in 5G | 380 |
| 16.3 | MEC in 5G Networks | 381 |
| 16.3.1 | Overview of MEC Computing | 381 |
| 16.3.2 | MEC in 5G | 383 |
| 16.3.3 | Use Cases of MEC Computing in 5G | 384 |
| 16.4 | Security Challenges in 5G Cloud | 385 |
| 16.4.1 | Virtualization Security | 385 |
| 16.4.2 | Cyber-Physical System (CPS) Security | 386 |
| 16.4.3 | Secure and Private Data Computation | 386 |
| 16.4.4 | Cloud Intrusion | 387 |
| 16.4.5 | Access Control | 387 |
| 16.5 | Security Challenges in 5G MEC | 388 |
| 16.5.1 | Denial of Service (DoS) Attack | 389 |
| 16.5.2 | Man-in-the-Middle (MitM) | 389 |
| 16.5.3 | Inconsistent Security Policies | 389 |
| 16.5.4 | VM Manipulation | 390 |
| 16.5.5 | Privacy Leakage | 390 |
| 16.6 | Security Architectures for 5G Cloud and MEC | 391 |
| 16.6.1 | Centralized Security Architectures | 391 |
| 16.6.2 | SDN-based Cloud Security Systems | 392 |
| 16.7 | 5GMEC, Cloud Security Research and Standardizations | 392 |
| 16.8 | Conclusions | 394 |
| | References | 394 |

| | |
|--------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| 17 Regulatory Impact on 5G Security and Privacy | 399 |
| <i>Jukka Salo and Madhusanka Liyanage</i> | |
| 17.1 | Introduction 399 |
| 17.2 | Regulatory Objectives for Security and Privacy 401 |
| 17.2.1 | Generic Objectives 401 |
| 17.3 | Legal Framework for Security and Privacy 402 |
| 17.3.1 | General Framework 402 |
| 17.3.2 | Legal Framework for Security and Privacy in Cloud Computing 403 |
| 17.3.3 | Legal Framework for Security and Privacy in Software Defined Networking and Network Function Virtualization 405 |
| 17.4 | Security and Privacy Issues in New 5G Technologies 405 |
| 17.4.1 | Security and Privacy Issues in Cloud Computing 405 |
| 17.4.2 | Security and Privacy Issues in Network Functions Virtualization 407 |
| 17.4.3 | Security and Privacy Issues in Software Defined Networking (SDN) 409 |
| 17.4.4 | Summary of Security and Privacy Issues in the Context of Technologies under Study (Clouds, NFV, SDN) 410 |
| 17.5 | Relevance Assessment of Security and Privacy Issues for Regulation 411 |
| 17.6 | Analysis of Potential Regulatory Approaches 412 |
| 17.7 | Summary of Issues and Impact of New Technologies on Security and Privacy Regulation 413 |
| References | 417 |

| | |
|--------------|------------|
| Index | 421 |
|--------------|------------|

The Editors

Madhusanka Liyanage

Centre for Wireless Communications, University of Oulu, Finland.

Madhusanka Liyanage received his BSc degree (First Class Honors) in Electronics and Telecommunication Engineering from the University of Moratuwa, Moratuwa, Sri Lanka, in 2009, his MEng degree from the Asian Institute of Technology, Bangkok, Thailand, in 2011, and his MSc degree from University of Nice Sophia Antipolis, Nice, France in 2011. In 2016, Liyanage received his PhD in Communication Engineering from the University of Oulu, Finland.

He is currently a post-doctoral researcher and project manager at the Centre for Wireless Communications, University of Oulu, Finland. In 2011–2012, he was a research scientist at the I3S Laboratory and Inria, Sophia Antipolis, France. Also, he was a visiting research fellow at Data61, CSIRO, Australia, Infolabs21, Lancaster University, UK, Computer Science and Engineering, The University of New South Wales, Australia and department of computer science, University of Oxford, UK during 2016–2018. His research interests are SDN, IoT, Block Chain, mobile and virtual network security. He is a member of IEEE and ICT.

Madhusanka is a co-author of over 40 publications including one edited book with Wiley. He is also a management committee member of EU COST Action IC1301, IC1303, CA15107, CA15127 and CA16226 projects. URL: <http://madhusanka.com>



**Ijaz Ahmad**

Centre for Wireless Communications, University of Oulu, Finland.

Ijaz Ahmad received his BSc degree in Computer Systems Engineering from the University of Engineering and Technology (UET), Peshawar, Pakistan. He completed his MSc (Technology) degree of Wireless Communications Engineering with a major in Telecommunications Engineering from the University of Oulu, Finland in 2012. After working as a research assistant in the Centre for Wireless Communications, he started his PhD at the University of Oulu, Finland in 2013.

Ijaz has received several awards including the Nokia Foundation Grant Awards, the Tuano Tonning Foundation Research Grant Awards, and the Achievement award as Inventor from University of Oulu, Finland, for excellent research during his PhD. He has contributed to over 20 publications including high impact factor journal articles, conference papers and book chapters. His research interest includes SDN, SDN-based mobile networks, AI for networking, network security, and network load balancing.

**Ahmed Bux Abro**

VMware Inc. USA.

Ahmed Bux Abro received his Bachelor degree in Computer Science in 1999 from the Shah Abdul Latif University and his Masters degree in Computer Science and Information Technology in 2002 from the University of Sindh with exceptional grades, he is currently a doctorate student at University of Wisconsin-Whitewater, USA. He holds top level professional recognitions and certifications from various industry leaders such as Cisco, IBM, ISC2, Juniper, VMware. A few to name here: CCDE (Cisco Certified Design Expert), CCIE (Cisco Certified Internetwork Expert) Security, VMware Certified Implementation Expert (VCIX), and CISSP (Certified Information Systems Security Professional).

Ahmed is a technologist, strategist and contributor in multiple technology fronts such as software-defined networking, security, cloud and data center. He has 16 years of widespread experience with focus around designing and architecting networks, cloud and virtualized data centers for Fortune 100 customers in diverse markets (North America, EMEA, Asia) and various industry sectors. Currently, he is playing a staff solution architect role at VMware, where part of his job is to help customers transform their legacy business into a digital business and legacy IT into a software-defined enterprise IT using an architecture led approach.

He has contributed in his current and previous role for various new frameworks, architectures and standards around Cloud, Network Function Virtualization, SDN and Security. Ahmed is a chapter co-author of a book on Software Defined Mobile Networks (SDMN), multiple drafts and research papers on the topic of SDN, security and mobility for IEEE and IETF organizations.



Andrei Gurtov

Department of Computer and Information Science,
Linköping University, Sweden.

Andrei Gurtov received his MSc in 2000 and his PhD in 2004 in Computer Science from the University of Helsinki, Finland. He is presently a Professor in Linköping University, Sweden. He is also an adjunct professor at Aalto University, University of Helsinki and University of Oulu. He visited ICSI in Berkeley multiple times. He is an ACM Distinguished Scientist, IEEE ComSoc Distinguished Lecturer and Vice Chair of IEEE Finland section. Andrei co-authored about 200 publications, including 4 books, 5 IETF RFCs, 6 patents, over 50 journal and 100 conference articles.

He supervised 12 PhD theses, serves as an editor of *IEEE Internet of Things*. URL: <http://gurtov.com>



Mika Ylianttila

Centre for Wireless Communications, University of Oulu, Finland.

Mika Ylianttila is a full-time professor at the Centre for Wireless Communications (CWC), at the Faculty of Information Technology and Electrical Engineering (ITEE), University of Oulu, Finland. Previously he was the director of the Center for Internet Excellence (2012–2015) and associate director of the MediaTeam research group (2009–2011), and professor (pro tem) in Information Networks (2005–2010). He is also adjunct professor in Computer Science and Engineering (since 2007). He received his doctoral

degree on Communications Engineering at the University of Oulu in 2005. He co-authored more than 100 international peer-reviewed articles on broadband communications networks and systems, including aspects on network security, mobility management, distributed systems and novel applications. His research interests include 5G applications and services, software-defined networking and edge computing. He is a Senior Member of IEEE, and Editor of Wireless Networks journal. URL: <http://www.ee.oulu.fi/~over/>

About the Contributors

Abdur Rashid Sangi served as Software Engineer/Product Manager in Hisense International Co. Ltd, Qingdao, China and was Assistant Manager, I.T. in the public sector R&D, Karachi, Pakistan. He received a Bachelor's degree in Computer Science and Engineering from Shah A. Latif University, Khairpur, Pakistan and his Master's degree in Communication Networks from Bahria University, Karachi Campus, Pakistan. He was awarded with full-scholarship and finished his PhD in Communication Network Security from Beijing University of Aeronautics and Astronautics (Beihang), China. Currently he is a Senior Engineer in the Huawei R&D center, Beijing, China. His current research interests include IoT security, Contiki, 6LoWPAN and Routing Protocol optimization and design.

Alessandro Piva (SMIEEE) received his MS degree in Electronics Engineering and his PhD in Computer Science and Telecommunications Engineering from the University of Florence, in 1995 and 1999, respectively. He is Associate Professor at the Department of Information Engineering of the University of Florence. His research interests lie in the areas of Information Forensics and Security, including data hiding, signal processing in the encrypted domain, and multimedia forensics, and Image and Video Processing. In the above research topics he has been co-author of more than 40 papers published in international journals and 100 papers published in international conference proceedings. He is currently a Senior Area Editor of the *Journal of Visual Communication and Image Representation*, Associate Editor of *IEEE Transactions on Dependable and Secure Computing* and *EURASIP Journal on Information Security*.

An Braeken obtained her MSc Degree in Mathematics from the University of Ghent in 2002. In 2006, she received her PhD in engineering sciences from the KU Leuven at the research group COSIC (Computer Security and Industrial Cryptography). In 2007, she became professor at Erasmushogeschool Brussels (currently since 2013, Vrije Universiteit Brussel (VUB)) in the Industrial Sciences Department. Her current interests include security protocols for sensor networks.

Anette Alén-Savikko is a postdoctoral researcher at the Faculty of Law, University of Helsinki and University of Lapland. Her research covers new media, digitization, intellectual property (IP) and data protection while she is particularly interested in EU law dimensions thereof. Anette has published and been involved in numerous projects in the fields of media law, IP and data protection law, with her research interests currently

including human centered models of personal data management. In addition, Anette has provided national expertise with regard to her areas of interest. She is currently involved in the Academy-funded project “Information Security of Location Estimation and Navigation Applications (INSURE)”.

Anna Brunstrom received her BSc in Computer Science and Mathematics from Pepperdine University, CA, in 1991, and her MSc and PhD in Computer Science from the College of William & Mary, VA, in 1993 and 1996, respectively. She joined the Department of Computer Science at Karlstad University, Sweden, in 1996, where she is currently a Full Professor and Research Manager for the Distributed Systems and Communications Research Group. Her research interests include transport protocol design, techniques for low latency Internet communication, cross-layer interactions, multi-path communication and performance evaluation of mobile broadband systems. She has led several externally funded research projects within these areas and served as the principal investigator and coordinator from Karlstad University (KaU) in additional national and international projects. She is currently the KaU principal investigator within two EU H2020 projects, the NEAT project aiming to design a new, evolutive API and transport-layer architecture for the Internet, and the MONROE project proposing to design and operate a European transnational open platform for independent, multi-homed, large-scale monitoring and assessment of mobile broadband performance. She is a co-chair of the RTP Medi Congestion Avoidance Techniques (RMCAT) working group within the IETF. She has authored/coauthored 10 book chapters and over 100 international journal and conference papers.

Bengt Sahlin received his MSc in Computer Science from Aalto University (former Helsinki University of Technology (TKK)). At TKK, he has also lectured on Modern Data Communications as well as on DNS and DNS security. He is a Certified Information Systems Security Professional (CISSP). Bengt has worked in the fields of data- and telecommunications for 19 years, mostly with security aspects. In 2000, he joined Ericsson where he has worked on mobile systems security and product security. He was also technical coordinator for Ericsson’s security implementation projects, and is now a manager of a security research group within Ericsson. Bengt Sahlin was 3GPP TSG SA WG3 chairman 2010–2013.

Chang Liu received a BS degree in Electronic Information Engineering from Dalian Maritime University, Dalian, China, in 2012. He is currently pursuing his PhD in the School of Information and Communication Engineering, Dalian University of Technology, China. From 2015 to 2016, he was a visiting scholar in Department of Electrical Engineering and Computer Science at University of Tennessee, Knoxville, USA. His research interests include Spectrum Sensing in Cognitive Radio, Statistical Signal Processing, Random Matrix Theory, Array Signal Processing and 5G networks.

Edgardo Montes de Oca graduated in Engineering in 1985 from Paris XI University, Orsay. He has worked as a research engineer in the Alcatel Corporate Research centre in Marcoussis, France and in Ericsson’s Research centre in Massy, France. In 2004, he founded Montimage, and is currently its CEO. He is the originator and main architect of MMT (Montimage Monitoring Tool). His main interests are future networks (SDN/NFV), network and application monitoring and security, detection and

mitigation of cyber attacks, and building critical systems that require the use of state-of-the-art fault-tolerance, testing and security techniques. He has participated in several EU and French national research projects (e.g. CelticPlus-MEVICO, SIGMONA and SENDATE; H2020-SISSDEN; ANR-DOCTOR). He is a member of NetWorld2020 and has published many papers and book chapters on SDN/SVN, testing, network monitoring, network security and performance.

Elena Simona Lohan received her MSc degree from the Polytechnic University of Bucharest (1997), a DEA degree at Ecole Polytechnique, Paris (1998), and her PhD in Wireless Communications from Tampere University of Technology (TUT) (2003). She is now an Associate Professor at TUT and has been a Visiting Professor at the Universitat Autònoma de Barcelona since 2012. She is the group leader for the signal processing for wireless positioning group at TUT. Her current research interests include wireless location techniques based on Signals of Opportunity, wireless navigation receiver architectures and multipath mitigation, and cognitive, privacy and security aspects related to user positioning. She is currently a working package leader in the Academy-funded project "Information Security of Location Estimation and Navigation Applications (INSURE)".

Fahad T. Bin Muhaya is a full Professor at Management Information Systems (MIS) Department, Business Administration College at King Saud University, Riyadh, Saudi Arabia. He co-founded the Center of Excellence in Information Assurance (CoEIA) and was appointed as a vice director of the Center. In addition, he was appointed as the Director of His Royal Highness Prince Muqrin Chair (PMC) for IT Security, which is the first research Chair in IT Security in the region. Meanwhile, he has served as department Chairman several times and also has served as a Dean. Bin Muhaya is a part-time Information Security Consultant for several government departments and national and international companies. Also he is a member of several scientific societies and founder and board council members of others.

Faisal Rehman is currently working on a research project at the University of Oulu, Finland, which is about the radio wave propagation issues through selective windows. Before working at the University of Oulu, he worked in the telecommunications field for almost 5 years, particularly in RF and optimization of mobile cellular networks. He also worked at transmission and switching departments of a PSTN. He holds Bachelors and Master's degrees in Telecommunications Engineering. His areas of interest include Radio Engineering, antennas, radio channels, and wireless networks.

Hammad Kabir is a doctoral student at the Department of Communication and Networking of Aalto University, Finland. His research focuses on intrusion detection, network security, mobile network, SDN and policy management.

Heidi Kuusniemi is a professor and director at the Department of Navigation and Positioning at the Finnish Geospatial Research Institute (FGI). She is also an Adjunct Professor at Aalto University, Department of Real Estate, Planning and Geoinformatics, and at Tampere University of Technology, Department of Electronics and Communications Engineering, Finland. She is the President of the Nordic Institute of Navigation. She received her MSc and DSc(Tech.) degrees from Tampere University of Technology, Finland, in 2002 and 2005, respectively. In 2003–2004, she was a visiting

researcher at the University of Calgary, Canada, and in the beginning of 2017 a visiting scholar at Stanford University, USA. Kuusniemi's research interests cover various aspects of GNSS and sensor fusion for seamless outdoor/indoor positioning, especially reliability monitoring and information security in positioning. She is the Coordinator of the Academy-funded project "Information Security of Location Estimation and Navigation Applications (INSURE)".

Helena Leppäkoski received her MSc degree in 1990 and her PhD in 2015 from Tampere University of Technology (TUT). She was with Metso Corporation, Helsinki, Finland, from 1990 to 2000 and joined TUT in 2000, where she is currently a Postdoctoral Researcher. Her research topics have varied from satellite positioning to various methods for pedestrian indoor positioning and machine learning for location related context inference. Currently she is working on a project on information security of location estimation and navigation applications. She is currently involved in the Academy-funded project "Information Security of Location Estimation and Navigation Applications (INSURE)".

Hoang Long MAI received a double degree in Engineering in Information Risk's Management and his Master's degree in Information Systems Security from University of Technology of Troyes in 2016. He is currently a PhD student in a CIFRE (Industrial Convention of Formation by Research) contract between Montimage France, University of Technology of Troyes and INRIA Lorraine. His PhD topic is focused on the Autonomous Monitoring and Control of Virtualized Network Functions for security and with an application to Named Data Networking.

Jan Melén is a Research Leader of Network Architecture group at Ericsson Research in Jorvas, Finland. He has over 15 years background on network protocol research and standardisation in the area of IP, mobility, routing and network architectures. Recently, Jan has done research on Internet-of-Things (IoT) and Machine-to-Machine (M2M) related topics on network design and architecture. He has participated and contributed to IETF and 3GPP standardisation and has had active role in Finnish strategic research agendas related to the field of IoT and future networks.

Jari Iinatti (SMIEEE) received his MSc and DTech degrees in electrical engineering from the University of Oulu, Finland, in 1989 and 1997, respectively. During 1989–1997, he was a Research Scientist at the Telecommunication Laboratory at the University of Oulu. During 1997–2002, he was an acting professor of Digital Transmission Techniques, and since 2002, Professor of Telecommunication Theory at Centre for Wireless Communications at the University of Oulu. He is also an IAS Visiting Professor at Yokohama National University, Yokohama, Japan. His research interests include future wireless communications systems, transceiver algorithms, wireless body area networks (WBANs) and medical ICT. He published more than 200 journal and conference papers and holds 6 patents. He supervised 13 Doctoral Theses and 64 Master's Theses. He has been a TPC member at about 30 conferences, and he was a TPC chair in the ISMICT2007, TPC co-chair in PIMRC2006, BodyNets2012 and PIMRC 2014, general co-chair in the ISMICT2011, 2014–2017.

Javid Taheri received his Bachelor and Masters degrees in Electrical Engineering from the Sharif University of Technology in 1998 and 2000, respectively. He received his PhD in the field of Mobile Computing from the School of Information Technologies at the University of Sydney, Australia. He is currently working as Associate Professor in the Department of Computer Science in Karlstad University, Sweden.

Jesus Llorente Santos is a doctoral student at the Department of Communication and Networking of Aalto University, Finland. His research focuses on mobile networks, software defined networking (SDN) and future internet architectures.

João André obtained his Diploma in Civil Engineering and his MSc in Structural Engineering from the Instituto Superior Técnico (part of University of Lisbon), and his PhD in Structural Engineering from Oxford Brookes University. He worked as a Professor for two years in the Universidade Lusófona, teaching courses on steel and reinforced concrete structures. He has been working in the Structures Department of the Portuguese National Laboratory Civil Engineering (LNEC) since 2005, where he currently serves as a Postdoctoral Research Fellow. He has published over 30 papers over a wide range of subjects, ranging from numerical and experimental analyses, robustness and risk analyses. He was appointed a member of the project team responsible for defining the “Robustness Framework” for the revision of the European Structural Eurocodes and he is the National Expert of WG6 of CEN/TC250. He is currently working in two European COST Action research projects concerning communication and bridge structures.

Jude Okwuibe received his BSc in Telecommunications and Wireless Technologies from the American University of Nigeria, Yola, in 2011. After graduation, he worked as a recruitment specialist with the American University of Nigeria for about a year before going for one year’s National Service where he served as an assistant instructor teaching computer science. In 2015, Okwuibe received his Master’s degree in Wireless Communications Engineering from the University of Oulu, Finland. Okwuibe is currently doing a doctoral program in Communications Engineering at the University of Oulu Graduate School (UniOGS), Finland. His research interests are 5G and future networks, IoT, SDN, Network security, and biometric verifications.

Jukka Salo received his MSc in Electrical Engineering at the University of Oulu in 1976, and joined Nokia Corporation in 1977, where he since then until the retirement in late 2016 held different positions in the research and product development of Nokia’s network systems. In 2008–2012, Jukka Salo was a Steering Board member in a Finnish Strategic Centre for Science, Technology and Innovation in the Field of ICT (TIVIT). In 2008–2016, he was Nokia’s representative in the Celtic (EUREKA cluster) Core Group and the Vice-chairman of Celtic. Celtic is an industry-driven European research initiative to define, perform and finance through public and private funding common research projects in the area of telecommunications. Jukka Salo was also involved in Policies, Governance and Regulation related research work in several international projects, including 4WARD (EU FP7), SAIL (EU FP7), MEVICO (EUREKA Celtic) and SIGMONA (EUREKA Celtic).

Julius Francis Gomes is pursuing his PhD in International Business from the University of Oulu. He currently works at the Oulu Business School as a Doctoral Student to research the futuristic business models for entities which will be involved in the tech-oriented business arena. His research focuses on using business models as a means to look into future industries. He is interested to research business ecosystems in different contexts, such as cyber security, healthcare, future's network, etc. with a business model perspective. He received his MSc (2015) in International Business from the University of Oulu. Prior to that, he acquired an MBA in 2011, specializing in managing information systems in business applications. Francis Gomes has enjoyed three years in a top tier bank in Bangladesh as a channel innovator.

Karl-Johan Grinnemo received his MSc in Computer Science and Engineering from the Linköping Institute of Technology, Sweden, in 1994. In 2006, he received his PhD in Computer Science from Karlstad University, Sweden. He worked almost 15 years as an engineer in the telecom industry; first at Ericsson and then as a consultant at Tieto. A large part of his work has been related to Ericsson's signaling system in the mobile core and radio access network. From the Fall of 2009 until the Fall of 2010, he was on leave from Tieto and worked as acting Associate Professor at the School of Information and Communication Technology, KTH Royal Institute of Technology. Between the Fall of 2010 and the Fall of 2014, he was an Associate Senior Lecturer at Karlstad University, and became a Senior Lecturer in the Fall of 2014. His research primarily targets application- and transport-level service quality. He has authored and co-authored around 40 conference and journal papers, and is a Senior member of IEEE.

Kimmo Järvinen received his MSc (Tech) degree in 2003 and the DSc (Tech.) degree in 2008 from Helsinki University of Technology (TKK), Finland. He was with the Signal Processing Laboratory at TKK from 2002 to 2008. In 2008–2013 and again in 2015–2016, he was a postdoctoral researcher in the Department of (Information and) Computer Science, Aalto University, Finland. In 2014/2015, he was with the COSIC group of KU Leuven ESAT, Belgium. Since November 2016, he is a senior researcher in the Department of Computer Science, University of Helsinki, Finland. His research interests lie in the domains of security and cryptography, especially in developing efficient and secure implementations of cryptosystems. He has authored more than 40 peer-reviewed scientific publications. He is currently a working package leader in the Academy-funded project “Information Security of Location Estimation and Navigation Applications (INSURE)”.

Lauri Isotalo received his MSc from Helsinki University of Technology (currently Aalto University) in 1992. He also has a postgraduate Diploma in Business Administration. At first, Lauri worked in Nokia Corporation in the Mobile Technology & System Marketing unit, specializing in Intelligent Networks. In 1992, he joined the Elisa Corporation, where he has held several managerial positions in value-added services business, system and process security and mobile network development. Since 2005, Lauri has also led Elisa SME teams in various international collaboration projects and acquired a deep knowledge of the cyber security of legacy telecommunication networks, in core, access networks, user terminals and modern virtualized data center IT platforms/cloud systems. From 2014, Lauri has headed SDN&NFV development in Elisa.

Liang Chen is a Senior Research Scientist in the Department of Navigation and Positioning at the Finnish Geospatial Research Institute (FGI), Finland. Before he joined FGI, he worked in the Department of Mathematics at Tampere University of Technology, Finland from 2009 to 2011. He received his PhD in Signal and Information Processing from Southeast University, China, in 2009. His research interests include statistical signal processing for positioning, wireless positioning using signals of opportunity and sensor fusion algorithm for indoor positioning. He is currently involved in the Academy-funded project “Information Security of Location Estimation and Navigation Applications (INSURE)”.

Lorenzo Mucchi (SMIEEE) received his D.Eng. degree (Laurea) in Telecommunications Engineering from the University of Florence, Italy in 1998 and his PhD in Telecommunications and Information Society in 2001. Since 2001, he has been with the Department of Information Engineering of the University of Florence as a Research Scientist. He is a Professor of Information Technologies at the University of Florence since 2008. His main research areas include theoretical modeling, algorithm design and real measurements, mainly focused on the fields of physical-layer security, visible light communications, spread spectrum techniques, localization, and interference management. Dr Mucchi is an associate editor (2016) of *IEEE Communication Letters*. He is also a member of the European Telecommunications Standard Institute (ETSI) Smart Body Area Network (SmartBAN) group (2013) and team leader (2016) of the special task force 511 “SmartBAN Performance and Coexistence Verification”. More details: <http://www.lorenzomucchi.info/>

Marika Iivari is a postdoctoral researcher at the Martti Ahtisaari Institute within the Oulu Business School. She defended her doctoral dissertation on business models in ecosystemic contexts. She received her MSc in International Business from the Ulster University, Northern Ireland. Her research interests are in the areas of open innovation, business models and strategy in the context of innovation ecosystems and smart cities, digital and ICT business ecosystems. She has been involved in several research projects around 5G and the Internet of Things, most recently in the healthcare sector. She is also an active member of the Business Model Community, the Open Innovation Community and the Society for Collaborative Networks.

Matti Hämäläinen (SMIEEE) received his MSc and DSc degrees in 1994 and 2006, respectively, from the University of Oulu, Finland. He contributed to more than 160 international scientific journal and conference publications. He is a co-author of “Wireless UWB Body Area Networks – Using the IEEE802.15.4-2011”, Academic Press and co-editor of “UWB: Theory and Applications”, Wiley & Sons. He holds one patent. Currently he is a University Researcher and Adjunct Professor at Centre for Wireless Communications, University of Oulu, Finland and IAS Visiting Professor at Yokohama National University, Yokohama, Japan. He is a member of External Advisory Board of Macquarie University’s WiMed Research Centre, Australia and International Steering Committee of International Symposium on Medical ICT. Dr Hämäläinen is also a contributor of ETSI TC SmartBAN. His research interests are in UWB systems, wireless body area networks and medical ICT.

Mehrnoosh Monshizadeh is finalizing her PhD in Telecommunication Networking at the Electrical School of Aalto University, Finland. She is working as a research security specialist at Nokia Bell Labs, Finland. Her research interests include cloud security, mobile network security, IoT security and data analytics.

Mikel Uriarte Itzazelaia received his BSc and MSc degrees in Telecommunication Engineering in 1998 from the University of the Basque Country (UPV/EHU). He spent one year in public R&D in Telecommunications enterprise (currently Tecnalia). From 1998 to the present, he worked at Nextel S.A., a telecommunications enterprise providing ICT engineering and consulting services. From 2001 to 2006, he worked as ICT director and as an information security lead auditor, subsequently becoming the head of the research and development unit. His research interests include ICT interoperability, resilience, performance and security in several areas such as identity and access control, networking, wireless sensing and cloud computing.

Mohammed Alkatheiri is an assistant professor in the Department of Computer Science, College of Computing and Information Technology, University of Jeddah, Saudi Arabia. Currently, he is a chair of the Information Technology Department. His current research interest focuses on the area of information security. Previously, he worked as a researcher in the Center of Excellence in Information Assurance at the King Saud University, Riyadh, Saudi Arabia. His research interest focusing on security and privacy related issues of information sharing, identification, and authentication. Also, he served as consultant for national projects and joined Prince Muqrin Chair for Information Security Technology (PMC) along with government departments on National Information Security Strategy project as a security consultant.

Oscar López Perez received his BSc in Telecommunication Engineering from the Polytechnic University of Catalonia in 1998. After finishing his studies, he worked in a technical school teaching different IT subjects in an Associate degree. In 2000, he joined Nextel S.A, covering different stages as technical, auditor and later providing consultancy services in ICT and cyber security. Since 2008, he has been working as a R&D researcher, participating in national and European research projects. His research work has been related to the evaluation of the operational security assurance, and in other initiatives such as enforcing security policies and in the result of an adequate security monitoring in different application environments.

Päivi Korpisaari is a professor in Communication Law at the Faculty of Law, University of Helsinki. She completed her Master of Laws in 1993, defended her Licentiate in 2000 and her Doctor of Law degree in 2007 from the University of Helsinki. She was appointed communications law professor at the University of Helsinki in 2014. Her research interests are in personal data protection law, freedom of expression, privacy, media law and communications law. She is currently a working package leader in the Academy-funded project “Information Security of Location Estimation and Navigation Applications (INSURE)” and TEKES-funded project MyGeoTrust.

Petri Ahokangas received his MSc (1992) and DSc (1998) degrees from the University of Vaasa, Finland. He is currently Adjunct Professor (International software entrepreneurship) and Senior research fellow at Martti Ahtisaari Institute, Oulu Business

School, University of Oulu, Finland. His research interests are in how innovation and technological change affect international business creation, transformation, and strategies in highly technology-intensive or software-intensive business domains. He has over 100 publications in scientific journals, books, conference proceedings, and other reports. He is actively working in several ICT-focused research consortia leading the business-related research streams.

Raimo Kantola is a Doctor of Science in Technology. He is a full, tenured professor of networking technology at the Department of Communications and Networking of Aalto University. After 15 years in Nokia Networks in positions in R&D and marketing, he joined Helsinki University of Technology as a professor in 1996 and was tenured in 2006. Professor Kantola's recent research is in trust in networks and customer edge switching. He has held many positions of trust at Helsinki University of Technology and Aalto University.

Rui Travanca has a Diploma in Civil Engineering and an MSc in Structural Engineering. Rui has a strong background within the telecommunication industry, which includes more than ten years working as a Civil Engineer and an Independent Engineering Consultant for major telecommunication operator companies. Rui is deeply involved in research, in structural engineering subjects, and has conducted several research works in the field of the structural behaviour of communication structures, mainly using structural health monitoring techniques. Main fields of interest/research are wind-sensitive structures, earthquake engineering, structural behaviour, structural simulation, numerical model calibration, dynamic analysis, structural health monitoring, optical sensors and wind tunnel testing.

Sadiqur Rahaman is completing his Master's degree in Wireless Communication Engineering in University of Oulu, Finland. Before that, he had taken his bachelor's degree in Electrical and Electronic Engineering and an MBA in Management Information Systems. He has published a number of international conference papers. His research interest lies in the field of antenna and radio engineering. He is currently working on a passive repeater for WLAN operation using various types of antenna and co-axial cable.

Satish Anamalamudi received his BEng degree in Computer Science and Engineering from Jawaharlal Nehru Technological University, Hyderabad, India, MTech in Network and Internet Engineering from Karunya University, Coimbatore, India and his PhD in Communication and Information Systems from Dalian University of Technology, Dalian, China. He worked as a Research Engineer in Beijing Huawei Technologies, Beijing, China, from November 2015 to August 2016. He is currently working as Assistant Professor in the Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian, China. His research interests include common-control-channel design for MAC and routing protocols in cognitive radio ad hoc networks, MAC and routing protocol design of IoT and 5G networks.

Shahriar Shahabuddin received his BSc from the University of Dhaka, Bangladesh and his MSc from the University of Oulu, Finland in 2009 and 2012 respectively. Afterwards, he started his PhD under the supervision of Professor Markku Juntti in University of Oulu, Finland. During the spring of 2015, he worked at the Computer

Systems Laboratory of Cornell University, USA, with Professor Christoph Studer. Shahriar received a distinction in his MSc and the best Master's thesis award of the Department of Communications Engineering, University of Oulu in 2012. He is the recipient of several scholarships and grants, such as Nokia Foundation Scholarship, University of Oulu Scholarship Foundation Grant, and UniOGS travel grant.

Simone Soderi (SMIEEE) received his MSc degree in 2002 from the University of Florence, Italy and his DSc degree in 2016, from the University of Oulu, Finland. Dr Soderi has more than 14 years' experience in embedded systems and safety related architectures. His skills range from electronic and electromagnetic compatibility to software engineering. During 2011–2014, he was a member of the Steering Committee of a joint research project between General Electric, Florence, Italy (GE) and the Centre for Wireless Communications, University of Oulu, Finland. During 2011–2015, he contributed in ETSI for ultra-wideband (UWB) devices in road and rail vehicles. Currently he is Cybersecurity Manager at Alstom, Florence, Italy. His research topics include UWB, electromagnetic compatibility, cyber-security for critical infrastructure systems and physical layer security. He has been TPC member of several conferences and served as reviewer of IEEE Transaction on Intelligent Transport Systems (ITS). Dr Soderi has published journal and conference papers, and various book chapters. He holds five patents regarding wireless communications and positioning.

Tanesh Kumar received his MSc degree in Computer Science from the South Asian University, New Delhi, India in 2014. Prior to that, he did his bachelors in Computer Engineering from the National University of Sciences and Technology (E&ME), Rawalpindi, Pakistan in 2012. Currently he is a doctoral student at the University of Oulu and a research scientist in the Centre for Wireless Communications (CWC), Oulu, Finland. His research interest includes IoT Security, Privacy in Hyperconnected Environment, Biometric Authentication and 5G security.

Van-Giang Nguyen received his Bachelor's degree in Electronics and Telecommunication Engineering from Hanoi University of Science and Technology, Vietnam in 2012 and his Master's degree in Information and Telecommunication Engineering from Soongsil University, South Korea in 2015. From 2013 to 2015, he worked as a research assistant at the Distributed Computing Network (DCN) laboratory, Soongsil University. Since 2015, he has been working towards his PhD degree in Computer Networks and Telecommunications at the Department of Computer Science and is working as a research assistant at the Distributed System and Communications (DISCO) research group, Karlstad University, Sweden. His current research interests include SDN (software defined networking), NFV (network function virtualization), future mobile packet core network, open source networking and 5G networking. He is a student member of IEEE SDN.

Vikramajeet Khatri graduated with an MSc IT from the Tampere University of Technology, Finland. He is working as a research security specialist at Nokia Bell Labs, Finland. His research interests include intrusion detection, malware detection, IoT security and cloud security.

Wenjing Chu is a Distinguished Engineer and Senior Director of Open Source and Standards at Huawei in Santa Clara, CA. Prior to Huawei, he was a Chief Architect of NFV in VMWare, Inc. and a Distinguished Engineer in Dell Research, Santa Clara, CA, driving its NFV strategy and advanced research in High Velocity Cloud. His work at Dell focused on high performance networking and real-time machine learning systems for the cloud. He is a Director of the Board for Open Platform for NFV (OPNFV) and was previously the Chair of the Compliance and Certification Committee and a member of the Technical Steering Committee. His long career in technology companies includes leading roles in startup multimedia network vendor Sentient Networks Inc. and enterprise Wi-Fi pioneer Airespace, Inc. Wenjing received his BSc in Computer Science from Peking University, China and received his MSc in Computer Science from the University of British Columbia, Canada.

Zaheer Khan received his PhD in Electrical Engineering from the University of Oulu, Finland, and his MSc degree in Electrical Engineering from the University College Boras, Sweden, in 2011 and 2007, respectively. Currently, he has a Lecturer/Tenure track position at the University of Liverpool, United Kingdom. He worked as a research fellow/principal investigator at the University of Oulu. He was the recipient of the Marie Curie fellowship for 2007–2008. His research interests include application of game theory to model distributed wireless networks, prototyping access protocols for wireless networks, IoT location tracking systems, cognitive and cooperative communications, and wireless signal design.

Foreword

5G cellular networks promise not only an enhancement of radio access technology but also to complete the trajectory to connecting billions of people and things, whether on motion or attached to an infrastructure. By reducing the cost and efforts to connect people and things, it will not only help accelerate economic growth across various industries and the public sector, but will also provide a platform based on cloud computing and IoT (Internet of Things) for many critical infrastructures that offer important utility, transportation and public safety services. It is currently, and will be for the next decade, the prime focus of research and development activities across multiple countries and continents, and the outcome sought will go beyond basic user connectivity services, also addressing opportunities to simplify networks and the deployment of new services.

As 5G networks become pervasive and foundational components of personal, public and enterprise systems, possibly replacing dedicated and isolated networks. Maintaining the confidentiality, integrity and availability of these networks will be among the key design and implementation challenges. Universal connectivity is attractive to both the intended beneficiaries of these networks likewise to the bad actors. They can wreak havoc from afar, across different industries and causing financial, privacy, safety and national security harms.

This book is one of the first attempts to comprehensively address the key security areas and domains for 5G networks, starting from a 5G security landscape overview and physical infrastructure security to an in-depth discussion around security mechanisms for different components of 5G mobile networks. It also provides important insights into the current and future threats to mobile networks and mapping those to the various threat vectors for different mobile generations, including 5G, by using a detailed threat analysis approach. Readers will find an opportunity to explore the evolved security model and lifecycle functions for 5G. This book has taken a fresh perspective on addressing security and privacy for new areas evolving with 5G, including Device to Device (D2D) connectivity, cloud services, SDMN (Software Defined Mobile Networks), NFV (Network Function Virtualization) and IoT (Internet of Things).

The book will be helpful to a range of 5G stakeholders: researchers looking for challenging new problems, mobile network operators (MNOs) and virtual network operators (MVNOs), seeking to plan for the new threat environment, owners of infrastructure investigating how 5G can improve their operations, telecom equipment vendors, and standardization bodies working on network and IoT standards.

Henning Schulzrinne
Chief Technology Officer
Federal Communications Commission,
The United States of America

Preface

The emergence of smartphones and tablets coupled with broadband wireless connectivity has changed our lives. More demands on high throughput, low latency, high-speed mobility and new services drive the development of 5G. The first commercial networks of 5G are expected for deployment by 2020, three years ahead of writing this book. However, initial proof-of-concept deployments are announced already for 2018. While multiple books already exist on 5G, this is the first book, to our knowledge that focuses on security aspects of future 5G ecosystem.

The book provides a reference material to a comprehensive study of 5G security. It offers an insight into the current and future threats to mobile networks and mechanisms to protect it. It covers the critical lifecycle functions and stages of 5G security, and how to build an effective security architecture for 5G based mobile networks. It addresses mobile network security based on Network-centricity, Device-centricity, Information-centricity and most importantly, People-Centricity views.

This book offers security considerations for all relative stakeholders of mobile networks, such as mobile network operators (MNOs), mobile virtual network operators (MVNOs), mobile users, wireless users, Internet-of-Things (IoT) and cybersecurity experts, security researchers and engineers.

5G Mobile Networks

5G networks are not only expected to be faster, but provide a backbone for many new services for Networked Society, such as IoT and the Industrial Internet. Those services will provide connectivity for autonomous cars and Unmanned Aerial Vehicles (UAVs), remote health monitoring through body-attached sensors, smart logistics through item tracking, remote diagnostics and preventive maintenance of equipment. Most services will be integrated with cloud computing and novel concepts such as mobile edge computing, which requires smooth and transparent communications between user devices, data centers and operator's networks. New classes of Quality-of-Service (QoS), such as low-latency ultra-reliable communication as well as energy-efficient sensor connectivity, will hopefully be supported by 5G.

New radio bands above 20GHz are being allocated for 5G. Since the current LTE systems already approach the theoretical limits of spectrum efficiency use, higher rates in the 5G can only be achieved by using millimeter-wavelength bands with challenging propagation properties in combination with very small cells. This is also needed to

achieve extremely high density of users per geographical area. Providing radio communication at high speeds and low power, as well as seamless roaming and network mobility, remain major challenges for 5G. Physical layer security on the radio level may prove to be an important challenge for 5G technology.

5G is presently under development by telecommunication vendors, EU projects (5G-ENSURE) and frameworks, and the 5G Infrastructure Public Private Partnership (5G PPP). Many of the vendors have provided their vision of 5G services and security models in White Papers. However, the standardization process of 5G is just starting within 3GPP, although other standardization bodies, such as the Internet Engineering Task Force (IETF), are continuously developing new secure protocols and architectures to be utilized in 5G. It is important that 5G networks are securely designed and standardized from the beginning, rather than adding security as an afterthought.

Although security models of 3G and 4G networks based on Universal SIM cards worked well, 5G security cannot be a carbon copy of existing designs, due to new requirements. Initially, the main motivation for security in cellular networks was the right functioning of the billing system, followed by encryption of the radio interface. Location and identity privacy of the user were also supported, followed by two-way authentication in 3G to prevent fake base stations. 4G added state-of-the-art cryptographic protocols and protection of physical tampering with the base stations, which could be installed on user premises. While all those security properties are still valid, 5G will face additional challenges due to increased user privacy concerns, new trust and service models and requirements to support IoT and mission-critical applications.

The Need for Security

Phone hacking was first spotted somewhere between the 1960s and 1970s, when phreakers demonstrated their skills to manipulate the functions of a telephone network. Methods to attack telecommunication systems have evolved since then and have changed shape from war dialers to viruses to worms to modern-day advance persistent threats. Tools to protect our telecommunication systems have also evolved from physical access control to antivirus to modern application and context aware firewalls.

Increased use of smartphones for data services and applications has exposed these devices to the same security threats that were once known and dedicated to personal computers (PCs). Mobile devices have replaced legacy system and have changed our ways to learn, work, entertain, shop and travel. Bring Your Own Device (BYOD) and cloud technologies have further diminished the enterprise boundaries and often challenged security experts to work out of the box strategies.

Motivations for attacking networks have also changed from fun-loving immature script kiddies to organized cybercrime rings and hacktivists with clear political and financial objectives. In this age of digitalization, whereafter connecting humans using Internet and mobile, we are talking about connecting things and machines. The mobile has not yet completely replaced the personal computer but has become an ideal place where personal information can be found for nefarious use. Therefore, security needs to be architected to not only protect from the current threats but to address the increasing and evolving threat landscape. Adequate security should include threat intelligence, visibility and real time protection.

On the other hand, today's networks host various values – examples include revenue streams and brand reputation. The accessibility of these values via the Internet has already attracted hacktivists, underground economies, cybercrime and cyber-terrorists. The values hosted in, and generated by, the 5G system are estimated to be even higher, and the assets (hardware, software, information and revenue streams) will be even more attractive for different types of attacks. Furthermore, considering the possible consequences of an attack, the damage may not be limited to a business or reputation; it could even have a severe impact on public safety.

This leads to a need to strengthen certain security functional areas. Attack resistance needs to be a design consideration when defining new 5G protocols. Security and privacy are cornerstones for 5G to become a platform for the Networked Society. Cellular systems pioneered the creation of security solutions for public communication, providing a vast, trustworthy ecosystem – 5G will drive new requirements due to new business and trust models, new service delivery models, an evolved threat landscape and an increased concern for privacy.

5G is going to offer similar impact to communication as once “fiber” technology did; it has the potential to transform the mobility concept. Applications for 5G are beyond the traditional mobile connectivity needs to new public communication, IoT, smart world based out of smart cities, smart transportation and more. One of the major challenges for 5G adoption is security related challenges.

To the best of our knowledge, no book is published yet that addresses the 5G security, comprehensively, and very little is written on this topic. Although the security is the mandatory requirement of 5G networks, many of the 5G security related issues are still under development. However, the rapid adaptation of 5G network will soon raise the requirement of a comprehensive handbook of 5G security.

5G Security Standardization

At the time of writing, 5G standardization has not yet started as the system architecture is still in research phase. Despite its name, the Third Generation Partnership Project (3GPP) continues its work for defining also 5G specifications. In February 2017, 3GPP published “Service Requirements for the 5G system” (TS22.261) that defines performance targets in various scenarios such as indoor, urban, rural and different applications (intelligent transport, remote monitoring, etc.). 3GPP plans to publish 5G Phase 1 specifications in 2018 as Release 15 and Phase 2 in 2020 as Release 16.

Since 5G is expected to be completely converged with Internet protocols, the standards produced by the Internet Engineering Task Force (IETF) in Request for Comments (RFCs) are expected to play a key role. The relevant Working Groups are, for example, IP Wireless Access in Vehicular Environments (IPWAVE) WG and Host Identity Protocol (HIP) WG on secure mobility protocols.

If 5G networks will serve safety-crucial applications as envisaged, the ISO (International Organization for Standardization) will introduce standards such as Common Criteria (ISO 15408) will apply. For instance, for car connectivity, a specific standard is ISO 26262, which covers car safety requirements. For tele-health, EU and USA-specific standards, such as HIPAA (Health Insurance Portability and Accountability

Act) and internationally ISO 27799. For smart cities and smart grids, standards of IEC (International Electrotechnical Commission) and compliance to, for example, North American Electric Reliability Corporation (NERC) will be needed.

ETSI (European Telecommunications Standards Institute) was the creator of GSM standard and key contributor of W-CDMA as 3G standard within 3GPP. As a co-founder of 3GPP, ETSI is actively involved in developing 5G through organizing such events as “ETSI Summit on 5G Network Infrastructure”, which focused on 5G standardization in 2017. ETSI identified priority applications for 5G as mobile broadband evolution, massive M2M communication, and ultra-reliable low latency communication. ETSI is also a known contributor to the Network Functions Virtualization Industry Specification Group (ISG) and is currently reforming a group focusing on 5G security.

ITU (International Telecommunication Union) receives input from regional organizations such as ETSI in Europe and ARIB in Japan and develops recommendations for standards-defining bodies. ITU Telecommunication Standardization Sector (ITU-T) created a Focus Group on International Mobile Telecommunications (IMT-2020), which operated in 2015-2016 and analyzed requirements and framework for the 5G ecosystem. ITU Study Group 17 (SG 17) focuses entirely on security aspects of telecommunication.

Several other relevant Standardization Bodies include IEEE 802, TCG and ONF. Interoperability and mobility with third-party networks such as WiFi involves standards from the IEEE (Institute of Electrical and Electronics Engineers) such as 802.11. At Trusted Computing Group (TCG), the Mobile Platform Work Group (MPWG) develops use cases, frameworks and analyses of 5G security. Open Networking Foundation (ONF) promotes the use of software-defined networking protocols and network operating systems. Its specifications, including OpenFlow, could become a part of 5G core architecture and therefore are also important from the security viewpoint.

Intended Audience

This book will be of key interest for multiple groups of researchers, engineers and business persons working on 5G development and deployment:

- *Mobile Network Operators (MNOs)*: as they will be looking to adopt 5G technology to offer new and state-of-the-art secure services to their customers. This book will offer the required guidelines, methods, tools and mechanisms to secure their network while embracing for 5G.
- *Mobile Virtual Network Operators (MVNOs)*: would like to equally reach the large customer base that is going to switch to 5G networks. Security is the key requirement while connecting MVNOs with the core networks of large operators.
- *Telecommunication researchers*: 5G security is one of the key areas of interest for telecommunications researchers, as security challenges outpace the traditional tools available on the market. This book will offer a single source of all the security related topics for 5G researchers and provide leads for basics of 5G security.
- *Academics*: Mobile network security has already been an area of research and study for major educational institutions across the world. With 5G evolution as the future of mobile networks, there is no such other reference book available that academics can use for teaching this critical area of interest.

- *Technology Architects and Standardization Bodies:* 5G is going to cross the traditional mobility borders and is going to have an equal impact on enterprises and organizations who are planning to transform into digital businesses. It would be critical for architects to start aligning their technology and security architectures to the future needs of 5G standards. This book offers resources to design and build a security architecture and maintain it.
- *IoT and Industrial Internet experts:* Internet of things is going to change the way industrial networks are built, and 5G is going to provide the underlying platform for IoT networks. Security has remained the top priority for industries due to criticality and sensitivity of the data and information flows in their networks. Advance knowledge of 5G security principles, components and domains is going to help industries lay a foundation of IoT security. This book will provide the guidelines and best practices for 5G-based IoT security.

Book Organization

The book is divided into four parts covering various aspects of 5G security ecosystem: Security Overview, Network Security, Device and User Security, and Cloud and Virtual Network Security.

The first part provides an introduction to 5G and history of preceding systems, an overview of 5G security architecture, and general aspects of telecommunication security. The first chapter describes the evolution of cellular systems. For each generation, from 1G to current 4G, its architecture and security mechanisms are presented. The second chapter focuses on 5G mobile networks from the viewpoint of requirements and enabling technologies. The main 5G system components (radio, core, end-to-end), standardization and research activities are surveyed. The third chapter on mobile network security landscape describes attacks possible in the existing and previous generation mobile communication systems. Severity and estimated frequency of threats are analyzed, concluded by the evolved 5G security model. The fourth chapter considers secure 5G software-defined network architecture. The fifth chapter concludes Part I with an overview of cyber-security preparedness framework.

Part II takes an in-depth look at security of core and radio interfaces. Chapter 6 introduces radio signal watermarking as a way to achieve security at the physical layer. The application of this concept to 5G architecture is proposed. Chapter 7 treats interoperability between 5G and Wireless LANs (WLANs) from the security viewpoint. This chapter compares possible attacks in 5G and WLANs and proposes a common interoperability architecture. Chapter 8 focuses on safety of physical infrastructure in 5G. Structural resilience of mast poles to natural disasters and deliberate attacks by humans are described. Chapter 9 is dedicated to Customer Edge Switching (CES). It is a security framework for 5G, which extends functionality of Network Address Translation (NAT) at the edges. Finally, Chapter 10 introduces a Software Defined Security Monitoring for 5G Networks. The use of novel Software Defined Networking (SDN) and Network Function Virtualization (NFV) concepts in 5G monitoring systems can address the classical weaknesses in legacy monitoring systems.

Part III considers security outside of the operator's part of the 5G network; namely, on user equipment such as smartphones and embedded modems and the users themselves.

The main topics concern security of the Internet of Things (IoT), privacy and authentication of users, and positioning security. Chapter 11 describes security of Internet of Things (IoT) when connected over 5G. A special consideration is given to wireless connectivity of robots, such as Unmanned Aerial Vehicles (UAVs). Chapter 12 handles user privacy in 5G, including location, identity and data privacy. Trust models and Identity management are considered. Chapter 13 performs a deeper investigation of secure device positioning in 5G. Outdoor and indoor positioning mechanisms are compared in the context of 5G, and their security threats and avoidance methods are analyzed.

Part IV is dedicated to cloud technologies and network virtualization security. Chapter 14 describes the roles and security models of Mobile Virtual Network Operators (MVNO) in 5G. Possible attacks on hypervisors, virtual machines and software-defined components to compromise availability, integrity and Authentication, Authorization, Accounting (AAA) are considered. Chapter 15 handles security of Network Function Virtualization (NFV) and related services in 5G networks. NFV driving forces, secure lifecycle and multi-tenancy issues, policy and machine learning in NFV are discussed. Chapter 16 introduces a concept of Mobile Edge Computing (MEC) in 5G. Various security challenges for MEC, possible attacks and secure architectures are described. The final chapter, Chapter 17 takes a look the regulatory aspects of privacy and security in 5G. The legal framework, relevance analysis, and technology implications can be found there.

Acknowledgements

This book focuses on 5G Security that is developed as a joint effort of many contributors. First of all, we would like to give our thanks to all of the chapter authors for doing a great job!

This book would not have been possible without the help of so many people. The initial idea for this book originated during our work in The Naked Approach (Nordic perspective to gadget-free hyperconnected environments), Towards Digital Paradise (TDP) and SECUREConnect (Secure Connectivity of Future Cyber-Physical Systems) projects. We thank the Finnish Funding Agency for Technology and Innovation (Tekes), Academy of Finland and Center for Industrial Information Technology (CENIIT) that funded the above research projects. We would also like to acknowledge all the partners in The Naked Approach, Towards Digital Paradise and SECUREConnect projects.

We also thank all the reviewers for helping us to select suitable chapters for this book. Moreover, we thank anonymous reviewers who have evaluated the proposal and given us plenty of useful suggestions for improving it. Professor Henning Schulzrinne, from the US Federal Communications Commission, wrote a nice foreword for this book and we really appreciate his efforts. We thank Sandra Grayson from John Wiley & Sons for her help and support in getting the book published.

Also, the authors are grateful to the Centre for Wireless Communications (CWC) and University of Oulu for hosting the 5G-related research projects, which helped us to gain the fundamental knowledge for this book. Last but not the least, we would like to thank our core and extended families and our friends for their love and support in getting the book completed.

Madhusanka Liyanage
Ijaz Ahmad
Ahmed Bux Abro
Andrei Gursov
Mika Ylianttila

Part I

5G Security Overview

1

Evolution of Cellular Systems

*Shahriar Shahabuddin¹, Sadiqur Rahaman¹, Faisal Rehman¹,
Ijaz Ahmad¹, and Zaheer Khan²*

¹ University of Oulu, Finland

² University of Liverpool, UK

1.1 Introduction

Wireless communication technologies are essential parts of our lives. From WiFi home networks to sophisticated machine-to-machine communication in the robotics industry, we live in a world of wireless connectivity and it is impossible to imagine a single day without using any wireless devices. The blessings of cellular technologies provided us with a great deal of mobility and thus made it possible to listen to the radio while travelling in a car or on the beach. The cellular devices are also convenient in that we no longer have to worry about the size of the cables to connect to the networks. We are now living in a world where conferences for business meetings, distance and online courses from universities, and medical help over long distances are considered as part and parcel of our daily lives. We have greater access to information than ever before and it is all possible due to the advancements and inventions in cellular communication.

The number of cellular users increased dramatically over the last decade compared to the other technologies and are still increasing. We can see from Figure 1.1 that the fixed broadband or fixed wired subscription did not increase that much in a last decade, while the mobile cellular subscriptions are increasing day by day. With the advent of sophisticated technologies, such as tactile computing, autonomous vehicles, wireless charging, smart living, etc., we can only envision how the use of cellular technologies will grow in the future.

This chapter is dedicated towards the evolution of cellular communication. In that respect, we start by discussing the initial developments and history of cellular systems. We subsequently go through the different generations of cellular systems and have a brief discussion about them. As the topic is broad, we try to confine ourselves to the basic information related to the radio interfaces and network architecture of different generations. We align the chapter with the focus of the book by discussing the evolution of security measurements during each generation.

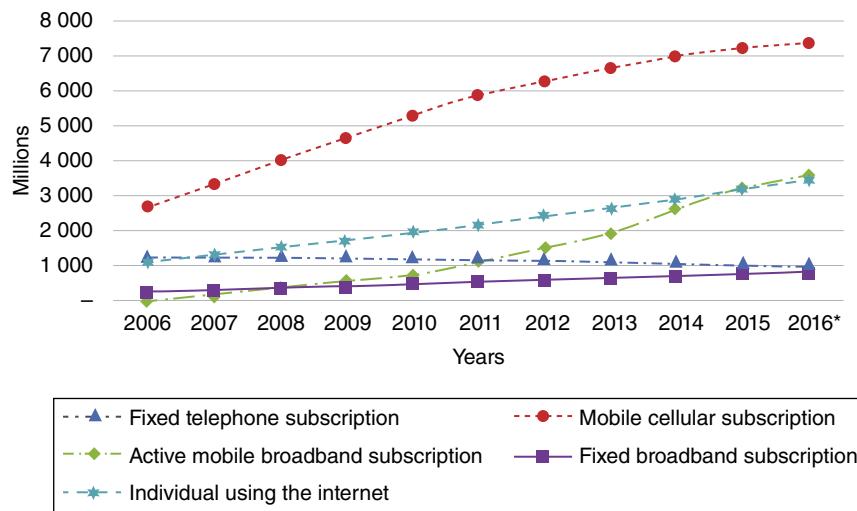


Figure 1.1 Growth of communication services encompassing the last decade.

1.2 Early Development

Wireless communication in its current practice is a very sophisticated technology, making long distance voice, data and multimedia communication possible between people, no matter which part of the world they reside in. The kind of evolution that wireless cellular technologies went through, in particular over the last three decades, and over the last two hundred years in general, makes for a fascinating journey. If we try to trace the initial efforts that became the foundation of the wireless communications of today, we have to go back as early as the ancient Greek, Roma and Chinese cultures, where electrical and magnetic properties of materials were experimented on. The early experiments on electrical and magnetic properties were not intended for wireless communication, since that sort of vision was not present as a motivation for these experiments.

We see that even in the 19th century, when the connection between electricity and magnetism was first developed, the intuition and imagination of what it could achieve was naturally missing amongst the researchers. It is good to say that it was mostly the random experiments that eventually led to the kind of communication systems we have now, and that is something which makes this journey more interesting. Even though, as mentioned above, the experiments towards trying to find electrical and magnetic properties in various ancient cultures, and considered as one of the foremost steps in this journey, it is also important to keep in mind that the last two hundred years present a more coherent and consistent picture that is paved with ground-breaking discoveries.

So, in our analysis the last two hundred years are of primary importance. We have to try to coherently present the connection of all those discoveries as to how one discovery led to another, and what became the motivation to carry out further discoveries. Until this decade, the story is not as linear and direct as it might appear when looking back to its destination. But as far as wireless communications are concerned, it would be unfair

and unimaginative to consider this point in time as the final destination, because as far as wireless communication is concerned, the sky is the limit, or even beyond [9].

Starting with the last two hundred years, say the year 1820, the Danish physicist Hans Christian Ørsted, during one of his lectures noticed that when the current from a battery was switched on and off, a compass needle showed the deflection. This observation led him to discover that an electric field creates a magnetic field; more particularly, an electric current produces a circular magnetic field as it flows through a wire.

The connection between electricity and magnetism was of immense importance that rapidly led to further developments. However, it is sometimes claimed that it was Gian Domenico Romagnosi who discovered this connection around two decades before, but the importance of this discovery cannot be considered insignificant. From the years 1823 to 1826, Dominique François Jean Arago, a French mathematician and physicist, discovered something called rotary magnetism, which was termed Arago's rotation. In simple words, he showed that a wire can become a magnet when current flows through it, and that most bodies could be magnetized. These discoveries were further explained by Michael Faraday later. André-Marie Ampère, another French physicist and mathematician, discovered electrodynamics. Ampère showed that two parallel wires carrying electric currents attract or repel each other, depending on whether the currents flow in the same or opposite directions. Ampere's initial plan was to gain more understanding between electricity and magnetism, and this had led him to these discoveries.

Michael Faraday's contributions are very significant in this journey, and he deserves all the credit that we can give him. After Ørsted had discovered the phenomenon of electromagnetism, it motivated many scientists to study this further, the efforts which helped Ampere in his discoveries. Similar motivation led Michael Faraday to carry out experiments, whereby he successfully managed to build two devices to produce electromagnetic rotation. Not only did he discover electromagnetic induction, but also predicted that electromagnetic forces extended the empty space around the conductor. In simple words, he predicted the existence of electromagnetic waves, which proved to be a true prediction later.

Samuel Finley Breese Morse, an American painter, invented the single-wired telegraph system. He was also a co-developer of the Morse code. This discovery also became possible because of the discovery of electromagnetism. The telegraph was important because it was a first attempt to use electromagnetism in an effort to communicate. The list of discoveries continued in the rest of the 19th century, and the German physiologist and physicist Hermann Ludwig Ferdinand von Helmholtz, worked on the phenomenon of electrical oscillation in 1847, which in itself was not a major contribution, but led to the major contribution by Heinrich Rudolf Hertz, one of his students, who later demonstrated electromagnetic radiations. In 1853, William Thomson also contributed in the form of calculating the period, damping and intensity, as the function of the capacity, self-inductance and resistance of an oscillatory circuit. Another proof of Helmholtz's work came from a discovery by Feddersen, who verified the resonant frequency of the tuned circuit, which was suggested by Helmholtz earlier.

James Maxwell is a prominent and influential name in the progression of wireless communication. He proved the existence of electromagnetic waves by formulating the electromagnetic theory of light and developed the general equations of the electromagnetic field, known as Maxwell equations. The most significant aspect of his work was that for the first time it was demonstrated that electricity, magnetism and also light are

manifestations of the same phenomenon. This discovery is of absolute importance, because it led to the prediction that radio waves exist, which was a very significant finding for the development of wireless communication. In 1866, the first transatlantic telegraph cable was installed and operated by using the Morse code, with a speed of five words per minute.

The first description of transmission of a wireless signal came in the form of a patent by the American dentist Dr Mahlon Loomis, in 1866. It was the idea of the wireless telegraph, from which he supposedly demonstrated the transmission of a wireless signal between two mountains. In 1882, another patent appeared in terms of wireless signal transmission, when American physicist Amos Emerson Dolbeam, transmitted a wireless signal using an induction coil, microphone, telephone receiver and a battery. In 1887, Hertz, a student of Helmholtz, sent and received wireless waves, using a spark transmitter and a resonator receiver. In 1895, Morse coded wireless signals were transmitted for more than over a mile by Guglielmo Marconi, and he carried out successful reception of a Morse coded wireless signal in 1901, which was sent across the Atlantic. In 1904, the patent of the diode came from J.A. Fleming. The triode amplifier was patented in 1906 by Lee DeForest. In the same year, Fessenden transmitted the first speech signal wirelessly. In 1907, the commercial Trans-Atlantic wireless service was started, which used huge ground stations. In 1915, wireless transmission of voice signals was carried out between New York and San Francisco.

Marconi carried out other ground-breaking and pioneering work in wireless communications by transmitting radio signals over long distances in 1920. Prior to that, Marconi was already working on the concept of wireless telegraphy. The breakthrough in his work came with his conclusion that if the height of the antenna could be raised, then the range of radio signal transmission could be extended, which he developed based on wireless telegraphy, where he grounded his transmitter and receiver. With these improvements, he managed to transmit a signal over 2 miles. He discovered short-wave radio, with wavelengths between the 10 and 100 meters range.

In 1920, we had our first commercial radio broadcast. In 1921, the police car dispatch radios came on the scene. In 1930, the television broadcast experiments were started by the BBC. In 1935, the first telephone call was made around the world. World War II led to rapid advancements in radio technology. In 1947, W. Tyrell proposed hybrid circuits for microwaves, and H.E. Kallaman constructed the VSWR indictor meter. In 1955, John R. Pierce proposed using satellites for communications. Sony marketed the first transistor radio. In 1957, the Soviet Union launched Sputnik I, which transmitted telemetry signals for about five months. The carterfone was a device invented in 1968 by Thomas Carter, which connected a two-way radio to the telephone system, letting one person on the radio talk to another person on the phone.

1.3 First Generation Cellular Systems

The prime developers of the first generation (1G) cellular network were the United States, Japan and some parts of Europe. It was based on analog modulation to provide voice services. In 1979, commercial cellular systems were implemented by Nippon Telephone and Telegraph Company (NTT) in Japan. Nordic Mobile Telephone (NMT-400) is a system developed in 1981 that supports international roaming and automatic handover.

Some European countries implemented this system at that time. Subscribers of NMT-400 were able to transmit up to 15 watts of power using car phones. Six countries – namely Finland, Sweden, Norway, Austria, Spain, and Denmark – adopted NMT-400.

The advanced mobile phone service (AMPS) and its alternative total access communication systems (ETACS and NTACS) were more successful for 1G. From the radio standpoint these above systems were identical. The main difference was the length of the channel bandwidth.

1.3.1 Advanced Mobile Phone Service

The advance mobile phone service (AMPS) was more advanced in comparison to the other 1G systems in the United States. It was deployed in Europe and Japan by an organization named Total Access Communication Systems (ETACS). As mentioned above, from the radio standpoint, the above-mentioned systems were identical, only differing in the length of channel bandwidth. For example, AMPS was based on a 30 kHz bandwidth, while the ETACS and NTACS used 20 kHz and 12.5 kHz for the channel bandwidth, respectively [11].

AT&T and Bell Labs first implemented the AMPS for commercial use in the year of 1983 in Chicago and its neighboring areas, then later in Israel in 1986, in Australia in 1987, and in Pakistan in 1990. By the mid-2000s, all commercial companies discontinued this system from the market around the world. This system was constructed using long base stations (height from 150 ft to 550 ft) with omnidirectional antennas. In the beginning, the carrier to interference ratio (CIR) was kept to 18 dB for better voice quality. Spectrum was assigned by FCC in the USA to two operators in each market, one for the incumbent telecommunications carrier and another for the non-incumbent operator. 20 MHz of spectrum was assigned for each operator, which could support a total of 416 channels. For voice communication, 395 channels were used and the remaining 21 channels were for control information. There were 7-cell frequency re-use patterns, where each sector consisted of 3 sectors per cell. The AMPS is based on the Frequency Modulation for voice communication and used Frequency Shift Keying (FSK) for managing the control channel. After the availability of 2G systems, AMPS were continued by the operators in North America for the purpose of a common fallback service for the entire region and for the roaming service between multiple operators that had implemented 2G systems.

1.3.2 Security in 1G

The first generation (1G) cellular system used analog communication, as stated before. Due to the vulnerable nature of analog signal processing, it was difficult to provide efficient security services for 1G. For example, eavesdropping was a pressing concern for 1G phones, as it was possible for anyone to listen in to a private communication between two users, because all it required was a simple receiver operating at the similar frequencies. There was absolutely no confidentiality in communication in 1G networks. Also, the identity of the cellphone could easily be duplicated, and all the call charges made from the duplicate phone could be directed to the original owner. Since the scale of the network was small, and a small number of users needed servicing, the 1G cellular

networks had a limited risk of mass cloning of the mobile sets. Although attempts had been made to completely get rid of mobile set cloning, they were proven to be unsuccessful. Even though the information about the number being dialed could be encrypted, the major problem was transmission through the air, as signals could easily be received by using any FM receiver, since the transmission used frequency modulation [16].

1.4 Second Generation Cellular Systems

The improvement of the processing abilities of hardware platforms made the development of 2G wireless systems possible. Digital modulation scheme was implemented in 2G, targeting the voice market. The overall system performance rapidly improved due to shifting from analog to digital modulation schemes. The total capacity in 2G was improved by using digital speech codecs, implementing time division and Code Division Multiplexing (CDM) techniques for multiplexing several users using a single channel. In 2G, stronger security systems were also introduced by applying encryption algorithms that were absent in the 1G.

Another attractive feature of the second generation along with other new applications was the short messaging service (SMS). The first SMS was sent using Vodafone GSM network on 3 December 1992 in the United Kingdom. Gradually, some European countries implemented this service to notify the users about the voice mail. Nokia released their first SMS supporting mobile phone, which was capable of sending SMS from one user to another. Today, over 23 billion SMS messages are sent from the mobile operator per day, all over the world. The SMS are used for news updates, business alerts, various payments, blogging, voting and for many other uses.

2G systems evolved to support packet data services, while the previous method was the circuit switched data service, which was similar in concept of dial-up modems. Wireless Access Protocol (WAP) was introduced to provide internet contents to handheld devices.

1.4.1 Global System for Mobile Communications

As soon as it became obvious that long-term economic goals in Europe had to be fixed, the CEPT was formed in 1982 by the “Conference Des Administrations Européennes Des Posts et Telecommunications” to address sector needs. The CEPT successively established the “Groupe Spéciale Mobile” (GSM), to develop the specification for a pan-European mobile communications network. The standardized system targeted spectrum efficiency, low mobile and base stations costs, international roaming, better voice quality, compatibility with other systems such as Integrated Services Digital Networks (ISDN), and the ability to support new services. Before GSM, the cellular market was scattered with a variety of mutually incompatible systems implemented in different countries. For example, Scandinavian countries had NMT-400 and NMT-900, the United Kingdom had TACS, Germany had C-450, and France had Radiocom.

The European telecommunications standards institute (ESTI) released the first version of the GSM standard, called the GSM Phase I in 1990. Consequently, many operators

implemented GSM and this standard gained acceptance outside of Europe. The standard was eventually renamed as the Global System for Mobile Communications.

The TDMA scheme is used in GSM air interface with a capability of multiplexing eight users in a single 200 KHz channel bandwidth, where the users were separated by different time slots. Gaussian minimum shift keying (GMSK) was introduced as a modulation technique of GSM. Because of the constant envelope property and significant power and spectral efficiency, the GMSK was convenient [1].

A circuit switched data of 9.6 kbps rate was also supported by GSM, along with the voice and SMS service. GSM packet radio systems (GPRS) were introduced by ETSI in the mid-1990s. It was an evolutionary step of GSM systems towards higher data rates. The GPRS and GSM systems both share the same frequency bands, signaling link and time slots. There were four different channel coding schemes to support the data, at the rates of 8 kbps to 20 kbps per slot. Theoretically, the GPRS was able to provide 160 kbps rate, where the 20–40 kbps rate was found in practice.

1.4.2 GSM Network Architecture

The GSM Network architecture is comprised of two major sub-components. This architecture forms the basis of the next generation (3G) systems and LTE. In Figure 1.2, the base station subsystem is comprised of the base-station transceiver (BTS) unit, with which the mobile stations (MS) and the base station controller (BSC) are connected over the air interface. BSC manages the traffic from several BTSs to the switching core. It also manages Mobility across BTSs. Another sub-component is Network Switching

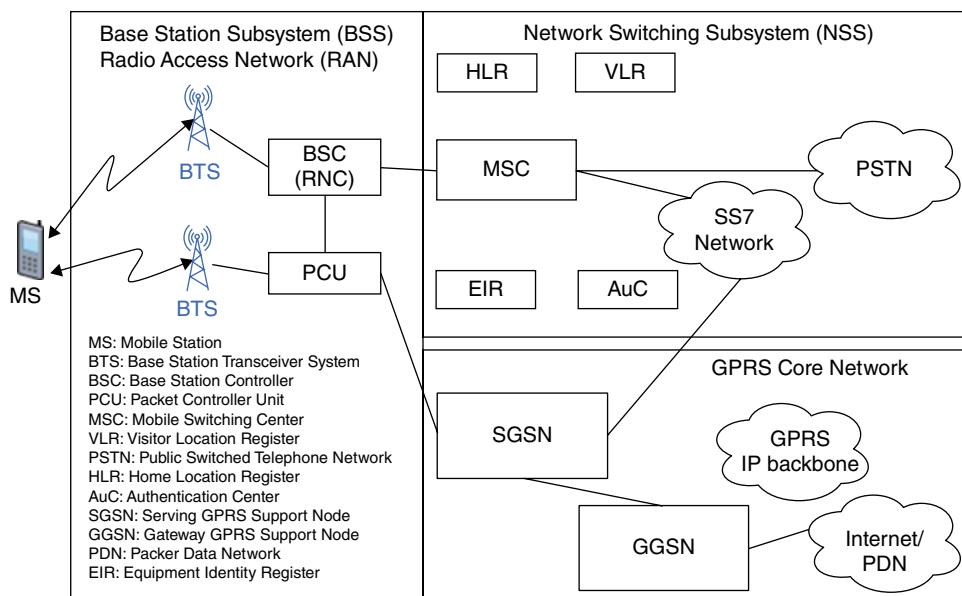


Figure 1.2 GSM network architecture.

Sub-system. Mobile Switching Center (MSC) and subscriber databases are parts of it. MSC carries out the switching to connect the calling party with the called party. MSC is connected with the Public Switched Telephone Network (PSTN), as shown in Figure 1.2. Home Location Register (HLR) and Visitor Location Register (VLR) are used to determine the suggested identity of the subscriber for the MSC.

The GPRS system can be upgraded from a GSM system by introducing new components, such as serving GPRS support node (SGSN), and gateway GPRS support node (GGSN), shown in Figure 1.2. For handling data, the packet control unit (PCU) is necessary in the BTS. SGSN was designed to provide location and mobility management. Providing IP access router functionality and connecting the GPRS network to the internet and other IP are the two tasks of GGSN [1].

The data rate of GSM was further increased with the introduction of an enhanced data rate for GSM evolution (EDGE) back in 1997. EDGE specified the use of 8PSK modulation that allows almost three times as much throughput compared to GPRS. An EDGE user could enjoy 80 kbps to 120 kbps of data rate.

1.4.3 Code Division Multiple Access

Code Division Multiple Access (CDMA)-based digital cellular technology was first proposed by Qualcomm in 1989. In 1993, Qualcomm obtained the acceptance of telecommunication industry association (TIA) to embrace their proposal as an IS-95 standard, which was the alternative to IS-54 TDMA, which was adopted earlier as the digital evolution of AMPS. Unlike GSM, multiple users share the same frequency band at the same time in IS-95 CDMA. A unique orthogonal spreading code is assigned for each user that helps to distinguish between different users on the receiver side. Spread signals showed noticeable improvement to multipath fading and interference. The channel bandwidth of IS-95 CDMA is 1.25 MHz for transmitting 9.2 kbps of lower voice signal.

The technical advantages of IS-95 CDMA were more capacity in per MHz of bandwidth, there was no limitation of built-in limit of number of users, power consumption was low so cell size of IS-95 was larger, and soft handoff was introduced. Another interesting feature was the ability to detect the period of silence so that transmission of data could be paused to save energy and increase overall efficiency. The above features gave CDMA systems a huge commercial and user acceptance.

Supplemental Code Channel (SCH) was introduced in the version of IS-95B. It is also known as packet mode transmission for increased efficiency. For example, it supports 14.4 kbps, which is allowed to combine 7 SCH to maintain the peak data rate of 115.2 kbps.

1.4.4 Security in 2G

The 2G cellular network was developed due to an increasing need for improved transmission quality, capacity and coverage. The advancements in semiconductor technology and microwave devices made digital transmission possible in mobile communications. 2G cellular networks incorporated data communications, unlike 1G, amongst other kinds of digital services such as text messages, picture messages and MMS (multimedia messages). With digitized services coming into play, data confidentiality and security

became of major concern. 2G cellular systems, in general, comprises of GSM, digital AMPS (D-AMPS), CDMA, and personal digital communication (PDC).

GSM is the most successful and widely-used standard in cellular communications throughout the world, as part of 2G cellular networks. It includes GSM900, GSM-railway (GSM-R), GSM1800, GSM1900, and GSM400. 2G phones using GSM were first introduced around 1990, first deployed in Finland in July 1991. IS-95, or CDMAONE, another technology under the 2G umbrella, based on CDMA, unlike GSM, which is Time Division Multiple Access (TDMA)-based. However, the use of GSM is much wider in scale than IS-95. The successor of GSM is wideband CDMA (W-CDMA), while the successor of IS-95 is CDMA 2000. In order to understand the security measures in 2G cellular networks, it is convenient to first focus on the security in the GSM. Supplemental Code Channel (SCH) was introduced in the version of IS-95B. It is also known as packet mode transmission for increased efficiency.

1.4.5 Security in GSM

GSM tries to focus on four aspects of security that include authentication of a user, ciphering of data and signaling, confidentiality of user identity, and the use of subscriber identity module (SIM) as a security module. SIM is another distinguishing feature of 2G cellular networks. The SIM is basically a detachable smart card containing subscriber information, and used for proving its identity with the operator along with the information regarding the kinds of services it is allowed to access. It plays a vital role in the security process. Authentication requires any particular user to prove that they are a valid customer requesting the service from a particular operator. Ciphering takes care of the interception of all the data and signaling. In order to handle confidentiality, GSM uses international mobile subscriber identity (IMSI) and, more particularly, uses Temporary Mobile Subscriber Identity (TMSI) to provide confidentiality for the user, by making sure that the information of any particular user being in any particular area is not disclosed to anyone to avoid any intrusion of confidentiality. The SIM card uses algorithms to develop a secure connection with the operator to carry out safe communication. In case the SIM card is taken by an unauthorized person, there is still a PIN code security measure in place.

GSM uses A3 and A8 algorithms between a mobile station and the GSM operator. These are the symmetric algorithms where the same key is used for the encryption and decryption. These algorithms have a one-way function, meaning that output can be found if the inputs are known, but the opposite is not possible. These algorithms are implemented in the SIM card. The technical details of these algorithms are further explained in [12].

1.4.5.1 IMSI

International Mobile Subscriber Identity (IMSI) represents the unique number for every subscriber in the world, and carries the information regarding the home network of the subscriber and country it belongs to. This particular information can be read from the SIM if local access to the SIM exists. It basically comprises of up to 15 decimal

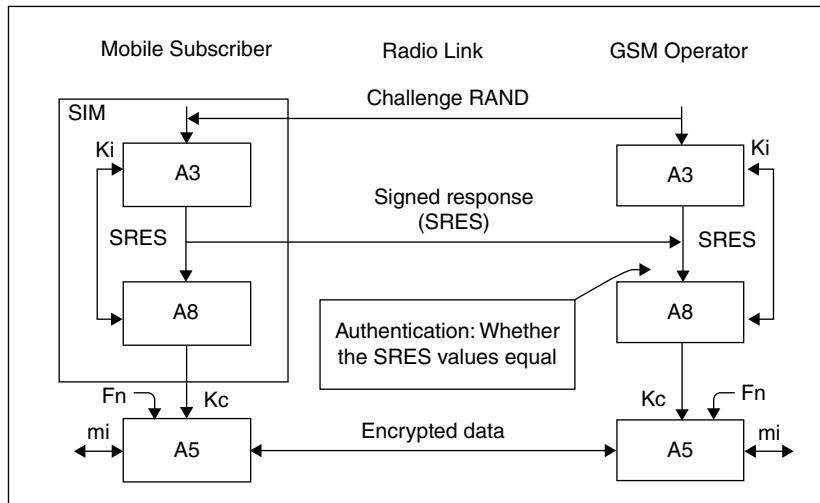


Figure 1.3 GSM authentication process.

digits, out of which the first 5 or 6 specify the network and the country. In order to prevent the eavesdropping, the IMSI is rarely sent, as instead the randomly-generated TMSI is used [14].

1.4.5.2 Ki

Ki is a root encryption key used in GSM. It is basically a randomly-generated 128-bit number assigned to a particular subscriber, and plays a large part in the generation of all the keys in GSM. The Ki is only known to the SIM and the Authentication center (AuC) for protection reasons. The mobile set also has no information about the Ki, other than just feeding the information to the SIM that it needs to know in order to perform the authentication or to generate the ciphering keys. The authentication and key generation is performed in the SIM.

1.4.5.3 A3 Algorithm

The A3 algorithm basically provides authentication to the user so that the user can access the system. The authentication between the network and the subscriber is carried out by the so-called challenge-response method.

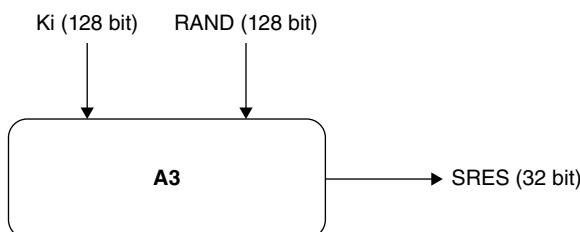


Figure 1.4 The A3 algorithm.

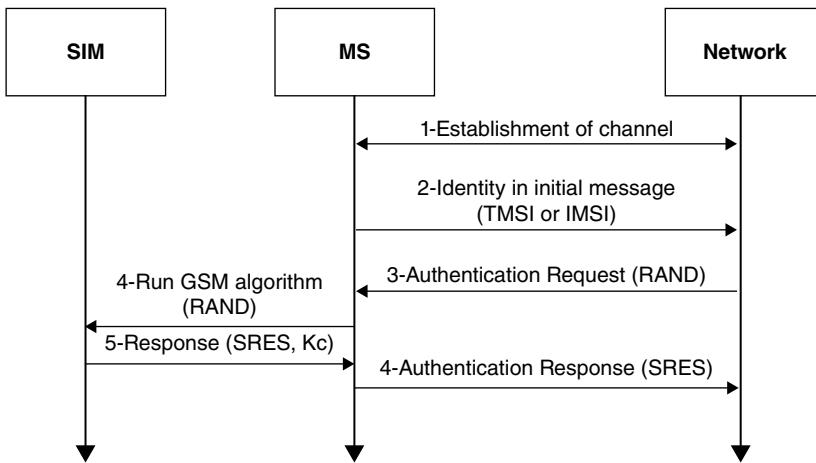


Figure 1.5 Working principle of A3 algorithm.

The 128 bit number (RAND) challenge is first transmitted from the network to the subscriber through the air interface, where it is processed at the SIM card. A3 authentication algorithm and Ki are responsible for sending the RAND to the SIM card in the phone. The SIM card processes RAND and the secret 128-bit key Ki, through the A3 algorithm, to produce a 32-bit signed response (SRES). The output of the A3 algorithm, that is, the SRES is transmitted back to the network from the subscriber again through the air interface. In the network, the AuC compares its value of SRES with the value of SRES that was received from the subscriber. If the two values match, authentication is considered to be successful, and the subscriber becomes eligible to join the network. The AuC does not store the copy of SRES, but takes the help of home location register (HLR) or visitor location register (VLR) whenever required.

1.4.5.4 A8 Algorithm

GSM uses ciphering to protect both user data and signaling at an air interface. Once the authentication has been successfully carried out, the RAND coming from the network together with the Ki coming from the SIM, are sent through an A8 ciphering key generating algorithm to create a ciphering key (Kc). This Kc created by the A8 algorithm is used with the A5 ciphering algorithm to cipher or decipher the data. The A5 algorithm is implemented in the hardware of the mobile phone as it encrypts and decrypts data in the air. Whenever the A3 algorithm is run to generate the SRES, the A8 algorithm also runs. Other than the A8 generating the ciphering key Kc, the network also generates the Kc, and shares it with the base stations handling the connection.

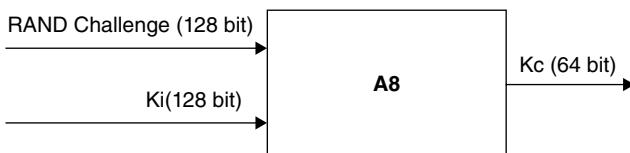


Figure 1.6 The A8 algorithm.

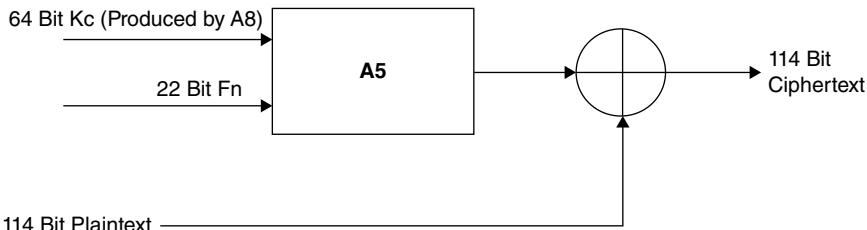


Figure 1.7 The A5 algorithm.

1.4.5.5 COMP128

COMP128 is technically a hash function, which is the implementation of A3 and A8 algorithms in the GSM standard. It is used to provide authentication and helps derive the cipher key (A3/8). GSM allows every operator to use their own A3/8 algorithm, and all the systems support this without the need for a transfer between the networks, even during roaming. However, the operators normally use, for example, COMP128 design, because it requires certain levels of expertise to make their own A3/8 algorithm.

1.4.5.6 A5 Algorithm

The A5 algorithm is basically a stream cipher and can be efficiently implemented on a hardware platform. Several implementations of this algorithm exist, and the most common ones are the A5/0, A5/1 and A5/2 (A5/3 is used in 3G systems). A5/1 is the most widely used, mainly in Western Europe and America, while the A5/2 is commonly used in Asia. A5/0 is used in so-called third-world countries, and countries under UN sanctions, which basically provides no encryption. A5 works on a bit-by-bit basis, which means that error in the received cipher text will only result in the event of the corresponding bit being erroneous.

GSM transmission is based on the sequence of bursts. Each burst has around 114 bits available for the information. A5/1 is used to produce for each burst a 114 bit sequence, which is XORed with the 114 bits before the modulation. A5/1 is executed using a 64-bit key together with a publicly known 22-bit frame number.

1.4.6 Security in IS-95

The procedures for authentication and security used in IS-95 are the same as in GSM; however, IS-95 uses an additional security technique known as the “private long code mask”.

For authentication, both the subscriber and the network use a secret key code. When any subscriber wishes to access the network, the network generates a random code and sends it to the subscriber. The secret key and the random code are used by the subscriber and network to generate another signal. This signed response is then sent to the network by the subscriber, where it is compared to the signed response stored in the network. If the signal matches, access is given to the system.

The additional feature of IS-95, private long code mask, just like the authentication key, is stored in both the subscriber and the network. It is like the public long code mask, which is an electronic serial number transmitted without protection used in analog

mode, except that it is more secure. The mobile or the system can initiate operation with a private long code mask by transmitting a “Long Code Transition” order after the call is set up.

1.5 Third Generation Cellular Systems

Third generation (3G) systems provided the higher data rates along with the higher voice capacity and also the advanced features such as applications like multimedia. The planning for the 3G was started in the early 1990s, with the invitation of proposals by International Telecommunications Union (ITU) known as IMT-2000. They started with the investigation of spectrum for these systems. The goal was to implement specifications for global harmony for mobile communication, which is able to initiate global interoperability by providing lower costs. ITU set the requirements for the data rates as the criterion for IMT-2000:

- in building or fixed environment data rates of 2 Mbps;
- for urban environments of 384 kbps of data rates; and
- 144 kbps for vehicular wide area environments.

Apart from the above requirements, the 3G systems were also intended to provide better quality of Service (QoS) for voice telephony and interactive gaming to internet browsing, e-mailing, and streaming multimedia applications.

1.5.1 CDMA 2000

The 3G standard for IS-95 was known as CDMA 2000 by the CDMA community. In 1999, the standard committee named as the third generation partnership project 2 (3GPP2), took the responsibility of official standardization process of CDMA 2000 from the development group Qualcomm and CDMA. CDMA 2000-1X was the first version of IS-95, where the channel bandwidth of 1.25 MHz was the same as IS-95. The data capability was enhanced by adding supplemental channels, which were actually separate logical channel. The capacity of each individual of fundamental channel was 9.6 kbps, where the capacity increased to 307 kbps by using the multiple supplemental channels. As this specification of channel capacity was in accordance with 3G requirements, it was instead called 2.5G. Gradually, in the version of CDMA 2000-3X, the data rate increased to 2 Mbps by using multiple carriers. Coherent modulation was introduced to improve the uplink channel quality. The capabilities of antennas were advanced by using transmit diversity and incorporating beam steering option. The key point of these upgrades was the backward compatibility. Both A and B versions of IS-95 and CDMA could be implemented in the same carrier, which is convenient for migration of those technologies [20].

1.5.2 UMTS WCDMA

As the popularity of GSM was at its peak, a joint collaboration was formed named 3GPP in 1998 by six regional telecommunication bodies from all over the world. The purpose was to continue the development of UMTS along with other standards of GSM.

The first UMTS standards of 3G were published in 1999, which is known as UMTS Release 99. It brought global success, which can be seen in the statistics of 3G Americas, and the UMTS Forum in May 2010 recorded that the total number of operators of the UMTS network were 346 in over 148 countries. The number of subscriber at that time was 450 million [8].

The architectural design of UMTS was kept the same as the GSM/GPRS network, but the 3G air-interface known as Wide-band CDMA (WCDMA) was a huge modification compared to the 2G air-interface. The design of WCDMA was provoked due to the success of IS-95. WCDMA is actually Direct Spread Spectrum CDMA systems, where user data is multiplied with pseudo random codes to provide synchronization, channelization and scrambling. This system is designed to operate in the 5 MHz bandwidth, which can support 100 different voice calls simultaneously. The peak data rate then varies from 384 to 2048 kbps. In addition, WCDMA supports using multi-code for increasing data rate for a single user.

1.5.3 UMTS Network Architecture

The UMTS network architecture consists of User Equipment (UE), UMTS Terrestrial Radio Access Network (UTRAN) and Core Network (CN). These three major subsections are shown in Figure 1.8. Looking at the first subsection, the components are similar to the 2G network. In UTRAN, Node Bs are connected to the RNC to manage radio resources in data link layer. This section has been developed for the service access point, which was absent in 2G. The CN part is also similar to the 2G, which controls different location registers. The function of SGSN and GGSN are kept the same as its ancestor, 2G.

There are a few differences in the architecture of 2G and 3G. For example, in 3G the base stations are known as Node B, which are controlled by the RNC, then the connection is made with a core switching network for voice calls and data traffic. RNC are connected to one another to confirm soft handover with lowest call drop rate. Except the RNC and Node B in the 3G architecture, there are no major differences with 2G in the overall design. It meant that those two architectures could connect with each other and work with the same core, packet switching and charging network. Only the modulation scheme is the major difference in 3G architecture.

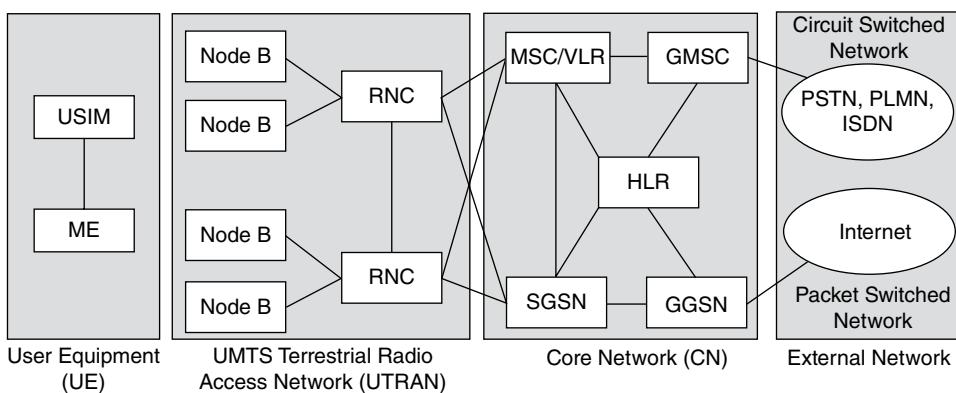


Figure 1.8 UMTS Radio Access Network.

1.5.4 HSPA

High-speed packet access is shortend HSPA. There are two key enhancements combined by 3GPP to UMTS-WCDMA, which are:

- 1) High-Speed Downlink Packet Access (HSDPA), introduced in the Release 5 in 2002; and
- 2) High-Speed Uplink Packet Access (HSUPA), which was available in Release 6 in 2004.

AT&T implemented the first HSDPA network, after which around 303 operators from 130 countries worldwide deployed HSDPA. Mostly, the HSPA was deployed as an upgrade of software to the existing UMTS systems [5].

According to internet usage patterns in the late 1990, the users demanded higher speed of downloads. So UMTS evolution focused on improving the downlink. Later, the HSDPA introduced a new downlink transport channel that was capable of providing up to 14.4 Mbps theoretically. It is named the High-Speed Downlink Shared Channel (HS-DSCH). Here the multiplexing technique was time division multiplexing with a limited use of CDM. HSDPA consists of 16 Walsh codes, 15 of which were used for traffic; 5, 10 or 15 codes could be used for a single user to gain higher throughput. The channel frame length was 2 ms, unlike WCDMA of frame length of 15, 20, 40 or 80 ms. In practice, the user of HSDPA was able to obtain throughputs in the range of 500 kbps to 2 Mbps.

1.5.5 Security in 3G

As mentioned above, 3rd generation cellular networks introduced services such as video, audio and graphics applications. It also introduced video telephony and video streaming via cellular networks communication. It was an attractive feature of mobile cellular networks, when looking at the evolution it went through. Extrapolating from the limitations of the 1st generation cellular networks, it was a landmark of sorts. CDMA 2000 and UMTS CDMA came under the 3G umbrella [13].

3G or UMTS (Universal Mobile Telecommunications), or in particular IMT-2000, provided a single compatible standard for cellular networks that could be used worldwide for all mobile applications. It provided support for both packet-switched and circuit-switched data communication. The security of CDMA 2000, and UMTS WCDMA, is covered below.

1.5.6 Security in CDMA2000

The entities participating in the CDMA 2000 security include the home network, the home location register and authentication center (HLR/AC), the serving network, the visitor location register and the Mobile station controller/packet data serving node (VLR and MSC/PDSN), the mobile subscriber (MS), and the user identity module (UIM).

The authentication and key management (AKA) protocol used in CDMA 2000 is the UMTS AKA mechanism described in the next section. The AKA procedure is executed in two stages. The first stage involves transfer of security credentials (authentication vector, AV) from the home environment (HE) to the serving network (SN).

The HE mainly contains the HLR and AC, and the SN consists of the parts of the core network that are directly involved in setting up connections.

In terms of access security, the SN network elements of interest are the PDSN, which handles packet-switched traffic, and the circuit switched nodes VLR/MSC. An operator with a physical access infrastructure will normally have both HE and SN nodes [8].

1.5.7 Security in UMTS

The UMTS security architecture is grouped together in five different sets of features, as shown in the Figure 1.9. The description of these groups of features is given as:

- 1) *Network access security*: provides the subscriber with secure access to the 3G services, and gives protection against attacks to the radio interface;
- 2) *Network domain security*: allows all the subscribers to be able to securely exchange signaling data and provides protection against attacks to the wireline network;
- 3) *User domain security*: deals with secure access to mobile stations;
- 4) *Application domain security*: makes sure that applications in the user and provider domain are able to communicate with each other securely;
- 5) *Visibility and configurability of security*: provides security information to the users, as to which security features are in place, and whether a certain security feature requires activation or not.

The network access security features described above can be further classified into the following two categories. The categories with their description are explained below:

- 1) *User authentication*: is the property of the network that provides service confirming the validity of the identity of the user, and
- 2) *Network authentication*: is the property that the user validates, which is connected to a serving network with is authorized by the user's home network.

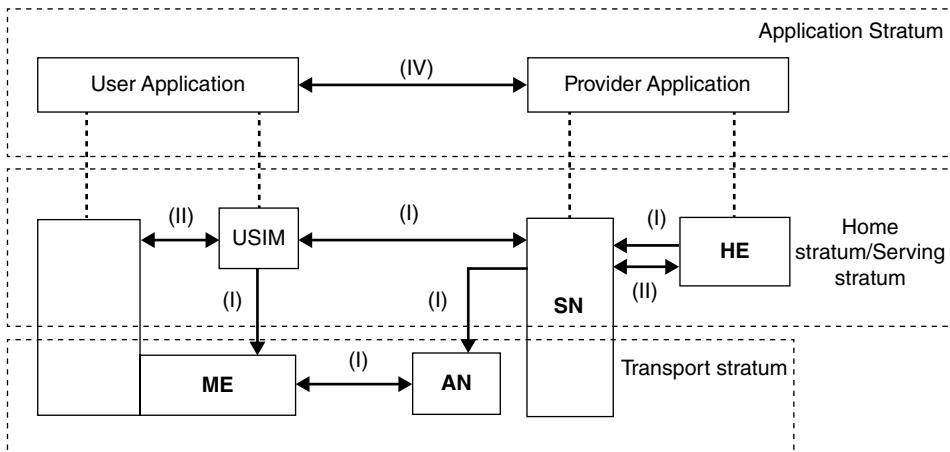


Figure 1.9 Overview of UMTS security architecture.

The following security features are associated with the confidentiality of the data on the network access link:

- *Cipher algorithm agreement*: the property that ensures that subscriber and serving network can securely decide on the algorithm that should be subsequently used;
- *Cipher key agreement*: the property that subscriber and the serving network mutually decide on the cipher key that should be subsequently used;
- *Confidentiality of user data*: the property that ensures that user data is protected on the overhead, such that it cannot be overheard; and
- *Confidentiality of signaling data*: the property that ensures that signaling data cannot be overheard on the radio interface.

The features provided to achieve integrity of data on the network access link are:

- *Integrity algorithm agreement*: the property that the subscriber and the serving network can securely decide on the integrity algorithm that should be subsequently used;
- *Integrity key agreement*: the property that the subscriber and the serving network agree on an integrity key that shall be subsequently used; and
- *Data integrity and origin authentication of signaling data*: the property that the subscriber or serving network is able to verify that signaling has not been modified later after it was sent by the sending entity, and that the origin of the signaling data received is the valid one.

UMTS AKA is a security mechanism used to accomplish the authentication features and all of the key agreement features described above. This mechanism is based on a challenge/response authentication protocol implemented in such a way as to achieve maximum compatibility with GSM's subscriber authentication and key establishment protocol, so that the transition from GSM to UMTS can be made. A challenge/response protocol is a security measure used by one entity to verify the identity of another entity, without revealing a secret password shared by the two entities involved [23]. Each entity must prove to the other that it knows the password without actually revealing the information that it has knowledge of the password.

The UMTS AKA process is started by a serving network after first registration by a user, after a service request, after a location update request, after an attach request, and after a detach request or connection re-establishment request. The information about the user must be transferred from the user's home network to the serving network in order to complete the process.

Table 1.1 Structure of an authentication vector.

| Field | Description |
|-------|----------------------|
| RAND | Random Challenge |
| Ck | Cipher key |
| Ik | Integrity Key |
| AUTN | Authentication Token |
| XRES | Expected Response |

Table 1.2 Structure of AUTN field of an authentication vector.

| Field | Description |
|-------|---------------------------------|
| SQN | Sequence Number |
| AMF | Authentication management Field |
| MAC-A | Message authentication code |

1.6 Cellular Systems beyond 3G

We present an overview of HSPA+, WiMAX and LTE in the following subsections. Although many industries started marketing the WiMAX as 4G systems, technically that was not the case. From an engineering perspective, both WiMAX and LTE represent a break from conventional 3G systems in terms of air-interface technology and network architecture both. These systems are capable of providing throughput level in megabit per second by using advanced signal procession techniques [6,22].

1.6.1 HSPA+

In June 2007, the Release 7 of 3GPP made an enhancement as a further evolution of HSPA. It is sometimes referred to as HSPA+. The key technical enhancements of HSPA+ are achieving higher-order modulation with multiple input multiple output (MIMO) by gaining higher peak rates, operation in dual-carrier downlink, packet connectivity when required to improve battery life, improved mobile receivers for capacity enhancement, and improved data rate and using single frequency network for better performance in multi-cast and broadcast. In May 2010, 56 operators in 34 countries deployed HSPA+ [7,25].

1.6.2 Mobile WiMAX

The institute of Electrical and Electronic Engineers (IEEE) established a group, named 802.16, to develop a standard for wireless metropolitan area network (WMAN). They introduced a standard for fixed wireless application in 2001, and gradually this was enhanced to support mobility. This revised version was known as 802.16e in 2005 and renamed as Mobile WiMAX. The industry consortium, named the Worldwide Interoperability for Microwave Access (WiMAX) Forum, was formed in 2001. The main purposes were to promote, develop, perform interoperability, conformance testing and certify end-to-end wireless systems based on the IEEE 802.16 air-interface standards. In 2007, WiMAX gained the approval of ITU as an IMT-2000 terrestrial radio interface option called IP-OFDMA. As the WiMAX network is designed using IP protocols, which does not offer circuit switched voice telephony, voice services can be provided using the VoIP (Voice over Internet Protocol). Within 2010, there were 504 WiMAX network operators in 147 countries. The notable thing is the number of motivational aspects in LTE design; for example, usages of OFDM and OFDMA technology, was inspired by the implementation of WiMAX [6,17,19].

1.6.3 LTE

The drastic growth of use of the Internet was the motivation for mobile broadband. As the mobile devices were continuously integrating various applications dealing with information, communication and medium of entertainments, it was the demand of time to enable the on-demand access to multimedia content from anywhere. Statistics showed that by the end of March 2009, the number of mobile broadband subscribers reached 225 million. To meet this huge number of services with higher performance, LTE design integrates some important radio and core network technologies. Amongst those technologies, the key features of LTE are discussed in three subsections below [10,24]:

1.6.3.1 Orthogonal Frequency Division Multiplexing (OFDM)

Orthogonal Frequency Division Multiplexing (OFDM) was the key difference between the existing 3G systems and the LTE. The traditional 3G systems were based on UMTS and CDMA 2000, where CDM techniques were used. OFDMA provides high data rates along with many more advantages. Due to high data rates, there are more probabilities of intersymbol interference because of multipath. OFDMA was the solution for the problem, by using multicarrier modulation, where high bit rate data streams are divided into several parallel lower bit rates. OFDMA also reduced the computational complexity, because of the implementation of Fast Fourier Transform (FFT). There were other advantages such as coding and interleaving diversity, efficient multicarrier scheme, efficient support of broadcast services, etc.

1.6.3.2 SC-FDE and SC-FDMA

To achieve better battery life, the Single Carrier Frequency Equalization (SC-FDE) transmission method used to transmit the data symbols are sent as a sequence of QAM symbols with an added cyclic prefix. For the uplink of LTE implements, SC-FDMA (multiple version of SC-FDE) allows multiple users to use parts of the frequency spectrum. The complexity of the transmitter and receiver is increased for using these systems.

1.6.3.3 Multi-antenna Technique

The multi-antenna technique provides the solutions of system capacity, link robustness and spectral efficiency. It is possible to combat multipath fading and obtain transmit diversity by using multi-antenna. Beamforming is possible by using multi-antenna so that the transmitted signals can be directed towards the most efficient direction of the receiver. It reduces the signal-to-interference ratio. Another important feature is multiuser MIMO, which allows multiple users in the uplink.

1.6.4 LTE Network Architecture

There are a few differences between the architecture of UMTS and the LTE systems architecture, which is depicted in Figure 1.10. Unlike the UMTS architecture, there is no RNC, SGSN and GGSN blocks in the LTE. In LTE, the Node B is known as eNode B, which is connected to Serving Gateway (S-GW) to terminate interface towards the 3GPP radio access network and Packet Data Network Gateway (P-GW) to control IP data services, including routing, allocating of IP address, enforcing policy, and

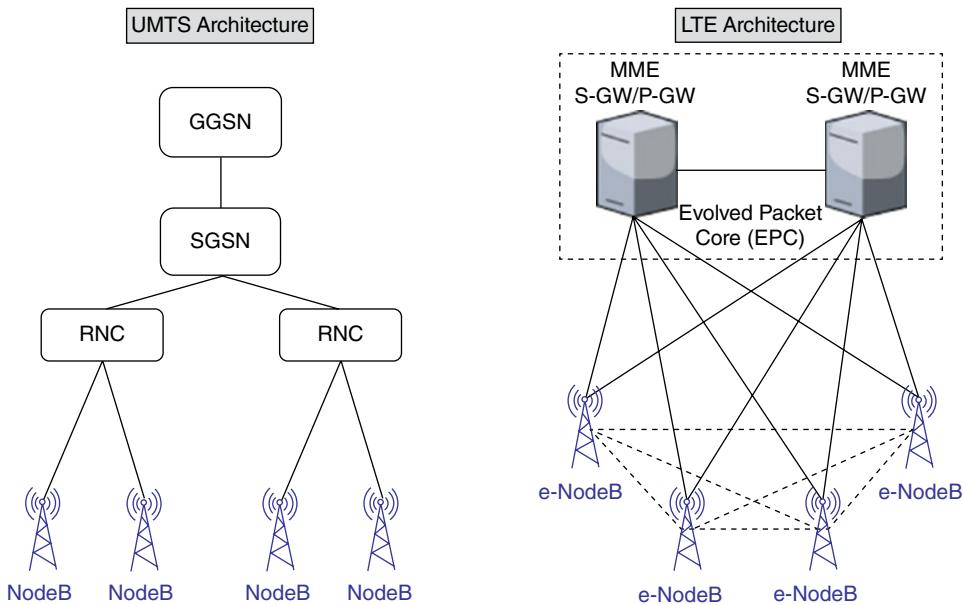


Figure 1.10 UMTS vs. LTE architecture.

providing access to non-3GPP access network. Function of Mobility Management Entity (MME) is to support equipment context and identity by authenticating to the authorized users [21]. The S-GW and P-GW are connected to each other and known as the Evolved Packet Core (EPC). EPC provides the function of access control, packet routing and transfer, mobility management, radio resource management, security and network management [15].

1.7 Fourth Generation Cellular Systems

4G is the fourth generation of mobile telecommunication technology. The requirements of 4G systems are defined by the ITU in IMT Advanced. The requirements are:

- a high degree of sharing features world-wide, which will support a vast range of services and applications with cost efficiency;
- internetworking compatibility within the IMT and also with other radio access networks;
- services compatibility with fixed and IMT networks;
- mobile devices with high quality;
- worldwide roaming capability;
- user-friendly equipment, services and applications; and
- 100 Mbps for high mobility and 1 Gbps for comparatively low mobility devices for supporting advanced services.

These requirements do not quantify the performance requirements, except the last one. In the detailed description of the IMT-Advanced, specific goals were set average and cell-edge performance to the usual peak data rates.

1.7.1 Key Technologies of 4G

1.7.1.1 Enhanced MIMO

Multiple-Input Multiple-Output (MIMO) is a key technique in the modern cellular system, which refers to the use of multiple antennas at both the transmitter and receiver sides. Therefore, base stations and terminals are equipped with multiple antenna elements intended to be used in transmission and reception to make MIMO capabilities available at both the downlink and the uplink.

Enhanced MIMO is considered as one of the main characteristics of LTE-Advanced that will allow the system to meet the IMT-Advanced rate requirements recognized by the ITU-R. The majority of the MIMO technologies already presented in LTE are expected to remain playing a vital role in LTE-Advanced, namely beamforming, spatial multiplexing and spatial diversity. However, further improvements in peak, cell-average and cell-edge throughput need to be obtained to significantly increase performance.

The above-mentioned techniques need some level of channel state information (CSI) at the base station, so that the system can adjust to the radio channel conditions and substantial performance improvement can be attained. For TDD systems, this information is easily collected from the uplink, provided the channel fading is adequately slow, due to the fact that the same carrier frequency is used for transmission and reception. Again, due to the asymmetry of FDD systems, feedback information over the reverse link is required. Full CSI could cause an additional overhead that might be too much, so quantization or statistical CSI are preferable in practice. In addition, terminal mobility can pose serious difficulties to the system performance, as the channel information arriving at the eNB may be outdated.

Multi-antenna techniques in a multi-user situation has the role of delivering streams of data in a spatially multiplexed fashion to the different users in such a way that all the degrees of freedom of a MIMO system are to be used. The idea is to perform an intelligent Space-Division Multiple Access (SDMA), so that the radiation pattern of the base station is adapted to each user to obtain the highest possible gain in the direction of that user.

1.7.1.2 Cooperative Multipoint Transmission and Reception for LTE-Advanced

4G cellular networks have to instantaneously provide a large number of diverse users with very high data rates, and the capacity of the new radio access systems needs to be enlarged. Conventionally, in cellular systems, each user is allocated to a base station on the basis of principles such as signal strength. At the terminal side, all the signals arriving from the rest of the base stations in the form of interference radically limit the performance. The user also connects with a single serving base station while causing interference to the rest of them. Due to the interference limitation of cellular systems, the task of high data delivery cannot be accomplished by simply increasing the signal power of the transmission. Each base station processes in-cell users independently, and the rest of the users are seen as inter-cell interference whose transmission power would also be increased.

CoMP in the framework of LTE-Advanced involves several likely coordinating schemes among the access points. Coordinated beamforming/scheduling is a simpler method, where user data are transmitted only from a single cell. Joint processing techniques require multiple nodes to transmit user data to the UE. Two approaches are

being considered as joint transmission, which requires multi-user linear precoding, and dynamic cell selection, where data is transmitted from only one cell that is dynamically selected.

1.7.1.3 Spectrum and Bandwidth Management

To meet the requirements of IMT-Advanced, as well as those of 3GPP operators, LTE-Advanced considers the use of bandwidths of up to 100 MHz in the following spectrum bands:

- 450–470 MHz band (identified in WRC-07 to be used globally for IMT systems);
- 698–862 MHz band (identified in WRC-07 to be used in Region 22 and 9 countries of Region 3);
- 790–862 MHz band (identified in WRC-07 to be used in Regions 1 and 3);
- 2.3–2.4 GHz band (identified in WRC-07 to be used globally for IMT systems);
- 3.4–4.2 GHz band (3.4–3.6 GHz identified in WRC-07 to be used in a large number of countries); and
- 4.4–4.99 GHz band.

1.7.1.4 Carrier Aggregation

In order to utilize the wider bandwidths of up to 100 MHz, a carrier aggregation scheme is required. It is designed in such a way that it consists of several component carriers of 20 MHz bandwidths, so that the LTE-Advanced devices can use greater amounts of data by using several carriers. For example, in a contiguous band, the scenario of using the component carriers for LTE and LTE-Advanced is shown. Also, there are proposed methods available for non-contiguous method with single and multiband operation.

1.7.1.5 Relays

Relay is implemented in the LTE-Advanced technology. The purpose of relaying is to provide coverage in new areas, cell-edge throughput, temporary network deployment, high data rate coverage and group mobility. Also, there are more advantages such as cost reduction, because it requires lower overhead costs than the eNB. The consumption of transmission power can also be reduced by using relay when the location of the relay is appropriate.

1.7.2 Network Architecture

In Release 8, 3GPP specified the rudiments and requirements of the EPS architecture that will serve as a basis for the next-generation networks. The disclaimers contain two major work items, namely LTE and SAE, which led to the specification of the Evolved Packet Core (EPC), Evolved Universal Terrestrial Radio Access Network (E-UTRAN), and Evolved Universal Terrestrial Radio Access (E-UTRA). Each of those corresponds to the core network, radio access network, and air interface of the whole system, respectively. The EPS provides IP connectivity between a UE and an external packet data network using E-UTRAN. Figure 1.12 provides an overview of the EPC, other legacy Packet and Circuit Switched elements and 3GPP RANs, along with the most important interfaces. In the services network, only the Policy and Charging Rules Function (PCRF) and the Home Subscriber Server (HSS) are included, for simplicity [4,28].

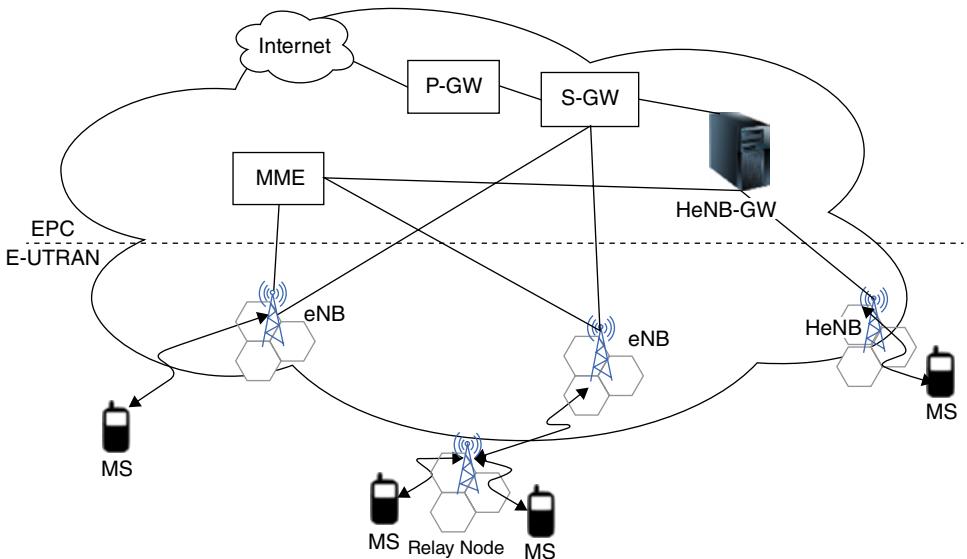


Figure 1.11 LTE-advanced E-UTRAN architecture.

In the framework of 4G systems, both the air interface and the radio access network are being improved or redefined, but so far the core network architecture, that is, the EPC, is not undergoing major changes from the previously standardized SAE architecture. Therefore, in this section, an overview of the E-UTRAN architecture and functionalities is given, which are defined for the LTE-Advanced systems and the main EPC node functionalities, shared by Releases 8, 9 and 10.

Enhanced Node B is the core part of the E-UTRAN architecture. It provides the air interface towards the UE. Each eNB is considered as the logical component that serves one or more several E-UTRAN cells. The target of this technology is to increase coverage, higher data rates, better QoS performance and fairness for the users. The EPC is a flat all-IP based core network. It can be accessed through 3GPP radio access, which allows the handover procedure. The Mobility Management Entity (MME), Serving Gateway (S-GW), and Packet Data Network Gateway (PDN-GW) work in a similar way to the LTE network architecture [26].

1.7.3 Beyond 3G and 4G Cellular Systems Security

Fourth generation cellular networks promise to provide higher user data rates, lower latency and a complete internet protocol (IP)-based network architecture. The major difference between the 3G and 4G cellular networks is that 4G operates entirely on IP protocol and architecture. For this reason, WiMAX is also considered as part of 4G networks. While discussing beyond 3G technologies, the major underlying technology in use is the LTE. Even though a similarity can be drawn between LTE and WiMAX, because of the IP-based protocol and architecture, they differ from each other in network architecture and security. The all IP-based infrastructure brings up increased security issues compared with the previous generation cellular technologies. For this

reason, in 4G networks, extra security mechanisms are expected to be carried out to provide the security for reliable communication.

The major concern in 4G security naturally includes that the user who wants to access the network must be authenticated along with the device that will be connected to the network. For this reason, security credentials, identity, certificates, username and password, are used for authentication. If we draw a comparison, starting from 2G when security was started to be taken as a major concern in cellular networks, a unique ID is used on the SIM card in 2G. While in 3G and 4G LTE, temporary ID and further abstraction is used to limit the possibilities of any sort of intrusion. In 4G, further secure signaling between the UE and MME (Mobile Management Entity) is introduced, and also security measures are taken care of between 3GPP and trusted non-3GPP users. As mentioned before, because of operating of open IP-based architecture, security remains the pressing concern in 4G cellular, and is given strong emphasis. It is important to point out that LTE- and LTE-Advanced are the same technologies. The label "Advanced" was primarily added to highlight the relationship between LTE release 10 (LTE-Advanced) and ITU/IMT-Advanced. This does not make LTE-Advanced a different system from LTE [3,28].

1.7.4 LTE Security Model

Figure 1.12 shows the authentication method of LTE with step-by-step details. The authentication process in LTE is initiated by the authentication server when it sends Enhance Authentication Protocol request/Identity message (EAP) to the UE. The UE responds by replying to the EAP-response/Identity message containing the identity message and Network Access Identifier (NAI). Upon receipt of the EAP-response/identity message, the authentication server tries to access the UE's certificate from its record. The authentication server generates the EAP-Request/Authentication and Key Agreement (AKA)-Challenge message using the standard AKA process.

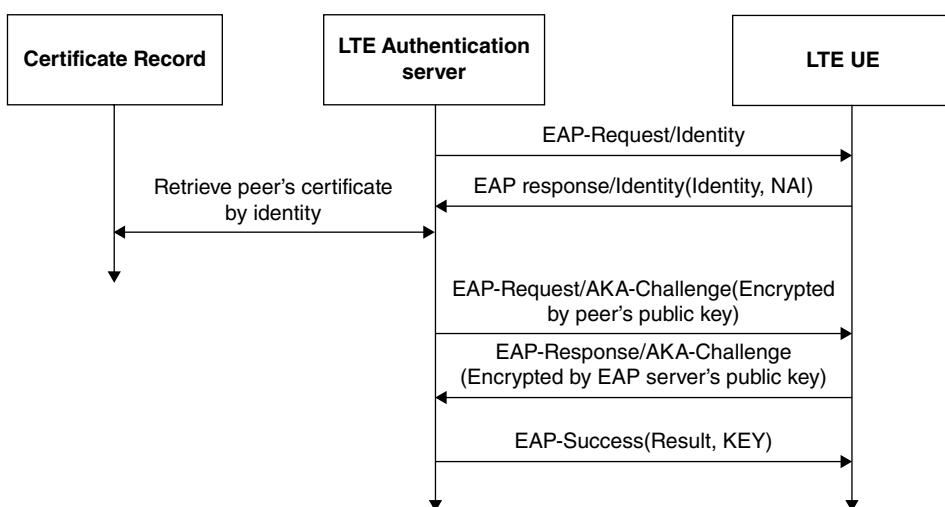


Figure 1.12 Authentication Process in LTE.

The authentication server sends the EAP-Request/AKA-Challenge message encrypted by the UE's public key to the UE. The UE then decrypts the EAP-Request/AKA-Challenge message using its own private key, and then sends the EAP-Response/AKA-Challenge to the authentication server. The authentication server decrypts the information using the server's private key and verifies the EAP-response/AKA-Challenge message using the AKA algorithm. If the message is correct, the EAP server sends the EAP success message to the UE [18].

1.7.5 Security in WiMAX

The IEEE 802.11 security issues were merged by the WiMAX group into IEEE 802.16 standards. This was done because as the WiMAX standard evolved from 802.16 to 802.16a to 802.16e, the requirements evolved from the line-of-sight to mobile WiMAX. Hence, the requirements for security and corresponding standards also evolved to address the changing demands. In order for the security features of the initial IEEE 802.16 standard to work for the IEEE 802.16e standard, additional features are added [29].

The key new features are listed as follows:

- 1) Privacy Key Management Version2 (PKMv2) protocol;
- 2) User authentication is carried out using Extensible Authentication Protocol (EAP) method;
- 3) Message authentication is carried out using Hash-based Message Authentication Code (HMAC) or Cipher-based Message Authentication Code (CMAC) scheme; and
- 4) Confidentiality is achieved using Advance Encryption Standards (AES) [2].

When it comes to WiMAX, over-the-air security is a major concern ensuring end-to-end network security. While security architecture has been developed to mitigate against threats over the air, there still remain a number of challenges. The main challenge seems to be the balance of security needs with the cost of implementation, performance and interoperability. Since WiMAX uses IP transport mechanisms in handling control/signaling and management traffic, network operators will also have to defend against general IP security threats [30].

1.8 Conclusion

In this chapter, we presented the evolution of cellular systems. We focused on the development related to radio interface, network architecture and security measurements for different generation of cellular systems. The very first 1G system to the most recent 4G system were briefly discussed. The fifth generation (5G) cellular system is the next major phase of cellular communication, also referred to as wireless technologies beyond 2020. The 5G standard has to cope with the demand of a 1000-fold capacity and seamless connectivity for at least 100 billion devices. A stand-alone technology will not be able to cope up with such a demand.

A combination of spectral efficiency, spectrum enhancement and network efficiency, etc. will meet the challenges of 5G. Various efficient networking technologies, such as small cells, device-to-device (D2D), and software-defined networks (SDN) technologies will be adopted. As a part of the spectrum enhancement, the unlicensed bands will be

used efficiently in addition to the licensed bands. Different technologies, such as massive MIMO and millimeter-wave MIMO, will play a key part for spectral efficiency in 5G networks and new radio interfaces have to be designed for these technologies.

The 5G system architecture and security models will be discussed in subsequent chapters. While this chapter provides basic information regarding the earlier cellular communications, we invite interested readers to go through the publications referenced in this chapter to gain a comprehensive knowledge of the evolution of cellular systems.

References

- 1 Al-Tawil, K. (King Fahd University of Petroleum and Minerals), Akrami, A. and Youssef, H. (1998) A new authentication protocol for GSM networks. *IEEE Conference on Local Computer Networks*, 11–14 October.
- 2 Andrews, J., Ghosh, A. and Muhamed, R. (2007) *Fundamentals of WiMAX*. Upper Saddle River, NJ: Prentice Hall.
- 3 Sankaran, C.B (2009) Network access security in next- generation 3GPP systems: A tutorial. *Communications Magazine, IEEE*, 47(2), 84–91.
- 4 Dahlman E. and Parkvall S. (2011) *4G: LTE or LTE-Advanced for Mobile Broadband*. West Sussex, UK: John Wiley & Sons, Ltd.
- 5 Johnston, D. and Walker, J. (2004) Overview of IEEE 802.16 security. *Security & Privacy, IEEE*, 2(3), 40–48.
- 6 Chin-Tser, H. and Chang, J.M. (2008) Responding to security issues in WiMAX networks. *IT Professional*, 10(5), 15–21.
- 7 Holma, H. et al. (2007) High-speed packet access evolution in 3GPP release 7. *IEEE Communications Magazine*, 45(12), 29–35.
- 8 Holma, H. and Toskala, A. (2002) High-speed downlink packet access. In: Chapter 11, *WCDMA for UMTS*. New York: John Wiley & Sons, Inc.
- 9 Hudderman A.A. (2003) *The Worldwide History of Telecommunications*. West Sussex, UK: John Wiley & Sons, Ltd.
- 10 IEEE Communications Magazine, Special issue on LTE–LTE Part I: Core Network, February 2009.
- 11 ITU Telecommunications indicators update 2016. www.itu.int/ITU-D/ict/statistics/
- 12 Josyula, R., Pankaj Rohatgi, R., Scherzer, H. and Tinguley, S. (2002) Partitioning attacks: or how to rapidly clone some GSM cards. *IEEE Symposium on Security and Privacy*, 12–15 May.
- 13 Korhonen, J. (2001) *Introduction to 3G Mobile Communications*. Norwood, MA: Artech House, Inc.
- 14 Lo, C.-C. and Chen, Y.-J. (1999) Secure Communication architecture for GSM networks. *IEEE Pacific RIM Conference on Communications, Computers and Signal Processing Proceedings*, 22–24 August.
- 15 Chang, M.J., Abichar, Z. and Chau-Yun, H. (2010) WiMAX or LTE: who will lead the broadband mobile internet? *IT Professional*, 12(3), 26–32.
- 16 Shin, M., Ma, J., Mishra, A. and Arbaugh, W.A. (2006) Wireless network security and interworking. *Proceedings of the IEEE*, 94(2), 455–466.
- 17 Marshall, P. (2008) HSPA+ challenges both WiMAX and LTE on the road to 4G. *Yankee Group Trend Analysis*, 19 September.

- 18 Seddigh, N., Nandy, B., Makkar, R. and Beaumont, J.F. (2010) Security advances and challenges in 4G wireless networks. In: *Eighth Annual International Conference on Privacy Security and Trust (PST)*, pp. 62–71.
- 19 Rengaraju, P., Chung-Horng, L., Yi, Q. and Srinivasan, A. (2009) Analysis on mobile WiMAX security. In: *Science and Technology for Humanity (TIC-STH), IEEE Toronto International Conference*, pp. 439–444.
- 20 Kasera, S. and Narang, N. (2005) *3G Mobile Networks: Architecture, Protocols and Procedures: Based on 3GPP specifications for UMTS WCDMA networks*. New York: McGraw-Hill.
- 21 Sauter, M. (2005) *From GSM to LTE: An Introduction to Mobile Networks and Mobile Broadbands*. West Sussex, UK: John Wiley & Sons, Ltd.
- 22 Smith, C. and Collins, D. (2002) *3G Wireless Networks*. Boston, MA: McGraw-Hill.
- 23 UMTS Forum. www.umts-forum.org
- 24 Leo, Y., Kai, M. and Liu, A. (2011) A comparative study of WiMAX and LTE as the next generation mobile enterprise network. *Proceedings of the 13th International Conference on Advanced Communication Technology (ICACT)*, pp. 654–658.
- 25 Muxiang, Z. and Yuguang, F. (2005) Security analysis and enhancements of 3GPP authentication and key agreement protocol. *Proceedings of the Wireless Communications, IEEE Transactions*, 4(2), 734–742.
- 26 3GPP, *Overview of 3GPP release 8 v.0.1.1*, Tech. Rep., June 2010.
- 27 3GPP TR 36.913, *Requirements for Further Advancements for E-UTRA*, v8.0.1, March 2009.
- 28 3rd Generation Partnership Program. *Network Architecture*. Technical Specification 23.002. Release 5. Version 5.5.0.
- 29 3rd Generation Partnership Program. *Security Architecture*. Technical Specification 33.102. Release 5. Version 5.2.0.
- 30 3rd Generation Partnership Program. *Cryptographic Algorithm Requirements*. Technical Specification 33.105. Release 4. Version 4.1.0.

2

5G Mobile Networks: Requirements, Enabling Technologies, and Research Activities

Van-Giang Nguyen, Anna Brunstrom, Karl-Johan Grinnemo, and Javid Taheri

Department of Computer Science, Karlstad University, Karlstad, Sweden

2.1 Introduction

In the previous chapter, we have seen the history and evolution of the four generations of mobile cellular systems. After more than three decades of evolution, mobile cellular systems have significantly changed from analog or circuit-based to packet-based communication systems, with also big changes in the speed and bandwidth improvement as well as the number of connected devices. According to a forecast by Ericsson [1], the number of connected devices is growing rapidly and will reach to close to 28 billion by 2021, with around 16 billion Internet of Things (IoT) related devices.

This big change in the demand from users and the emergence of new services, and the increase of the communication needs from vertical industrial sectors such as automotive, agriculture, health, and transport, impose a huge challenge on the current generation of mobile cellular system (i.e. 4G). As a consequence, it requires the development of the next generation mobile system or fifth generation (5G), which is expected to be an ecosystem for every Internet-enabled device. In the following, we will explore more about the vision of the 5G system and its typical use cases to see how it differs from today's 4G mobile networks.

2.1.1 What is 5G?

Since 5G is still in its infancy and the progress on standardizing is still ongoing, there exist many different definitions of what 5G is and what it will be supporting. In 2015, the Next Generation Mobile Network (NGMN) Alliance, an association of more than 80 partners from the mobile telecommunications industry and research, published a white paper [2], where 5G is defined as “an end-to-end ecosystem to enable a full mobile and connected society. It empowers value creation towards customers and partners, through existing and emerging use cases delivered with consistent experiences, and enabled by sustainable business models”. It should be recognized that 5G is not just a one-step evolution from today's 4G network (i.e. LTE 4G and IMT-Advanced), but instead it is a big paradigm shift. According to the 3rd Generation Partnership Project

(3GPP) [3], the 5G system will support most of the existing Evolved Packet System (EPS) services (i.e. 4G related services), in addition to many new services.

2.1.1.1 From a System Architecture Perspective

5G will be a mix of multiple access technologies, which include both current and existing access radio technologies such as LTE, and the New Radio (NR) for 5G [4], as well as the evolution of Wireless Local Area Network (WLAN) technologies. In addition, 5G will integrate most of the current emerging network paradigms such as cloud computing, Software Defined Networking (SDN) [5], and Network Function Virtualization (NFV) [6], into a unified, programmable software-centric system and infrastructure.

2.1.1.2 From the Spectrum Perspective

5G will be allocated with a significant amount of new spectrum to cope with massively connected devices, highly dense networks, and to support a variety of use cases, in which the traditional cellular frequency bands (typically < 6 GHz) might not provide sufficient bandwidth for all 5G applications. Particularly, during the World Radiocommunication Conference in November 2015 (WRC-15), an agenda item was approved for studying and investigating additional spectrum above 6 GHz (from 24 GHz up to 86 GHz) for 5G mobile services, and the results of those studies will be considered at the next WRC in 2019. In addition, WRC-15 also identified a range of new spectrum bands below 6 GHz or sub-6 GHz that can be used for 5G mobile services.

2.1.1.3 From a User and Customer Perspective

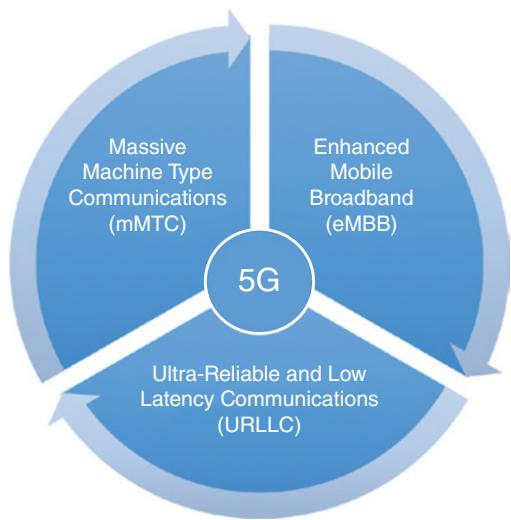
This is when 5G realized that subscribers can, on demand, have access from their devices from anywhere, at any time with high data rates (e.g. 1G bps), very low latency (e.g. at around 1 ms), much improved quality of service and quality of experience. With 5G, they can experience many newly defined service types such as virtual reality, augmented reality for gaming or other purposes, ultra-high definition and 3-dimensional videos, and autonomous driving, etc. In the following, some typical 5G use cases will be described in detail.

2.1.2 Typical Use Cases

So far, there are a large number of use cases, which are considered as main drivers for the upcoming 5G ecosystem. These use cases come from different projects, organizations, and industrial vertical sectors, based up on their visions and their needs. However, in general, there are primarily three main categories of use cases, which have been agreed by most of the standardization groups, including the International Telecommunication Union (ITU) and 3GPP. These include enhanced mobile broadband (eMBB), massive machine type communications (mMTC), and ultra-reliable and low latency communications (URLLC), as shown in Figure 2.1:

- *Enhanced mobile broadband:* This usage scenario refers to the improvement of the current mobile broadband services, which are typically human-centric, in terms of user experienced data rates, traffic volume, coverage, and seamless mobility compared to services delivered in today's system. This scenario covers both wide and dense (e.g. hotspot) area coverages, which have different requirements. Some typical examples

Figure 2.1 Three main use case categories as defined by ITU in [7] and 3GPP in [8].



of 5G services in this category are ultra-high definition video (e.g. 4 K/8 K), virtual reality, augmented reality, virtual presence, etc.

- *Massive machine type communications:* This usage scenario relates to deployments of a large number of connected devices, which typically transmit relatively small amount of data such as sensors and utility meters. These devices are required to be low-cost and have a long battery life. Some typical examples of 5G services in this category are inventory control, smart city, smart metering, video surveillance, etc.
- *Ultra-reliable and low latency communications:* This usage scenario is about the capability to provide a given service with stringent requirements in terms of ultra-low latency, ultra-high reliability and high availability, as well as high throughput. Some typical examples of 5G services in this category are autonomous driving cars, smart grids, eHealth, Tactile Internet, remote surgery, industrial automation and control, etc.

The rest of the chapter will be organized into two main parts. In the first half, typical technical requirements that need to be met when 5G is realized are identified, followed by a description of technologies, which are considered as key enablers for 5G. The second part summarizes ongoing 5G activities, which are recently developing from global, regional standardization organizations to open research forums and research communities worldwide.

2.2 5G Requirements

A huge number of requirements for 5G have been identified by different organizations, companies, and research communities. The requirements emanate from end user's experience, system performance, services, and operation and management. For example, ITU-R has identified some key requirements for 5G such as peak data rate, mobility, connection density, network and spectrum efficiency, etc. [7]. Figure 2.2 summarizes key requirements of 5G and some example values for each requirement. In principle, 5G requirements can be seen from a user performance perspective such as data rate,

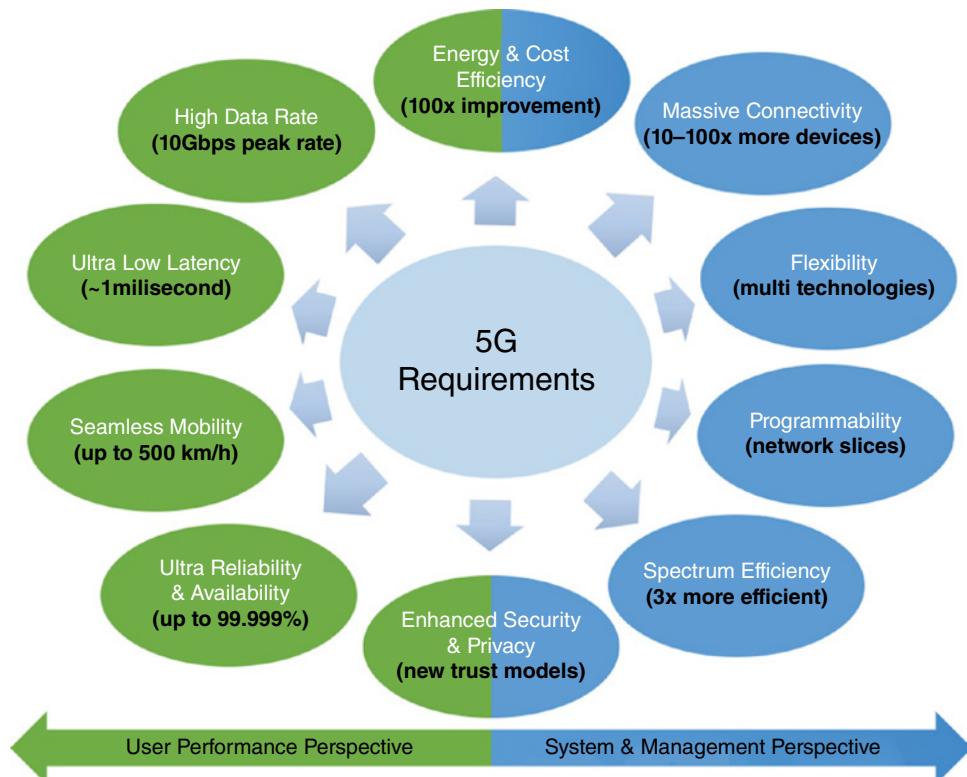


Figure 2.2 5G key requirements and some example values.

latency, mobility, reliability, etc. and from networking and system management perspective such as connection density, network flexibility, energy and cost efficiency, etc. In the following, these key perspectives will be discussed in detail.

2.2.1 High Data Rate and Ultra Low Latency

Data rate and latency are two major evaluation metrics to assess the user quality of experience in wireless communication systems. When it comes to the development of the next generation system (i.e. 5G), these two metrics are key to satisfy the user quality of experience.

The requirements on the data rate are expressed in terms of the peak data rate, which is the maximum achievable data rate for a user under ideal conditions, and the user experienced data rate, which is the achievable data rate for a user in the real network environment. Currently, 4G networks are offering users the maximum peak data rate of 1G bps, while the maximum data rate experienced by the user is around 10 Mbps [7]. However, the introduction of new bandwidth-hungry applications and services such as virtual reality, ultra-high definition video streaming (e.g. 4 K and 8 K), demands an extreme improvement of 5G networks compared to current 4G networks. According to [7], the peak data rate is expected to be enhanced by up to 20 Gbps, while the user experienced data rate will give 100 times improvement over 4G networks and reach up to 1 Gbps.

Another fundamental requirement is latency, which is typically expressed as end-to-end latency perceived by the end user. With the advent of newly defined services such as Tactile Internet, self-driving car, and automatic traffic control, which require real-time responses and interactions, minimizing the latency is becoming more crucial. More specifically, the 5G system is expected to reduce the latency ten times in the user plane, down to 1 millisecond, and half in the control plane, down to 50 milliseconds, compared to the 4G system [9].

2.2.2 Massive Connectivity and Seamless Mobility

Massive connectivity refers to the requirement of supporting a large number of connected devices and consequently a large or massive number of connections in an area unit (e.g. connection per square kilometer). In the 5G era, the increased number of devices in the network is not only coming from the emergence of new types of services and new types of devices such as sensors, meters, wearable devices, and vehicles, but it also from the exponential increase in the number of existing device types such as smart phones and tablets. Due to the proliferation of these smart things, the 5G system is expected to support a connection density of up to 1 billion connected devices per square kilometer, or put differently, 100 times more devices compared to the 4G system. In addition to the number of connected devices, the network densification can also be reflected by the traffic density, which is measured by the total amount of traffic exchanged by all devices over the considered area [2]. The expected value for this metric in the 5G era is tens of Gbps per square kilometer.

Besides the requirement to support a massive number of connected devices, the 5G system is also expected to provide seamless service experience to mobile users. However, not all devices and users in the 5G era are mobile, thus seamless mobility is not necessary. Therefore, on-demand mobility solutions should be supported, depending on the types of devices and services [2]. For example, it is expected to enable an acceptable service experience for mobile devices moving at speeds up to 500 km/h.

2.2.3 Reliability and High Availability

Reliability and high availability are two other important requirements that need to be guaranteed in the 5G system. In general, the reliability of a system refers to its capability of guaranteeing the success rate of data transmission under stated conditions (e.g. a latency budget) over a certain period of time. Depending on different use cases and services, the reliability rate will vary. As described previously, there are a number of services and applications in the third usage scenario (i.e., ultra-reliable and low latency communications), such as public safety, eHealth, automatic traffic control, and mission critical services, which require extremely high reliability for the communication. In order to support these kinds of services, the 5G system is expected to guarantee a reliability rate of up to 99.999%.

In order to provide services to end users anywhere, at any time, the 5G system must ensure its availability, which refers to the ability to endure against possible outage scenarios. The availability is usually expressed as a percentage of uptime in a given period of time (e.g. a year) and assessed based on the number of nines in the digits (e.g. 99.99%). The 5G system should guarantee the availability rate with as many nines as possible, for example, five nines or 99.999%.

2.2.4 Flexibility and Programmability

Flexibility and programmability are two network-driven requirements. Since the 5G system will be the integration of multiple technologies to support a large number of devices and services, its network architecture should be flexible in order to satisfy a range of different requirements in connection to properties, and attributes exposed by those devices and services. The flexibility of the network can be expressed in terms of the ability to support various types of radio access technologies, the capability of scaling the network resources on demand and independently between the radio access network and the core network, as well as between the control plane and the data/user plane, the capability of installing new services and applications in a very short time, and the capability of re-shaping the network infrastructure in real time to adapt to a change in the user or customer demands.

In addition, the network infrastructure in 5G should be programmable and reconfigurable. Indeed, the 5G network infrastructure will be constructed as a set of different logical virtualized networks or “slices” over the same physical infrastructure. The programmability of the network allows on-demand and autonomic networking, where mobile operators can define, program and configure their own network “slices” according to their policies and their defined use cases.

2.2.5 Energy, Cost and Spectrum Efficiency

Along with the enhancement of the network capacity and the improvement of user experience, energy and cost efficiency must be taken into account in the design of 5G. In particular, the 5G system is expected to have a 100-fold improvement on energy efficiency compared to today’s 4G system. Meanwhile, the cost efficiency, which represents the economical aspect of the 5G system, must be increased in order to guarantee mobile operator’s revenue.

In addition, as described previously, 5G will be fueled by not only the human-centric devices such as smartphones, but also the massive number of “things”, such as sensors, smart meters, etc. These things are required to have much longer battery life in order to operate in the field, without any further power supply, for example at least 10-year battery lifetime.

Last but not least, spectrum efficiency should also be significantly improved compared to today’s 4G system. For example, it should be three times higher for enhanced mobile broadband [7].

2.2.6 Security and Privacy

Apart from the aforementioned requirements, security is another important aspect that needs to be taken into account in the development of 5G. Indeed, the introduction of diversified services and devices will pose many challenges for guaranteeing the security in the 5G era. More specifically, the security for 5G will need to be guaranteed in different levels, including access level, infrastructure level, and service level. For example, at the infrastructure level, being enabled by a variety of technologies such as SDN, NFV, network slicing, etc., 5G network infrastructure will be more open and programmable, thus driving new security requirements such as how to securely

guarantee the communication channel between the control and data planes as SDN is adopted, and how to isolate and manage network slices securely as network slicing is enabled, etc. At the service level, there will be many kinds of newly defined 5G services, which will require different security levels. For example, some critical services, such as public safety and eHealth, require more security than other services, such as virtual reality, augmented reality, etc. In addition, the involvement of new actors, and the introduction of new business models, will derive new service delivery models and new trust models. For example, in the current system, the trust model is mostly formed between users and mobile operators. However, more actors such as vertical service providers will play an important role in defining new trust models in 5G. The careful design of these models, together with guaranteeing the aforementioned requirements, is necessary to provide an end-to-end secured communication channel for users in the 5G ecosystem.

Last but not least, privacy concerns also need to be considered in the development of 5G. Indeed, 5G networks will accommodate massive numbers of user devices, meaning that a great amount of user privacy information such as user identifiers will be carried over the 5G network. In addition, new types of device identities such as identifiers for IoT devices will also be introduced in 5G. Therefore, it requires an efficient way to manage this massive amount of information as well as to protect and prevent the leakage of user personal information. Research activities and proposals on 5G related security and privacy will be discussed in detail in other chapters.

2.3 5G Enabling Technologies

In order to meet the strict requirements discussed previously, a number of technology candidates have been considered and widely discussed. Table 2.1 provides the mapping between the 5G requirements and corresponding potential technology enablers. The development in technology will happen both in the radio access network (RAN), the core network, and the end-to-end system. In the following, these developments will be described in detail.

Table 2.1 5G key requirements and corresponding technology candidates.

| Requirements | Technology Candidates |
|-----------------------------------|--------------------------------------------|
| High Data Rate | mmWave, Massive MIMO, Small Cell |
| Ultra-Low Latency | Mobile Edge/Fog Computing, D2D |
| Massive Connectivity | Massive MIMO, D2D, M2M, Small Cell |
| Reliability and High Availability | Cloud-RAN, SDN, NFV, MANO, Cloud Computing |
| Flexibility and Programmability | Cloud-RAN, SDN, NFV, Network Slicing, MANO |
| Energy and Cost Efficiency | Cloud-RAN, SDN, NFV, Network Slicing, MANO |
| Spectrum Efficiency | Massive MIMO, Small Cell, D2D |
| Security and Privacy | See Chapters 5–7 |

2.3.1 5G Radio Access Network

The enabling technologies for 5G RAN include mmWave communication, massive Multiple Input Multiple Output (MIMO), ultra-dense small cell, Machine-to-Machine (M2M) and Device-to-Device (D2D) communications, cloud-RAN, and mobile edge and fog computing. These technologies will be described in the following.

2.3.1.1 mmWave Communication

As mentioned previously, one of the key features of the 5G system is to have higher capacity in terms of data rate, for example, up to tens of Gbps at peak data rate. In order to achieve those targets, more spectrum availability is required. However, current wireless systems are typically operating in a spectrum band, ranging from hundreds of MHz (e.g. 700 MHz) to below 3 GHz (e.g. 2.6 GHz). These spectrum usages are not sufficient enough for 5G. One of the most effective solutions for expanding the bandwidth range is to exploit the very high spectrum bands, which have not been occupied yet (e.g. > 10 GHz). In particular, during the meeting at the WRC-15 conference hosted by ITU, several proposed frequency bands above 10 GHz for 5G have been approved to be studied ahead of the next WRC conference in 2019 [10], for example, 24.25–27.5 GHz, 50.4–52.6 GHz, 81–86 GHz, etc. In this sense, millimeter wave communication (mmWave) [12] is the best technology candidate.

The mmWave research was first conducted by Jagadis Chandra Bose in 1897, which refers to the use of frequencies in the range of 30 to 300 GHz, with the corresponding wavelengths in between 10 mm and 1 mm, as shown in Figure 2.3. Due to some reasons, such as high propagation loss, the mmWave communication was commonly used for indoor environments or backhaul links. However, many research initiatives have illustrated the feasibility of mmWave technology for 5G mobile networks by adopting many recent advances in propagation modeling [11] or channel modeling [12], to create a larger amount of bandwidth. Apart from the benefits of allowing larger bandwidth, higher data rate that makes the mmWave a promising technology for 5G, there are still a number of challenges and open issues that need to be solved in the future, such as interference and heterogeneity [13].

2.3.1.2 Massive MIMO

In order to meet the 5G requirements in terms of network density and capacity enhancement, one of the most prominent solutions is to densify the number of deployed antennas, which refers to a technical solution called massive MIMO. Fundamentally,

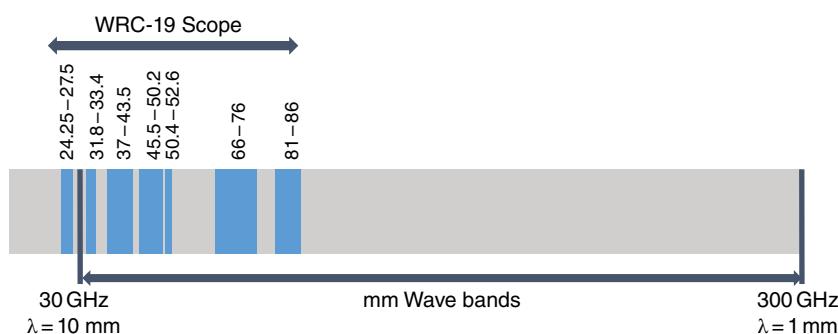


Figure 2.3 Millimeter-wave bands and potential 5G bands to be studied ahead of WRC-19.

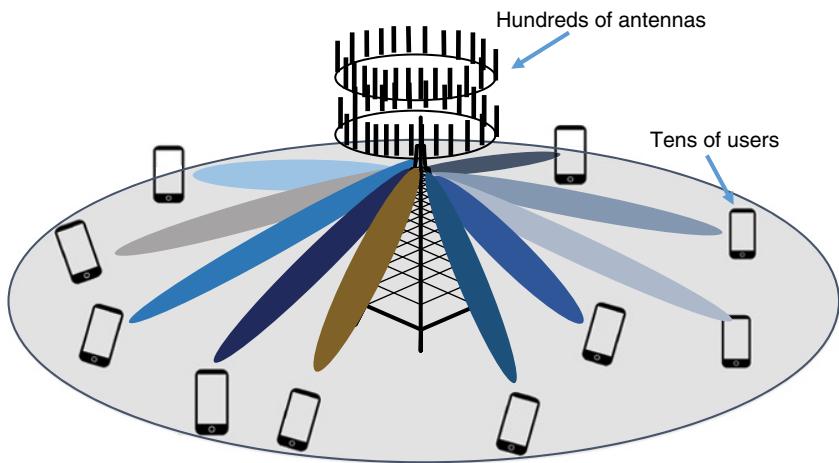


Figure 2.4 An illustration of massive MIMO concept.

MIMO is an antenna technology for wireless communications in which multiple antennas are used to transmit and receive data. In fact, the MIMO concept has been commonly utilized in current 4G networks, which refers to multi-user MIMO communication [14], where several users are simultaneously served by a multiple-antenna base station; whereas, massive MIMO is defined as a multi-user MIMO system, where the number of base station's antennas and the number of users are large [15]. Such a feature as having more antennas at the base station promises to increase the network capacity and density. More importantly, massive MIMO is said to significantly enhance spectral and energy efficiency [15]. These reasons make massive MIMO an essential technology for 5G [16]. Figure 2.4 depicts the concept of massive MIMO. Apart from the benefits of massive MIMO, there are still several research questions that need to be addressed, such as mitigation of pilot contamination, channel estimation, implementation-aware algorithmic design, etc.

2.3.1.3 Ultra-Dense Small Cells

Another way of increasing the network density and improving the throughput is to densify the number of wireless nodes, which have a smaller coverage range than the macro-cell base stations used in the 3G and 4G legacy systems. The technical solution behind this idea is denoted as the small cell technology. As defined by the Small Cell Forum, the “small cells” is an umbrella term for operator-controlled, low-powered radio access nodes with a coverage range in between ten to several hundreds of meters, including those that operate in licensed spectrum and unlicensed carrier-grade WiFi. An example of small cell deployment is shown in Figure 2.5.

With small cells, the size of the cell is reduced, meaning they bring the network much closer to the user, thus better serving high traffic areas such as indoor and hotspot areas. In addition, the higher number of low-powered transmission points on the small cell network enables better use of available frequency resource, thus improving the spectral efficiency. Furthermore, the 5G system will be constructed in a heterogeneous fashion, where macro and small cells are co-located and maybe connected to each other

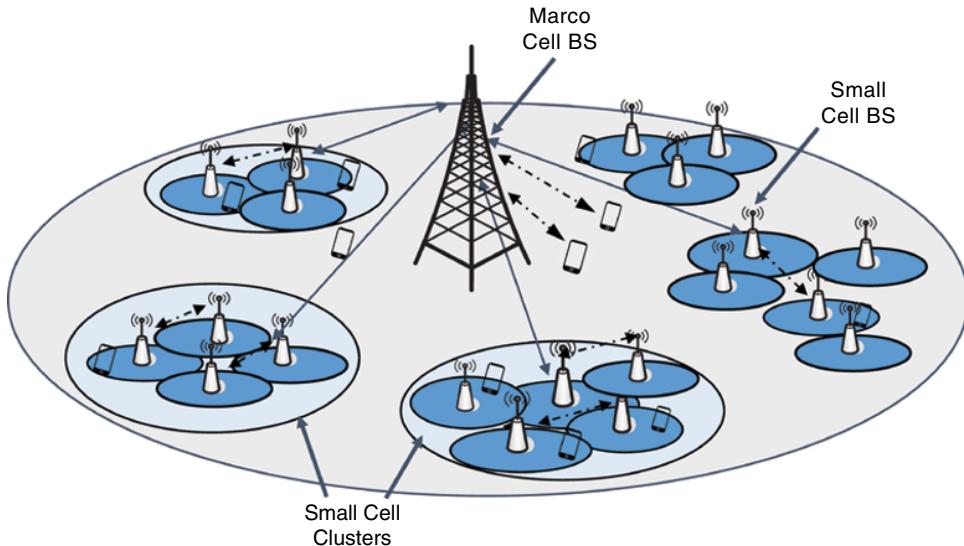


Figure 2.5 An illustration of small cells deployment.

via wireless backhaul links, thus providing increased levels of network capacity through traffic offloading. However, the heterogeneity of small cells in the network will pose challenges in terms of interference and mobility management, thus affecting system performance as a whole. These issues will need to be addressed in future research. Some other ongoing research on small cells for 5G would include load balancing, wireless backhauling, mmWave and massive MIMO in small cells, etc. [17].

2.3.1.4 M2M and D2D Communications

a) *M2M Communication*

As mentioned previously, two-thirds of the use case categories of 5G will be related to the IoT and Machine Type Communication (MTC), including massive and critical communications. Therefore, although the concept of M2M or MTC communication was introduced in 4G LTE systems by 3GPP some time ago [18], it is still considered as one of the key enablers for 5G [19]. Fundamentally, M2M communication refers to the automated data communications among devices and the underlying data transport infrastructure [20]. The data communications may occur between an MTC device and a server, or directly between two MTC devices. There are a number of services and applications enabled by M2M communication, such as monitoring and metering, home and industry automation, health care, and automotive, as shown in Figure 2.6 (a). Several open issues and challenges need to be resolved in future M2M related research, such as scalability, security and privacy, energy efficiency, etc.

b) *D2D Communication*

D2D communication [21] refers to direct communication between two mobile users/devices, without traversing through a network infrastructure. It has been specified by 3GPP in LTE Release 12. Exploiting direct communication between devices, D2D communication can help improve spectrum efficiency, user data rate gain, and reduce latency as well as energy consumption, thus being considered as

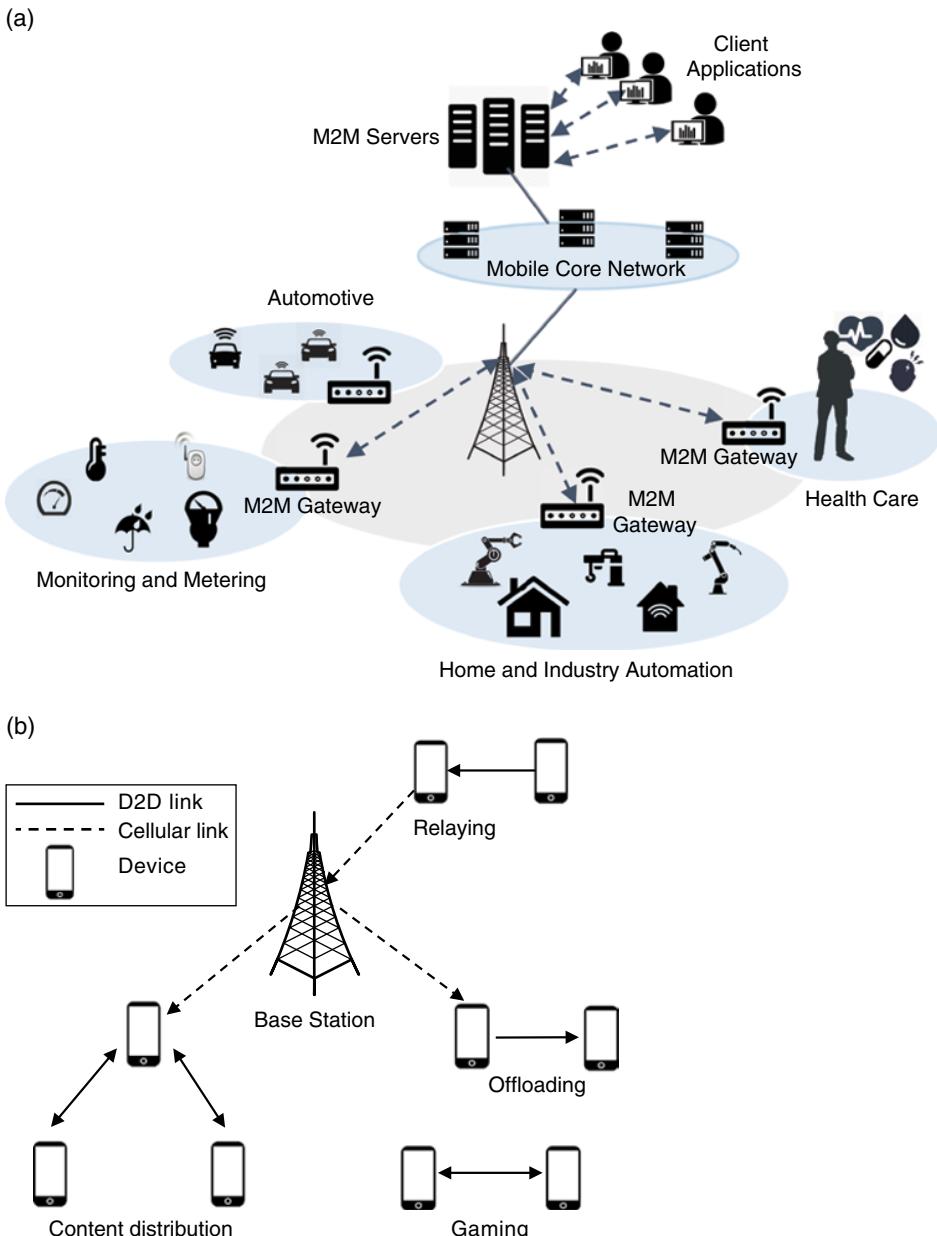


Figure 2.6 (a) M2M communications and use case examples; and (b) D2D communications and use case examples.

one of the key components of the 5G system [22]. In general, the operation of D2D communication can be in-band D2D on licensed cellular spectrum (e.g. LTE) and out-of-band D2D on unlicensed spectrum (e.g. WiFi). There are a number of use cases and application scenarios for D2D, such as proximity-based services,

gaming, public safety, vehicular communications, and offloading, as shown in Figure 2.6(b). However, there are a number of open issues that need to be solved in the future, such as interference management, resource management services and device discovery, security and privacy. Some other directions for future research would be the integration of D2D communication with mmWave and massive MIMO technologies.

2.3.1.5 Cloud-based Radio Access Network

Cloud-based Radio Access Network (Cloud-RAN) is an ideal solution to design the radio access part of 5G networks, since it enables energy efficiency, cost savings on baseband resources, as well as improvements on network capacity, increased throughput, etc. [23]. The Cloud-RAN is essentially the decoupling of the Remote Radio Head (RRH) from the Baseband Unit (BU) of a base station, and the implementation of BU in a centralized cloud computing environment. RRHs are connected to a BBU pool by using high speed fiber or microwave-link fronthaul networks. In fact, there are several options to split the functionality of the base station, which refers to the RAN-as-a-Service (RANaaS) [24]. These two concepts are shown in Figures 2.7(a) and (b). This simplified base station architecture is paving the way for dense 5G deployment by making it affordable, flexible and efficient [25].

Apart from the benefits that Cloud-RAN offers to the design of the 5G system, there are various challenges that need to be overcome before fully exploiting its benefits; such as fronthaul constraints and performance optimization, placement optimization of RRHs, efficient scheduling and elastic scaling of BBUs in the BBU pool. Some other research directions in the future could be the incorporation of C-RAN and distributed RAN (D-RAN) or research on heterogeneous CRAN (H-CRAN).

2.3.1.6 Mobile Edge and Fog Computing

As we move to 5G, many of its services and applications will require very stringent latency in the order of milliseconds. One of the most prominent solutions is to bring the IT services and processing capabilities down to the edge of the mobile network, within the RAN and in close proximity to mobile users. This refers to the concept of Mobile Edge Computing (MEC) technology and its sibling Fog Computing. Figure 2.8 illustrates the concept of MEC and its architecture. As specified in the ETSI white paper [26] published 2015, the aim of MEC is to reduce latency, ensure highly efficient network operation and service delivery, and offer an improved user experience. With this capability, MEC will open new frontiers for network operators, application service providers, and content providers, by enabling them to introduce innovative services and applications. Some typical examples of services enabled by MEC are augmented reality, RAN-aware video optimization, connected cars, and IoT, etc. [26].

A similar concept to MEC is Fog Computing (FC) [27] defined by Cisco in 2012, a paradigm in which cloud computing resources are extended to the edge of the network, to create a highly virtualized platform that provides compute, storage, and networking services between end-devices and traditional data centers. Some of the prominent features of FC, which are suitable for 5G communications, are low latency, location

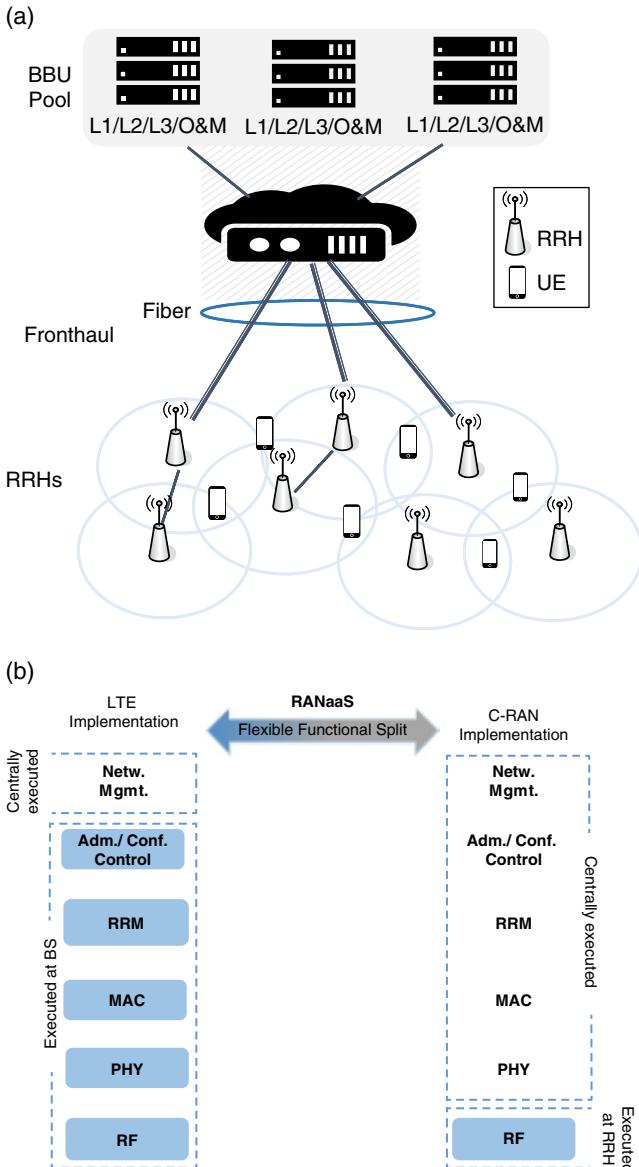


Figure 2.7 (a) Cloud-RAN concept, adapted from [23]; and (b) RANaaS concept, adapted from [24].

awareness, real-time interactions, mobility support, geographical distribution, and the predominance of wireless access [27].

Although both MEC and FC are extremely fit for the development of the 5G system, there are several research issues that need to be further explored, such as the interworking between edge clouds, between edge clouds and centralized clouds, mobility management to allow users to seamlessly access edge applications, and other open challenges such as security and performance.

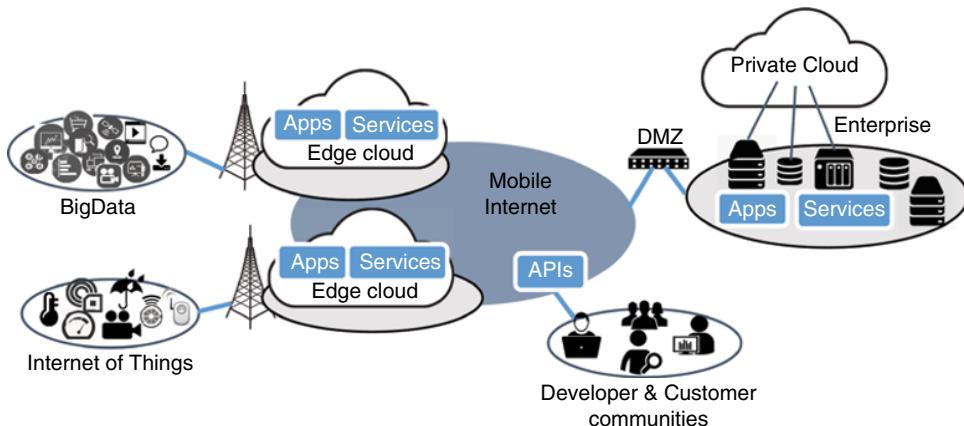


Figure 2.8 Mobile Edge Computing architecture (adapted from [26]).

2.3.2 5G Mobile Core Network

SDN, NFV, and cloud computing are considered as key technologies to design the core part of 5G networks. These technologies are described as follows:

2.3.2.1 Software Defined Networking

In terms of network flexibility and programmability, SDN [5] is widely recognized as the best technology candidate for the development of 5G networks. The concept of SDN was first proposed in the campus and data center network areas. It features the separation of the data plane from the control plane, and facilitates the network management through the abstraction of network control functionalities, as shown in Figure 2.9 (a). Being adopted by 5G, SDN will enable a more agile and flexible core network architecture. In addition, programmability and openness characteristics of SDN will help mobile operators shorten the life cycle of introducing their new services and innovation into markets. By separating the control and data planes, the network infrastructure can be constructed on demand and on the basis of service requirements (network-as-a-service), thus improving the resource efficiency. It is worth to note that the SDN concept can also be used in the RAN domain, where the SDN controller could control and schedule the radio resources for base stations, thus improving the spectrum efficiency as well as mobility management.

However, there are still many challenges and issues with SDN that need to be addressed, such as the scalability problem due to the centralization of network intelligence, extra latency between devices and the SDN controller, security problem of the communication channel between the control and data planes, the lack of standardization on designing the protocol communicating between the control and data planes, policy and charging enforcement. Other aspects related to the adoption of SDN into mobile networks, such as placement problem of SDN controller, mobility management, load balancing, etc., have been detailed in [28].

2.3.2.2 Network Function Virtualization

As described previously, the 5G system is not just about high data rate, low latency, and flexibility. It is also about cost efficiency, which will have impact on the revenue of mobile operators. 5G mobile operators will expect the cost for the deployment, which refers to

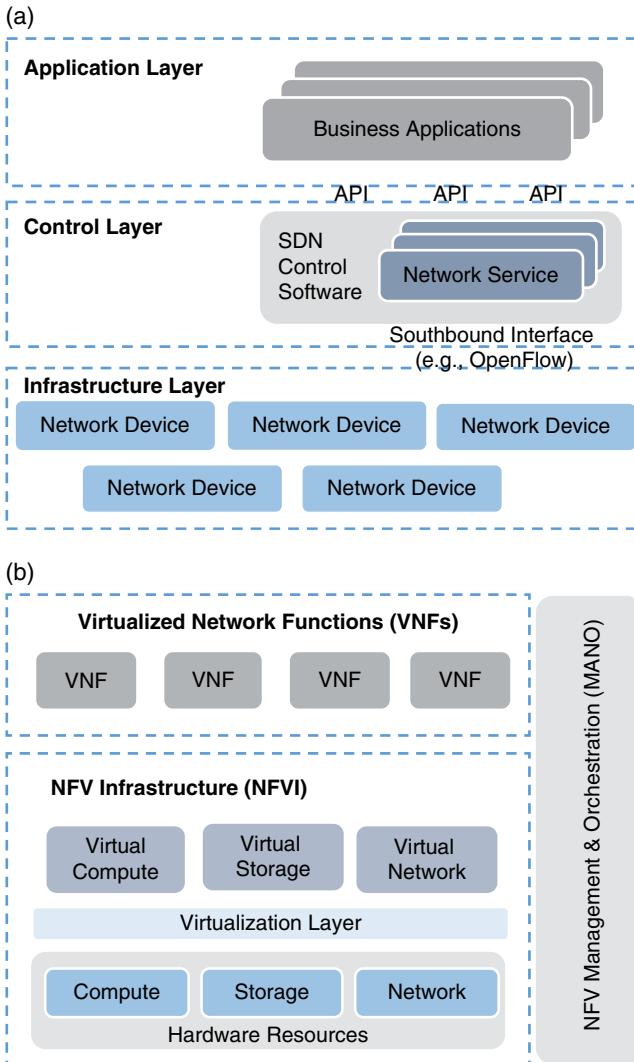


Figure 2.9 (a) SDN architecture, adapted from [5]; and (b) NFV architecture adapted from [6].

capital expense or CAPEX, and the cost for operation and management, which refers to operational expense or OPEX, to be as low as possible. NFV [6] is the foundation for these capabilities and is identified as the cornerstone of 5G core network solutions. Figure 2.9(b) shows the reference architectural framework of NFV. Essentially, NFV refers to the relocation of network functions, which are traditionally implemented on dedicated costly hardware platforms to software appliances running in the cloud environment or on general-purpose commodity servers. By operating the network function as software, it is easier for mobile operators to dynamically scale the resources (computing, storage, and networking) according to changes in traffic demands, and to faster time-to-market of new services. In addition, the combination of SDN and NFV has encouraged the development of new networking paradigms, such as network service chaining, and network slicing.

Although NFV has been proven as the key enabler for the development of 5G, especially the core part, there are many challenges that need to be further studied in future work, such as optimization of network functions placement, resource allocation, management and orchestration, and network performance [29].

2.3.2.3 Cloud Computing

As described in the previous section, cloud computing has been considered as an ideal solution for re-designing the current RAN architecture. With its anticipated benefits, such as on-demand and elastic provisioning of services and resources over the Internet, cloud computing has made it as one of the key enablers for designing 5G core networks. In this case, 5G core network functions will be realized as virtual machines or containers controlled by the cloud manager. The capability of providing resources in a multi-tenant model of cloud computing allows mobile operators to implement the concept of mobile virtual network operators (MVNO) much more easily than in the past. In addition, a pay-as-you-use business model and the ability to move and consolidate the resources offered by cloud computing, can help the mobile operators reduce their capital and optimize operational expenses. Compared to NFV, cloud computing was born to virtualize the commodity IT hardware, while NFV refers to the inspiration of cloud computing to virtualize network functions. In fact, in the recent development of NFV, many cloud technologies such as OpenStack¹ or VMware² are serving as the resource backend for virtual network functions.

Centralizing all resources will ease the management and provisioning process, but it will result in the long delay for end-to-end communication, which may not be suitable for some of the newly defined 5G services. Therefore, combining cloud computing and other computing paradigms, such as mobile cloud computing, fog and edge computing, will be a promising direction to investigate in future research.

2.3.3 5G End-to-End System

The key technology enablers for constructing a 5G end-to-end system include network slicing, and management and orchestration.

2.3.3.1 Network Slicing

Today's 4G system have been optimized mostly for serving human-to-human communication where mobile phones are the main players. However, in the future, the 5G system is expected to support diverse services and applications with various characteristics and requirements, where IoT devices will become dominant. Such IoT-related services will require different types of features and network capabilities in terms of latency, data rate, mobility, reliability, security, etc. Therefore, in order to guarantee these requirements and improve the network performance and resource utilization, each service type should be provided as an end-to-end, isolated, and infrastructural environment to operate on. In this sense, the network slicing concept will become the foundation of all capabilities.

¹ OpenStack. <https://www.openstack.org/>

² VMware. <http://www.vmware.com/>

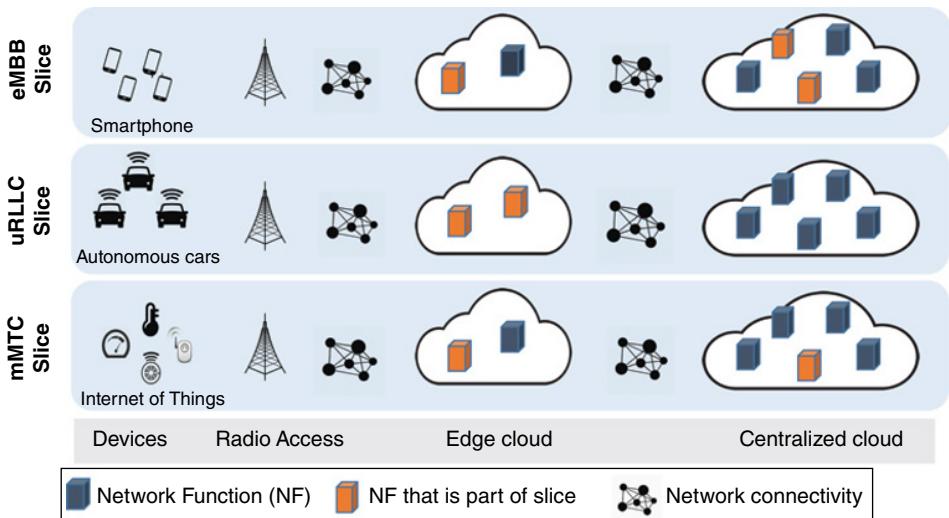


Figure 2.10 An example of network slicing.

Although network slicing has been widely recognized as the key characteristic of 5G by many network operators and vendors, it has not yet been standardized, thus there are variants of defining what slicing is. According to ITU-T in [30], slicing is the basic concept of the Network Softwarization. It allows logically isolated network partitions (LINP), with a slice being considered as a unit of programmable resources such as network, computation and storage. As NGMN defined in its white paper [2], a network slice, namely “5G slice”, is composed of a collection of 5G network functions and specific radio access technology settings that are combined together for the specific use a case or business model. Figure 2.10 illustrates an example of the network slicing concept, with three different slices corresponding to the three main 5G use case categories discussed in Section 2.1. To this end, the implementation of the network slicing concept is on an end-to-end basis. The 5G system will be composed of multiple end-to-end slices, where dedicated resources and quality of service (QoS) are guaranteed.

However, network slicing still presents many challenges and gaps that must be fulfilled in future studies such as slice definition, lifecycle management of a slice, resiliency of slice control, resource allocation and optimization within a slice and between slices, strongly guaranteeing security within a slice and between slices, end-to-end QoS management, integrating with other technologies (e.g. information centric networking (ICN), D2D), etc.

2.3.3.2 Management and Orchestration

When the 5G mobile networking era arrives in the next few years, due to the diversity of use cases, services, and the number of network slices created with different resource requirements, the management and orchestration (MANO) of the network becomes more and more crucial. The role of MANO will be managing the whole network infrastructure in terms of fault management, configuration, accounting, performance, and security. More importantly, MANO will be in charge of lifecycle management and

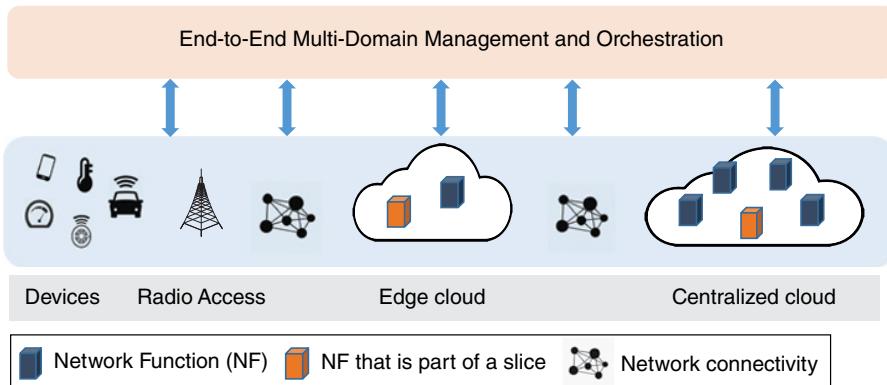


Figure 2.11 The illustration of the end-to-end multi-domain management and orchestration.

provisioning the network resources for the end-to-end connectivity of network slices in a dynamic, automated, and efficient manner. As illustrated in Figure 2.11, the role of the end-to-end management and orchestration will be multi-domain, multi-operators, and multi-technology spanning from the infrastructure layer to the application (service) layer, and spanning from the RAN to the core of the network.

In 2014, the NFV MANO working group in the European Telecommunications Standards Institute (ETSI) has specified in its technical specification [31], an architectural framework showing main components and their functionalities, as well as the operations within the framework. In the meantime, there are several efforts that have implemented the NFV MANO concept as open source platforms, such as OpenStack Tacker, OpenBaton, OSM, Open-O, etc. [30]. These open source platforms are being applied to today's 4G core network and will be integrated into the deployment of the 5G network architecture.

However, due to the diversity of network resources from RAN to the core, the current NFV MANO framework should be extended in order to manage not only virtualized network functions and resources, but also physical nodes. In addition, dynamically managing and orchestrating the network services and slices would also be challenging.

2.4 5G Standardization Activities

A number of research activities are currently being carried out within standardization organizations, such as International Telecommunication Union (ITU), 3GPP, ETSI, the Institute of Electrical and Electronic Engineers (IEEE), and the Internet Engineering Task Force (IETF). In the following, the status of ongoing activates within these organizations will be summarized.

2.4.1 ITU Activities

After success in developing international standards and specifications for the previous International Mobile Telecommunications (IMT) generations, including IMT-2000 (i.e. 3G) and IMT-Advanced (i.e. 4G), ITU continues its leading role to outline the concept and vision for 5G, including identifying key parameters, with a target date set for 2020, thus it is

called IMT-2020. Currently, the 5G related activities within ITU are happening in two of its three main sectors: ITU Radiocommunication sector (ITU-R), which focuses on wireless and radio system aspects, and ITU Telecommunication standardization sector (ITU-T), which focuses on wireline and network architecture aspects.

2.4.1.1 ITU-R

In 2012, ITU-R, having the leading role of Working Party (WP) 5D, initiated work to develop the next generation IMT for the year 2020 and beyond, or IMT-2020. As initial results from the work, in September 2015, ITU-R released a recommendation ITU-R M.2083-0 [7], which defines the framework and overall objectives, including the description of key usage scenarios, key capabilities, technology trends, spectrum implications, and timelines for the development of IMT-2020. By June 2017, several deliverables are expected to be completed at the ITU-R WP 5D 27th meeting, including reports on the technical performance requirements, evaluation criteria and evaluation methods, specification submission requirements, and a circular letter.

From the spectrum regulation perspective, ITU-R published a report in July 2015, which provides information on the study of technical feasibility of IMT in the bands above 6 GHz. As already discussed, the WRC-15 hosted by ITU-R has come up with several agenda items for investigating additional spectrum above 24 GHz, which will be finalized in the WRC-19.

2.4.1.2 ITU-T

In May 2015, ITU-T started working on IMT-2020 by establishing a new Focus Group (FG IMT-2020) within the ITU-T Study Group 13, to identify the network standardization requirements for the 5G development of IMT for 2020 and beyond. The main objectives of FG IMT-2020 are to identify the needs for standardization of the “wireline elements” of 5G networks, to be a launching point for ITU-T’s contribution to IMT-2020 standardization, and to align deliverables with those of ITU-R. As the initial results from the first phase, FG IMT-2020 has completed a report on gap analysis in December 2015 [32]. The report identifies 85 gaps, such as high-level network architecture, network softwarization, etc. In 2016, FG IMT-2020 was restructured into four working groups (WG), including Architecture and Framework WG, Network Softwarization WG, Information Centric Networking WG, and End-to-End Network Management WG. Also, in 2016, several proofs of concepts were demonstrated, such as end-to-end network slicing.

2.4.2 3GPP Activities

Overall, the 5G research activities within the 3GPP can be divided into three phases: Pre-5G phase, 5G phase I, and 5G phase II, happening at both 3GPP RAN and Service and System Aspect (SA) Technical Specification Groups (TSG). The outputs of these phases will appear in technical specifications Release 14, Release 15, and Release 16, respectively. The status of ongoing 5G related activities in 3GPP is summarized below.

2.4.2.1 Pre-5G Phase

The Pre-5G phase is also known as the study phase, which started in early 2015. At that time, the 3GPP Services WG within TSG SA (SA1) initiated a Study Item (SI) on New Services and Markets Technology Enablers (SMARTER), with the objectives of developing

high-level use cases and identifying the related high-level potential requirements for 5G. According to [8], use cases within the SMARTER study are grouped into four main building blocks, including Enhanced Mobile Broadband (eMBB), Massive Internet of Things (mIoT), Critical Communications (CriC), and Network Operation (NeO). The System Architecture WG (SA2) also initiated several feasibility studies on the enhancement of the existing architecture, as well as a study item on designing a new system architecture for the next generation mobile networks [33]. In addition, the SA Telecom Management WG (SA5) completed a study item on the NFV concept in the management of mobile networks.

From a RAN perspective, TSG RAN has completed a study on channel model for frequency spectrum above 6GHz. In September 2015, 3GPP organized a workshop on RAN for 5G and established a study item on the requirements and scope of the new radio technology. All results from the study items by SA and RAN TSGs will be inputs for the technical specification Release 14, which will be completed in June 2017.

2.4.2.2 5G Phase I

After finishing the pre-5G phase with a lot of proposed study items, this phase will start the first phase of standardization of 5G related solutions, in which work items are proposed for approval. The 5G phase I will focus on the use case building block of enhancing mobile broadband, as well as low latency and high reliability. The concept of New Radio (NR) will be standardized from this phase, where the frequency range will be below 6GHz and above 6GHz. It will also address the operation of standalone and non-standalone NR. From the SA perspective, the requirements from the SMARTER study and study items on SA2 will be followed in the 5G phase I. This phase will be completed in June 2018, with the release of Release 15.

2.4.2.3 5G Phase II

The 5G phase II will be optimized, not only for enhanced mobile broadband, but for all 5G use cases. The NR will be further studied and standardized with new requirements and features, such as the operation at higher frequency band using mmWave technology. The SA work items will be further enhanced and improved, including security aspects from the SA Security WG (SA3). Although the concrete work item for the second phase has not been decided upon yet, the results from this phase should be ready by December 2019 for the IMT-2020 submission and addressing all identified use cases and requirements.

2.4.3 ETSI Activities

There are several industry specification groups (ISGs) and working groups within ETSI related to the development of 5G, such as NFV ISG³, MEC ISG⁴, mmWave Transmissions (mWT) ISG⁵, and Next Generation Protocols (NGP) ISG⁶.

ESTI NFV ISG was founded in 2012 with the objective to develop the required standards for NFV. So far, this group has published three white papers providing information, including a reference architectural framework, research activities and progress, scope and

³ ETSI NFV ISG. <http://www.etsi.org/technologies-clusters/technologies/nfv>

⁴ ETSI MEC ISG. <http://www.etsi.org/technologies-clusters/technologies/mobile-edge-computing>

⁵ ETSI mWT. <http://www.etsi.org/technologies-clusters/technologies/millimetre-wave-transmission>

⁶ ETSI NGP. <http://www.etsi.org/technologies-clusters/technologies/next-generation-protocols>

perspective for the future action. Currently, ETSI NFV ISG has been working on different aspects of NFV, including architectural models, management and orchestration, software architecture, and reliability and availability. In November 2015, they set up a new group, ETSI OSM, for the development of open source software for NFV MANO.

ETSI MEC ISG was founded in 2014 and aims at creating a standardized, open environment to allow the efficient and seamless integration of applications across multi-vendor MEC platforms. Many activities are going on within this ISG, including defining service scenarios, technical requirements, framework and reference architectures, as well as proof of concepts.

ETSI mWT ISG was launched in 2015, with the purpose of studying mmWave technology for higher frequency bands in between 30 and 300 GHz. Currently, the mWT ISG is working to facilitate the use of the V-band (57–66GHz), the E-band (71–76 and 81–86GHz) and, in the future, higher frequency bands (50 GHz up to 300 GHz) for large volume backhaul and fronthaul applications. The mWT ISG also expects to publish a specification on the channelization of W-band (92–114,5GHz) and D-band (130–174,8GHz) in the near future.

ETSI NGP ISG was created to review the future landscape of Internet protocols, with the goal of creating more efficient Internet protocols for future IP networks including 5G. In October 2016, the first group specification on the next generation protocols was published, which lists example use cases and compares the existing IP suite protocols with next generation ones.

2.4.4 IEEE Activities

Currently, the IEEE has various 5G initiatives⁷ taking place in many working groups, such as the IEEE 802 standards group, the IEEE Communication Society (ComSoc), and the IEEE SDN initiative.

The IEEE 802 standards group originally focuses on local and metropolitan networks, including both wired (IEEE 802.3 for Ethernet) and wireless networks (IEEE 802.11 for Wi-Fi, IEEE 802.15 for WPAN, and IEEE 802.16 for WiMaX). Currently, the IEEE 802 standing committee group has been discussing the creation of an IEEE 5G specification and the relationship between IEEE and IMT-2020 [34]. Several potential 5G-related projects have been considered, such as IEEE 802.11ay for mmWave bands operation, and IEEE 802.1ah for IoT.

IEEE ComSoc has, since May 2015, been organizing a series of summits that focus on 5G. Recently, IEEE ComSoc has formed an IEEE GET5G Committee to discuss various issues and challenges related to 5G, and develop special interest groups (SIGs) in various areas such as mmWave, cloud-based mobile core, Tactile Internet, end-to-end latency, network architecture, and gigabit service enablement, which are crucial to 5G.

The IEEE SDN Initiative was launched in 2014 by the IEEE Future Directions Committee, addressing specific issues and challenges raised by the adoption of SDN and NFV that goes beyond technical issues to also encompass skill development and economics. The IEEE SDN initiative has recently published a white paper, which identifies technical challenges, business sustainability and policy issues of 5G software-defined ecosystems [35].

⁷ IEEE 5G Initiative. <http://5g.ieee.org/about>

2.4.5 IETF Activities

IETF is a standards body for the development and promotion of standards for the evolution of the Internet. It has many groups working on specific topics related to the development of 5G standards. For example, in the IoT research area, IETF has 6TiSCH and CoRE. There are working groups researching service function chaining (SFC), and distributed mobility management (DMM). In addition, affiliated with IETF, the Internet Research Task Force (IRTF) has several research groups working on new emerging technologies, such as SDNRG, NFVRG, Thing-to-Thing (T2RG), and ICNRG. Recently, several research topics, such as network slicing and IP protocols for 5G, have also been actively discussed at side meetings at the regular IETF 96 and IETF 97 meetings in 2016.

2.5 5G Research Communities

Along with the activities being carried out in the standardization organization, 5G-related research topics are also being discussed globally within research communities in Europe, Asia and America. It should also be noted that they have proven the strong cooperation between each other through the signing of Memorandum of Understanding (MoU) contracts. In the following, the ongoing 5G related activities within these communities will be summarized.

2.5.1 European 5G Related Activities

The following will provide a snapshot of the status of ongoing activities related to the development of 5G that takes place within the EU 7th Framework Program (FP7), the EU Horizon 2020 (H2020) Program, and the Celtic Plus Program.

2.5.1.1 5G Research in EU FP7

The earliest research related to 5G was started under the umbrella of EU FP7. Since 2013, the European Commission has granted 50 million euros to research into developing 5G technology, with the mission of making the EU a leader in 5G research and delivering 5G mobile technology by 2020. The grant was funded to more than ten projects⁸, which address the architecture and functionality needs for 5G and beyond 4G networks, such as METIS-I, 5GNOW, iJOIN, Mobile Cloud Networking, etc. Among these projects, the METIS project phase I (METIS-I)⁹ is the biggest, with the total cost of around 27 million euros. The METIS-I project has been recognized as a reference for the development of 5G worldwide. So far, these projects have been completed, with their final reports now available.

2.5.1.2 5G Research in EU H2020

The EU H2020 program is the successor of the previous EU FP7 program, which has been funded with nearly 80 million euros available over a 7-year period from 2014 to 2020. In 2013, to continue the research on 5G, the EU Commission, the EU ICT industry,

⁸ EU FP7. <https://5g-ppp.eu/projects/>

⁹ METIS Project. <https://www.metis2020.com>

and other partners, formed 5G Infrastructure Public Private Partnership (5G-PPP)¹⁰ with the funding of 700 million euros from the H2020 program and 700 million euros committed by the private side, to continue developing 5G technology and to make EU take a global leading position in 5G. So far, the 5G-PPP has published white papers, which identify the vision, architecture, key capabilities, design principles, key enabling technologies, and the time plan for 5G. It has also published a series of white papers investigating use cases and requirements from vertical sectors such as eHealth, Factory, Automotive, and Energy. The 5G research within 5G-PPP will go through three major phases until 2020, with three corresponding proposal calls. The first phase was dedicated to the research and innovation and the first call has resulted in 19 projects being selected. The budget of 128 million euros funds the 19 projects, working with different research topics from 2014 to 2016, including radio network architecture and technologies, convergence beyond last mile, network management, and network virtualization and software networks [36]. With the budget of 148 million euros, the second phase will focus on proofs of concept, experiments, and vertical industry, spanning from 2017 to 2019. The third phase will focus on large-scale trials with vertical industry. The research in this phase will be started by 2018, with the fund of around 425 million euros.

2.5.1.3 5G Research in Celtic-Plus

Celtic-Plus¹¹ is a telecommunication and ICT cluster under the umbrella of EUREKA, which is an intergovernmental organization formed by several European countries in 1985. The Celtic-Plus mission is to facilitate innovative research projects in the area of telecommunications, new media, future Internet, and applications and services towards a new “Smart Connected World”. The current research focus within Celtic-Plus spans two key areas: “Networking and Clouds” and “Services and Applications”. Each research area has many research topics, which are closely related to the development of 5G, such as cloud computing, SDN, NFV, IoT, eHealth, smart cities, and smart homes, etc. Some examples of 5G related projects within the Celtic-Plus are UNITED about NFV support of IoT services, WINS@HI about wearable IoT network solution for work safety, SIGMONA about SDN and NFV for mobile network¹², etc. [37].

2.5.2 Asian 5G Related Activities

Ongoing activities related to the development of 5G are also being conducted in some research communities in Asia, including 5G Forum in South Korea, 5G Mobile Communications Promotion Forum (5GMF) in Japan, and IMT-2020 Promotion Group in China.

2.5.2.1 South Korea: 5G Forum

In May 2013, the 5G Forum¹³ was formed on the basis of the collaboration between mobile network operators, global manufacturers, research institutes, universities, and the government. Its main mission and objectives are to promote 5G technology research

¹⁰ 5G-PPP. <https://5g-ppp.eu>

¹¹ Celtic-Plus. <https://www.celticplus.eu>

¹² SIGMONA Project. <http://sigmona.org/>

¹³ 5G Forum. <http://www.5gforum.org/>

and development, as well as international collaboration on 5G technology. During the years 2015 and 2016, the 5G Forum has published several white papers outlining the vision, requirements, enabling technologies, spectrum considerations, and a service roadmap for the development of 5G by 2020 and beyond. As planned, the first world trial of 5G will be deployed by the members of 5G Forum such as ETRI, SKT, KT at the Pyeongchang Winter Olympics in 2018.

2.5.2.2 Japan: 5GMF Forum

The 5GMF Forum¹⁴ was founded on September 30, 2014 with the objectives of promoting 5G research in Japan, as well as the global collaboration based on a roadmap on the 5G implementation policy by the government of Japan. In May 2016, the 5GMF Forum published the first white paper entitled “5G Mobile Communications Systems for 2020 and beyond”, which identifies its visions, key concepts and key technologies with their release of 5G. In addition, the 5GMF Forum has also been organizing several workshops on 5G related issues and global collaboration with other 5G initiatives. As planned, the world’s first 5G implementation will take place at the Olympic and Paralympic Games in Tokyo in 2020.

2.5.2.3 China: IMT-2020 5G Promotion Group

The IMT-2020 5G Promotion Group¹⁵ was jointly established by three Chinese ministries, including the Ministry of Industry and Information Technology, Ministry of Science and Technology, and the National Development and Reform Commission in February 2013, based on the original IMT-Advanced Promotion Group. The main objectives of the group are to promote the development of 5G technologies in China and to facilitate cooperation with other global 5G initiatives. So far, the group has published several white papers presenting their vision, key concepts and key technologies for 5G. The group has also been organizing several 5G Summits every year to open the door for collaboration with foreign companies and organizations.

2.5.3 American 5G Related Activities

The primary coordinated 5G effort in the Americas is in the 5G Americas¹⁶ organization. The 5G Americas is an industry trade organization composed of telecommunication service providers and manufacturers, aiming at leading 5G developments for the Americas. Its mission is to advocate for and foster the advancement and full capabilities of LTE wireless technology and its evolution beyond to 5G. So far, the 5G Americas has published several white papers related to 5G requirements, solutions, spectrum, as well as the evolution of 4G.

14 5GMF Forum. <http://www.5gmf.org/>

15 IMT-2020 (5G) Promotion Group. <http://www.imt-2020.cn/en>

16 5G Americas. <http://www.4gamericas.org>

2.6 Conclusion

The number of ever-increasing number of connected devices, the emergence of new services, and the increase of communication needs from vertical industries require the development of a next generation mobile system, 5G. Although 5G is still under the research and development process, it will most likely be ready for the users by 2020. In the 5G system, there will be a variety of advanced technologies, such as massive MIMO, Cloud-RAN, SDN, NFV, etc., as well as the usage of unoccupied frequency bands above 6 GHz. In addition, the 5G network infrastructure will be constructed as multiple slices by adopting the network slicing technology. These technologies promise to fulfill the requirements, which come from both the user perspective and the networking perspective, in order to serve a large number of use cases in 5G era.

This chapter presents the visions and main use-case classes, key requirements, and enabling technologies for 5G. The chapter also summarizes the status of ongoing activities related to 5G development from all over the world, including standardization organizations as well as regional research communities. In the remaining chapters, potential security related problems in future 5G networks will be analyzed and discussed.

2.7 Acknowledgement

This work has been funded by the project High Quality Networked Services in a Mobile World (HITS), funded by the Knowledge Foundation of Sweden.

References

- 1 Ericsson (2016) Ericsson mobility report, White paper, Available at: <https://www.ericsson.com/assets/local/mobility-report/documents/2016/ericsson-mobility-report-november-2016.pdf> (accessed 12 February 2017), November 2016.
- 2 NGMN Alliance (2015) NGMN 5G white paper, White paper, Available at: https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf (accessed 12 February 2017), February 2015.
- 3 3GPP (2016) Service requirements for next generation new services and markets; stage 1, Technical specification, TS 22.261 (Release 15), August 2016.
- 4 3GPP (2016) Study on scenarios and requirements for next generation access technologies, Technical report, TR 38.913 (Release 14), October 2016.
- 5 Open Networking Foundation (2012) Software-defined networking: The new norm for networks, White Paper, Available at: <https://www.opennetworking.org/images/stories/downloads/sdn-resources/white-papers/wp-sdn-newnorm.pdf> (accessed 12 February 2017), April 2012.
- 6 ETSI NFV ISG (2012) Network functions virtualization: An introduction, benefits, enablers, challenges, and call for action, White Paper, Issue 1, Available at: https://portal.etsi.org/nfv/nfv_white_paper.pdf (accessed 12 February 2017), October 2012.
- 7 3GPP SA1 (2016) Feasibility study on new services and markets technology enablers; stage 1, Technical report, TR 22.891 (Release 14), September 2016.

- 8 ITU-R (2015) IMT vision – framework and overall objectives of the future development of IMT for 2020 and beyond, Recommendation, REC M. 2083-0, September 2015.
- 9 Gupta, A. and Jha, R.K. (2015) A survey of 5G networks: Architecture and emerging technologies. *IEEE Access*, 3, 1206–1232.
- 10 ITU-R (2015) Studies on frequency-related matters for International Mobile Telecommunications identification including possible additional allocations to the mobile services on a primary basis in portion(s) of the frequency range between 24.25 and 86 GHz for the future development of International Mobile Telecommunications for 2020 and beyond, Resolution 238 (WRC-15), November 2015.
- 11 Rappaport, T.S. et al. (2013) Millimeter wave mobile communications for 5G cellular: It will work! *IEEE Access*, 1, 335–349.
- 12 Sooyoung, H. et al. (2016) Proposal on Millimeter-Wave Channel Modeling for 5G cellular system. *IEEE Journal of Selected Topics in Signal Processing*, 10(3), 454–469.
- 13 Rangan, S. et al. (2014) Millimeter wave cellular wireless networks: Potentials and challenges. *Proceedings of IEEE*, 102(3), 366–385.
- 14 Gesbert, D. et al. (2007) From single user to multi-user communications: Shifting the MIMO paradigm. *IEEE Signal Processing Magazine*, 24(5), 36–46.
- 15 Ngo, H.Q. (2015) *Massive MIMO: Fundamentals and system designs*. Linkoping University, PhD Dissertation.
- 16 Papadopoulos, H. et al. (2016) Massive MIMO technologies and challenges towards 5G. *IEICE Transactions on Communications*, E99-B(3).
- 17 Kamel, M., Hamouda, W. and Youssef, A. (2016) Ultra-dense networks: A survey. *IEEE Communications Surveys and Tutorials*, 18(4), 2522–2545.
- 18 3GPP TR 23.887 (2013) Study on Machine-Type Communications (MTC) and Other Mobile Data Applications Communications Enhancements, v.12.0.0, December 2013.
- 19 Boccardi, F. et al. (2014) Five disruptive technology directions for 5G. *IEEE Communications Magazine*, 52(2), 74–80.
- 20 Shariatmandari, H. et al. (2015) Machine-type communications: Current status and future perspectives toward 5G systems. *IEEE Communications Magazine*, 53(9), 10–17.
- 21 Asadi, A., Wang, Q. and Mancuso, V. (2014) A survey on device-to-device communication in cellular networks. *IEEE Communications Surveys and Tutorials*, 16(4), 1801–1819.
- 22 Tehrani, M.N., Uysal, M. and Yanikomeroglu, H. (2014) Device-to-device communication in 5G cellular networks: Challenges, solutions, and future directions. *IEEE Communications Magazine*, 52(5), 86–92.
- 23 Checko, A. et al. (2014) Cloud RAN for mobile networks – A technology overview. *IEEE Communications Surveys and Tutorials*, 17(1), 405–426.
- 24 Rost, P. et al. (2014) Cloud technologies for flexible 5G radio access networks. *IEEE Communications Magazine*, 52(5), 68–76.
- 25 Agyapong, P. et al. (2014) Design considerations for a 5G network architecture. *IEEE Communications Magazine*, 52(11), 65–75.
- 26 ETSI ISG MEC (2015) Mobile edge computing: A key technology towards 5G, White paper, issue 11, September 2015.
- 27 Bonomi, F., Milito, R., Zhu, J. and Addepalli, S. (2012) Fog computing and its role in the Internet of Things. *Proceedings of 1st ACM Workshop on Mobile Cloud Computing (MCC)*.

- 28 Liyanage, M., Gurkov, A. and Ylianttila, M. (eds) (2015) *Software-Defined Mobile Networks (SDMN): Beyond LTE Network Architecture*. John Wiley & Sons.
- 29 Nguyen, V.G., Brunstrom, A., Grinnemo, K.-J. and Taheri, J. (2017) SDN/NFV based mobile packet core network architectures: A survey. *IEEE Communications Surveys and Tutorials*, vol. 00, no. 99.
- 30 ITU-T (2014) Requirements of network virtualization for future networks, Recommendation ITU-T Y.3012, April 2014.
- 31 ETSI ISG NFV (2014) Network function virtualization (NFV); Management and orchestration, Group specification, version 1.1.1, December 2014.
- 32 ITU-T Focus Group IMT-2020 (2015) Report on standard gap analysis, December 2015.
- 33 3GPP SA2 (2016) Study on Architecture for Next Generation System, TR 23.799, version-0.5.0, June 2016.
- 34 IEEE 802 5G/IMT-2020 (2017) Standing Committee [Online] Available at: http://ieee802.org/Stand_Com/5G/index.html (accessed 12 February 2017).
- 35 Manzalini, A. et al. (2016) Towards 5G software-defined ecosystems: Technical challenges, business sustainability and policy issues. IEEE SDN Initiative White Paper. Available at: <http://sdn.ieee.org/images/files/pdf/towards-5g-software-defined-ecosystems.pdf> (accessed 12 February 2017).
- 36 The 5G-PPP (2017) First wave of research and innovation projects. Available at: <https://5g-ppp.eu/wp-content/uploads/2015/10/5GPPP-brochure-final-web.pdf> (accessed 12 February 2017).
- 37 Celtic-Plus (2017) Available at: <https://www.celticplus.eu/running-projects/> (accessed 12 February 2017).

3

Mobile Networks Security Landscape

Ahmed Bux Abro

VMware

3.1 Introduction

The mobile network started merely as a voice communication channel but has surpassed any other medium in history to become the center of our daily life, economy and governance. Future of societies and major economies such as Europe rely heavily on the mobile infrastructure where a direct financial and economic impact of the ICT (Information and Communications Technology) industry results in approximately 5% of GDP or hundreds of € billions [1]. The mobile population is going to hit 5.5 billion by the time 5G is officially launched in 2020 [2]. That means more connectivity and more information generation and sharing using the mobile networks.

As a result of the telecommunication boom and its critical role in the overall public and economic health, this industry will soon become a prime target for anti-state and criminal actors who want to leverage this platform to disrupt its growth and in turn use it to launch attacks against a wider user base of mobile networks. Impacts of breaches in this new generation of the connected world can be huge and impactful, as happened recently in the case of the Yahoo massive data breach that impacted 1 billion users in a single attack or the LinkedIn 2016 breach that impacted 117 million, resulting in a leak of sensitive personal and professional data [3].

Mobile network security has gradually evolved in parallel to the evolution of the telecommunication industry. In this chapter, we will cover the overall security threat landscape of mobile networks as it evolved with the various generations of mobile networks. We will also discuss the history of mobile security threats, relative protection mechanisms and impacts to the mobile networks.

3.2 Mobile Networks Security Landscape

The mobile network security landscape (as shown in Figure 3.1) should be viewed in the light of how the various generations of mobile networks have evolved. There is a direct correlation between the evolution of the mobile network technology and the relative

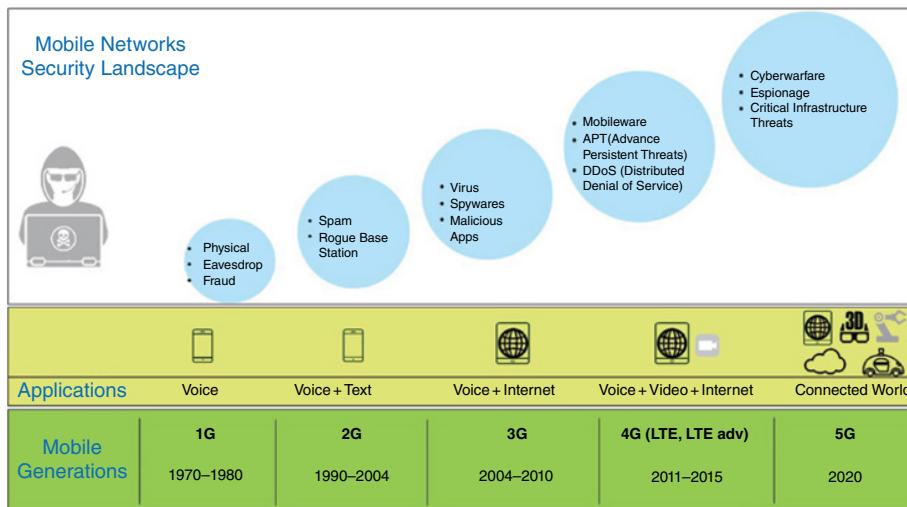


Figure 3.1 Mobile network security landscape.

evolution of security threats in terms of the technology architecture, technical capabilities, services offered, and associated threat vectors.

Mobile networks started witnessing serious threats and challenges immediately after the introduction of the first generation (also called 1G) of mobile technology and has kept on growing as a complex and challenging threat landscape. 1G was primarily introduced to offer mobility for voice users. Consumers started witnessing the freedom to attend and make calls while mobile. Criminals discovered an opportunity and methods to commit mobile frauds and impersonate the legal subscribers to hack their phone to make free calls. Cell phone cloning became an industry by making and selling illegal cloned phones. Some hackers identified new ways to hijack and eavesdrop on the calls while being made, and listen in to the private conversations for various nefarious reasons.

With 2G, the era of message spamming arose in the mobile world. Spamming was used as a pervasive attack to inject false information or send unwanted marketing jargon to the mobile users. Message inboxes were occupied with spam messages targeting a certain group or the wider community. Fraudsters used mobile spamming for their own vicious purposes. Rogue base stations (also called IMSI Catchers) were invented to intercept mobile traffic by offering fake network authentication.

As the user devices become smart and resourceful, data applications and internet became the key services offered by mobile service providers in new 3G networks. An average 3G data connection speed was somewhere between 500 and 700 kbps, which was sufficient to provide connectivity for internet facing applications. Threat vector in 3G targeted the user phones, computer system and its operating system. Mobile OS vulnerabilities were exploited, as mobile applications were injected with malicious code to gain unauthorized access to sensitive personal information, such as contacts, user passwords and location data. As the data speed increased so did the type of infections in the shape of malwares and spywares.

LTE was the first time the mobile network was switched to an all IP-based end-to-end architecture. It helped mobile service providers with speed of innovation, offered new services and scale, and also increased the threat vector for 4G networks. DDoS

(Distributed denial of service) and APT (Advance persistent threats) were the new realities for the mobile network, as impact to the service was critical with huge financial losses as the result of such attacks. Attackers became more organized and started following a systematic approach in their threat of execution. It has become harder to detect their stealth presence in the mobile network, to protect and mitigate, with an average attack consisting of months' duration.

5G is coming forward with the promise of connecting billions of devices, phones over a highly reliable, widely dense, bandwidth capable, fast and fault tolerant network infrastructure, which will service multiple sectors and industries. Key use cases for 5G are critical infrastructure, IoT, smart cities and the connected world. With these use cases, 5G will be an ideal target for attackers who may want to cause major economic and social disruption within a minimal timeframe. 5G threats will be constituted around financial and politically motivated gains, executed by groups of professionals and criminals with extensive technological knowledge and resources. The threat landscape for 5G will be dynamic and based on sophisticated and complex threats, such as Stuxnet and flame malwares.

3.2.1 Security Threats and Protection for 1G

1G offered voice-only services that were based on analogue technology, also referred to as Advance Mobile Phone System (AMPS). It used separate frequencies by using the FDMA (Frequency Division Multiple Access) method for each call and required a huge bandwidth to serve several mobile units. There were no built-in authentication or identification mechanisms to uniquely identify the user phone, which left the mobile network exposed, with zero protection against channel hijacking, cloning and eavesdropping. Mobile carriers suffered millions of dollars in loss because of cloning and eavesdropping; later it became a grown-up illegal industry with a well established source of revenue for unlawful actors [4]. Cloning was used not only for system misuse but also for many other criminal and fraudulent activities.

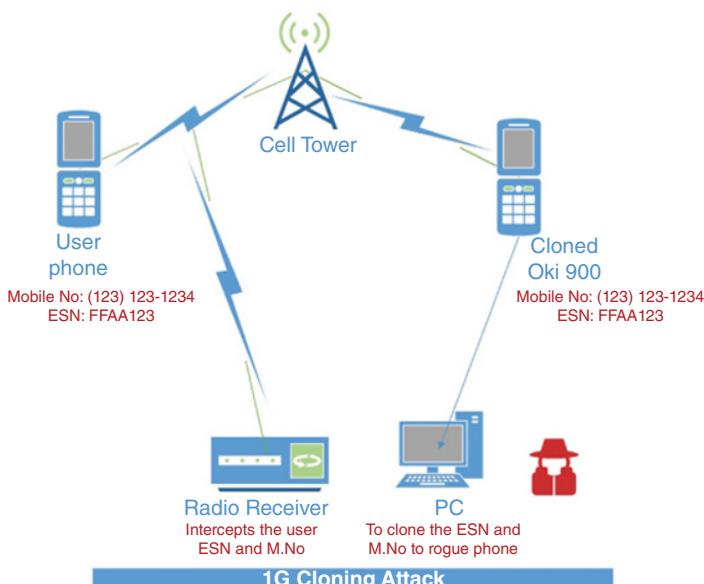


Figure 3.2 Cell phone cloning attack in 1G network.

As demonstrated in Figure 3.2, a cell phone cloning attacker requires certain hardware and software tools to impersonate a legitimate network subscriber. A radio receiver is used to sniff and intercept the user call targeting at the cell tower. The ESN (Electronic Serial Number) and mobile number information is stolen through the receiver. The attacker will use a PC with software to burn the stolen ESN and mobile identity information to clone a copy of the original cell phone. Attackers are able to use off-the-shelf tools like Oki 900 to impersonate and eavesdrop on communications over AMPS [3]. It was hard to distinguish between the victim and the cloned phone.

Such threats have forced carriers to disrupt many legitimate subscriber connections and introduce new identification mechanism, such as a unique pin code for each customer to remove nefariously cloned mobile devices from their network.

3.2.2 Security Threats and Protection for 2G

User authentication remained a key focus while developing 2G standards, as the goal was to reduce the call charge fraud, channel hijack and mobile cloning attack surface. The subscriber identity module (SIM) was first used to assign unique identification for mobile phones and could be securely stored as a computer chip inside the cell phone.

As user identity protection began to be taken seriously, a new threat was introduced in 2G, masquerading as a carrier base station with a rogue base station (also called IMSIcatcher) to perform a man-in-the-middle-attack (MitM).

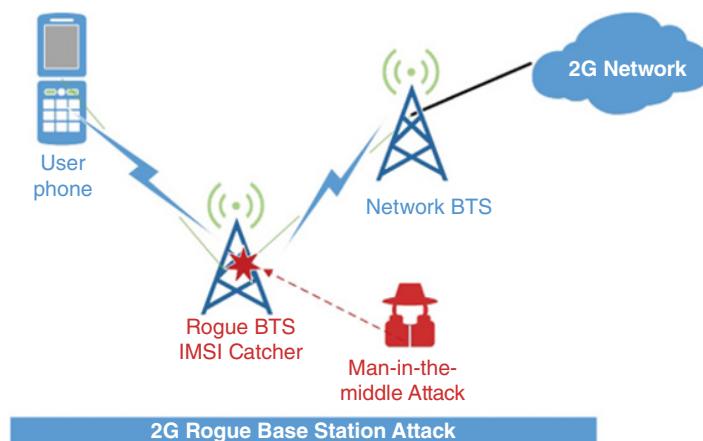


Figure 3.3 IMSI Catcher attack in 2G network.

The IMSI (International Mobile Subscriber Identity) catcher threat is demonstrated in Figure 3.3. This attack exploited the lack of 2-way authentication between the cell phone and the mobile network. Attackers would install a rogue base station that could impersonate a valid BTS (Base Transceiver Station) and let the user transmit the critical identity information such as IMSI over an unsecure channel. A rogue base station attack

was not only used to steal the IMSI information, but was also used to sniff the voice traffic over GSM, and to tap the sensitive user data transmitted over GPRS and EDGE. An attacker could easily copy the user internet traffic and extract the information such as the password through analyzers. Interestingly, the tools required to launch a GSM MitM attack are readily available on the market, that is, a regular BTS (Base Transceiver Station) and open source software OSMOCOMBB.

2G also introduced encryption on a limited scale to protect the traffic between user equipment and base station. Although it was unable to fully protect against cryptanalytic attacks, it was able to provide some basic encryption protection for signaling and user data.

Other common security threats were in the form of mobile short messaging spam traffic for false advertising and marketing.

3.2.3 Security Threats and Protection for 3G

Mobile network third generation (3G) decided to baseline its security following the CIA (Confidentiality, Integrity and Availability) framework. As a result, the AKA (Authentication and Key Agreement) protocol was adopted for two-way authentication between user equipment and the network, and also to protect against attacks, such as rogue base stations. AKA used strong 128 bit auth keys and hash functions to maintain the authenticity and integrity of signaling messages sent over the radio.

Even though two-way authentication reduced the chances of such attacks, attacks such as MitM would still be possible in the UMTS environment using such advanced tools like mobile jammers and OSMOCOMBB.

As 3G was designed to offer the next generation of data services and Internet connectivity for mobile users, new challenges and vulnerabilities were introduced to the system.

Mobile networks transitioned to a packet switching model, IP-based RAN (Radio Access Network) and IP Core network, and these changes unleashed the IP-based threat vector that had not been present in previous generations of mobile networks.

Mobile devices were replaced with small-sized computers called smartphones, which became host to typical OS vulnerabilities and weaknesses. Smartphones now required regular patching and updates against system vulnerabilities, as any failure will expose the phone to threats by attackers who can exploit the vulnerabilities to leaked data or install viruses and spywares.

Installation of unauthorized or malicious application caused phones to be hacked and were sometimes used to degrade the mobile service providers' service performance and attack the network. Some phone manufacturers were able to implement a strict application security policy for a centralized app store, but others found it hard to cope with the ever-growing number of malicious codes hosted on their app store platform.

3.2.4 Security Threats and Protection for 4G

4G LTE and LTE Adv. has overall introduced a great leap and evolution to the mobile network performance and throughput. It offers voice, data, video and internet services through a common mobile architecture.

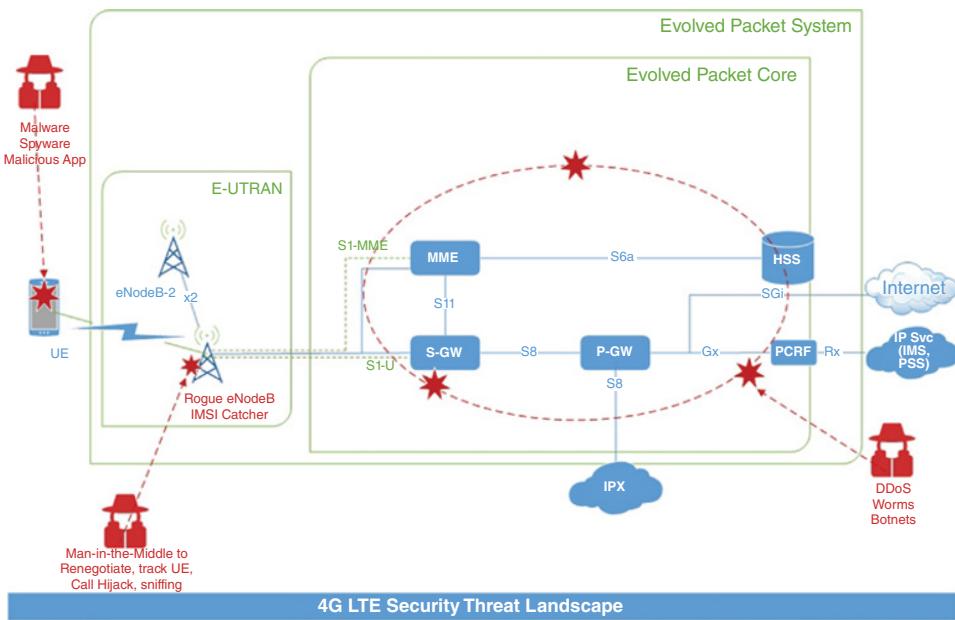


Figure 3.4 4G end to end security threat landscape.

4G threats were scattered throughout multiple domains of the 4G network, as shown in Figure 3.4. There were new types or viruses and malware targeting smartphones, in order to steal user data and passwords. Millions of malicious apps have been developed to impersonate user games, utility or fake major banking apps. As the install base of the smartphone operating system went into millions, attackers started identifying vulnerabilities and developed threats accordingly.

With IP core networks, 4G networks were targeted with well-designed DDoS (Distributed Denial of Services) attacks to cause a larger impact on the mobile services. 4G LTE security can be divided into multiple domains such as UE, RAN, Core Network and Internet Services. Individual domain security threats for 4G are covered in the following sections.

3.2.4.1 LTE UE (User Equipment) Domain Security

Today's UE are a powerful internet connecting small handheld computers with high speed CPU (Central Processing Unit) and memory capacity. It serves us in our daily life, social and financial activities, not only to interact with each other, but to make online payments and bank transactions. Smartphones are always connected via wireless LAN or cellular connection and can run an interdependent operating system and software applications that allow users to access their data anywhere, at any time, and in any place. A simple vulnerability in the mobile operating system can have a significant impact; as a reference, a recent "XCODEGHOST" vulnerability found in an iOS tool affected 500 million users [5].

Today, about 87% of the time spent on mobile devices is using apps and at least 24.7% of mobile apps carry one high risk security flaw [6]. Malicious applications can be downloaded and installed intentionally or accidentally by the users or attackers respectively.

These malicious apps can be used to gain stealth access to user personal data and passwords (stored on the phone) for financial and other criminal gains.

Malwares and worms can be installed to launch an attack on the local user network or widely target the mobile service provider network, as the smartphone is considered a trusted network device by mobile service providers.

3.2.4.2 LTE (Remote Access Network) Domain Security

LTE E-UTRAN can be exploited to gain access to UE locations using its Cell Radio Network Temporary Identifier (C-RNTI), while UE resides in a single cell or roams across multiple cells. This attack can be protected by encrypting the traffic carrying control signal, and command and confirm messages for C-RNTI [7].

3.2.4.3 LTE Core Network Domain Security

LTE is designed with an IP-based end-to-end open network architecture that helps simplify the overall network operations, but on the other hand, it opens the mobile network to IP-based security threats.

The LTE core network can be a potential target for a distributed denial of service (DDoS) attack with an impact as large as the loss of network services to millions of subscribers. DDoS can target critical EPC (Evolved Packet Core) components to cause a loss of service, such as sending overwhelming authentication and authorization requests to the HSS database, causing system overflow conditions. A DDoS can be run against an entire IP network backbone by injecting false routing information or manipulating the network database to cause service downtown. The simplest form of DDoS attack is the TCP SYN attack, where a network device is sent with millions of false TCP SYN packets to cause a denial of service condition. Complex DDoS attacks may involve installation of malware on the affected systems and will require extensive efforts to mitigate its after effects from the system.

Protection for DDoS can be done through a specialized Anti-DDoS security system, or newer network technologies such as Software Defined Network (SDN) can be used to push security policies through a centralize controller.

3.2.4.4 Security Threat Analysis for 4G

Table 3.1 below covers the detailed threat analysis of common 4G threats, their impact severity and the probability of threat occurrence:

Protection mechanisms against 4G security threats involve the following measures and mechanisms:

- Always install authorized apps on your mobile devices.
- Download apps from the vendor app store only.
- Protect app access with a passcode.
- Define a service access policy for each app, i.e. limit location and contact access to certain app types.
- Enable encryption on mobile devices.
- Business and critical apps should use secure containers to encrypt and control sensitive data.
- Install antivirus.
- Always keep the operating system patched and updated.

Table 3.1 4G Security Threat Analysis.

| Threat | Threat description | Impact Severity (Minor, Moderate, Severe, Extreme) | Threat Occurrence (1-5, Low-High) |
|------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------|--------------------------------------------------|
| Insecure Mobile OS (Operating System) | Mobile operating systems carry vulnerabilities that are fixed, using vendor issued patches and updates. If not fixed, can cause attackers to exploit vulnerabilities to hack into mobile systems | Moderate | 4 |
| Download unauthorized apps | Users download app from app store that are not verified by the vendor or checked by their IT department, and can be malicious | Moderate | 5 |
| Insecure App with sensitive data | A legitimate app that leaks sensitive personal or business data and with no mechanism to encrypt or protect | Severe | 5 |
| Virus | Malicious software code with a specific purpose to damage mobile functions or files | Severe | 2 |
| Malware | An advanced virus or malicious app that can propagate and self-reproduce causing large-scale, network-wide damage | Extreme | 3 |
| Spyware | A malware type used to steal end user data, sensitive information to transmit to remote attackers | Extreme | 3 |
| DDoS (Distributed Denial of Service) | Launched as a coordinated attack involving hundreds of thousands of devices infected with malicious code. Targets the availability of mobile networks | Extreme | 2 |

3.2.5 Security Threats and Protection for 5G

With the wide spectrum of 5G applications and services and its critical role to serve society for social, economic growth and public safety, the threat vector for 5G can be wide. Motivations to threaten and attack the 5G will now be higher than in previous network generations. There is a greater chance that the 5G will be a key target for criminal activities driven by various different motives, such as state-sponsored political motives, adversaries, organized crime cartels, espionage and cyberwarfare.

Attackers keep innovating and finding new ways to evade detection, having learnt to exploit the social and financial system for their needs. With the evolution of digital payment systems, and technologies like Bitcoin getting into the mainstream, it will be a lot easier for criminals to stay under the blanket and keep gaining the financial benefits.

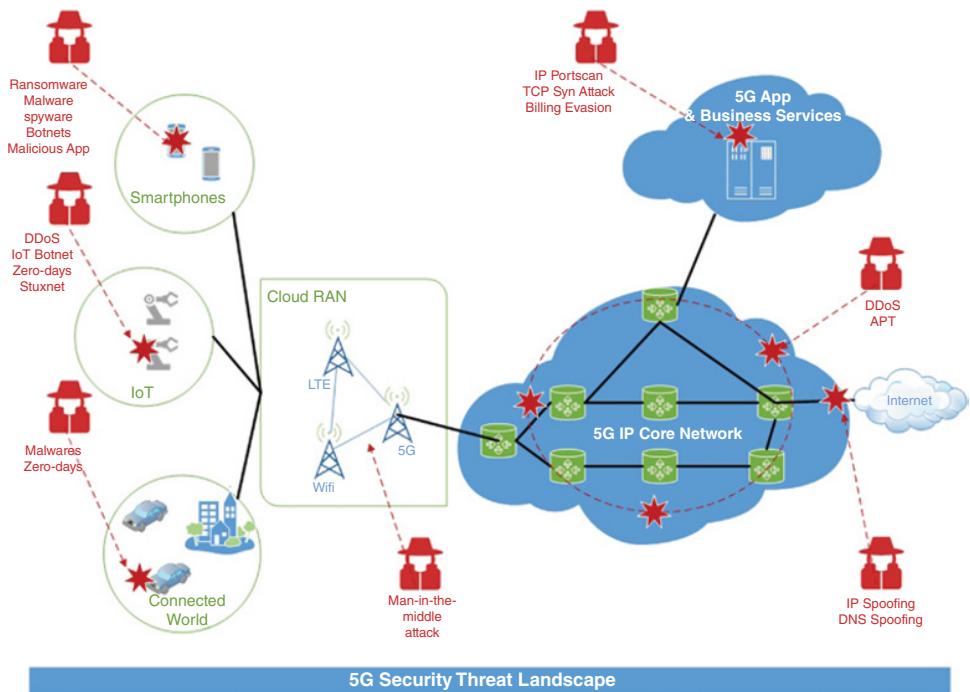


Figure 3.5 5G security threat landscape.

The 5G threat vector will have no boundaries and will range from the end user equipment such as mobile phones, industrial items, sensors, home automation, automated cars, enterprise networks to mobile networks. With such a larger threat landscape, as demonstrated in Figure 3.5 above, it will span from end user devices to RAN (Radio Access Network) to mobile core network to the Internet. Figure 3.5 also covers the positioning of various threat types within the network, such as smartphone threat types including ransomware, spyware malware and Bots. An MitM attack could be launched at the Cloud RAN domain, while DDoS may target the IP core network. Every single domain of 5G network will be under threat in such a landscape.

It will be a serious challenge for security practitioners to step up and build a defense system that can protect 5G networks end to end.

Applications and services will be 10x for 5G networks, as it is going to be the platform for the digital and connected world. 5G is going to serve as the critical infrastructure like other traditional utilities such as electricity, and will be used to provide connectivity for community health and governance, finance, trade and industrial systems. This enlarges the threat landscape of 5G beyond any other communication network that has ever existed.

5G is planned for 2020, but based on the ongoing 5G standardization and research work so far, we have the understanding that the 5G architecture is going to be laid on the top of IP-based architecture and will inherit and expand the characteristics of traditional IP-based mobile networks.

3.2.5.1 Next Generation Threat Landscape for 5G

5G is not only going to be the next generation of mobile network, but it will still be a platform to introduce the next generation of security threats. Next generation security threats are foreseen to carry the following characteristics:

- *Sophisticated*: complex in nature, uses a multi-staged mix of various attack vectors and tools. For example, the Angler exploit kit that is packaged to exploit a multiple vendor into a single attack;
- *Obfuscatory*: attacks that are obscured by multiple layers and very hard to detect;
- *Evasive*: hard to detect and ability to hide itself, e.g. ransomware cryptowall attack; and
- *Persistent*: such attacks are meant to be consistent and evolve themselves after every failed attempt.

APTs (Advance Persistent Threats) are attacks that carry out the above characteristics, are hard to mitigate and protect from and eventually become serious threats causing great damage. APTs usually target crucial infrastructure facilities, large service providers that serve masses and cause major disruption and multi-dimensional impact.

3.2.5.2 IoT Threat Landscape

5G will serve as the network platform for future industrial systems, critical infrastructure and IoT (Internet of things). Attacks targeting such critical networks are foreseen as advanced in nature and will require a highly complex skillset and resources to execute. Politically motivated and sponsored attacks will often be encountered in the 5G environment. Such attacks will have the potential to leverage the undisclosed system vulnerabilities, also called zero day flaws, and operated through a centralize command and control (C&C) system. There will be no single tool or approach used for such attacks and it will often be a mixture of threat techniques such as DDoS, phishing and advance rootkits. The impact of these kinds of attacks is widespread and goes beyond economic damage and may involve national security, public safety and loss of human life. As per US Presidential Executive Order 13010, telecommunications is considered as a critical infrastructure and states that: “it is so vital that their incapacity or destruction would have a debilitating impact on the defense or economic security of the United States” [8].

Most of critical infrastructures today are using a centralize control system, such as SCADA (Supervisory Control and Data Acquisition), which is further connected to a network. All remote systems need to send regular signaling and health information to the system for normal operation, so a malware type of threat can be used to disrupt the control system and result in the loss of services. In the future, 5G is going to serve these systems for the underlying network and connectivity services and will be exposed to the threats to these systems.

3.2.5.3 5G Evolved Security Model

To protect 5G from advanced and complex threat landscapes, we need an evolved security model (as show in Figure 3.6) that offers in-depth protection, not only from existing threats but also from evolving and zero-day threat types.

It will require a well-defined security strategy and plan to protect the various components for the 5G network, including the end devices and end users. An effective security strategy can be designed based on telemetry data gained through a well-designed surveillance system. A security position plan would be needed to place the security mechanisms in effective positions. Processes and tools need to be identified to

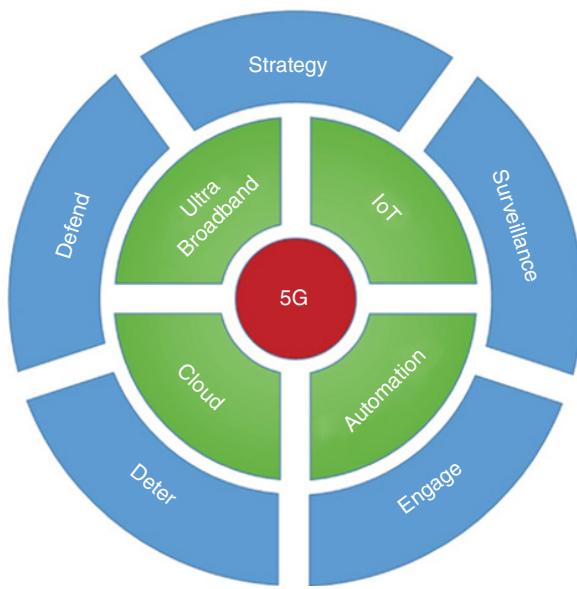


Figure 3.6 5G Evolved Security Model.

engage in case an attack is detected, block it and evolve the system. A deterrence and defense system should be in place for the continuous protection of the mobile network and dependent services.

3.2.5.4 5G Security Threat Analysis

One of the best approaches to prepare for security challenges in 5G is to master and learn the current challenges and threats to LTE and LTE Advance networks. As mentioned earlier, due to the core IP-based nature of 5G, it will inherit most of the threats that currently exist in 4G (LTE and LTE Advance) networks. At the top of threats listed under Section 3.2.4 for 4G LTE threats, Table 3.2 below lists some advance threats with the potential to target 5G networks:

Table 3.2 5G security threat analysis.

| Threat | Threat description | Impact Severity (Minor, Moderate, Severe, Extreme) | Threat Occurrence Probability (1-5, Low-High) |
|-----------------|-------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------|-----------------------------------------------------|
| Ransomware | Specialized malwares use exploit, encrypt and lock access to critical data. Access granted after paying demanded ransom money | Severe | 3 |
| Advance Malware | Advance malwares targeting billions of mobile and IoT devices with capability to exploit the OS and network vulnerabilities | Extreme | 3 |

(Continued)

Table 3.2 (Continued)

| Threat | Threat description | Impact Severity (Minor, Moderate, Severe, Extreme) | Threat Occurrence Probability (1-5, Low-High) |
|---------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------|--------------------------------------------------------------|
| IoT Botnets | IoT and mobile devices hosting a control agent/bot receiving remote commands and continuously leaking telemetry information to a remote bot-master running a central command and control (C&C) system. Used for both passive and active attacks | Severe | 2 |
| Critical Infrastructure Threats | Threats that are focused, damaging critical infrastructure services such as SCADA, i.e. Stuxnet, Shamoon attacks | Extreme | 3 |
| Zero-day Attacks | An advance attack exploiting the undiscovered vulnerabilities of a system. Can be a combination or package of multiple attack types, malware, rootkits and botnets | Extreme | 1 |

3.3 Mobile Security Lifecycle Functions

Security lifecycle functions for mobile device and networks are developed to protect the end-to-end security posture of mobile systems and networks. It addresses the security at individual stages of mobile provisioning, configuration, assessment and security monitoring. Lifecycle functions leverage security systems, tools and processes are required to protect the confidentiality, integrity and availability of mobile devices and networks.

As 5G networks are going to introduce new tools to carry out mobile-e-commerce and new business use cases for mobiles, such as IoT and ultra-broadband. Such enterprises will especially look towards a well-defined security management and governance system for their mobile end users to further protect their critical data and applications residing on end-user personal mobile device uses as an extension of a corporate network, as in the case of BYOD (Bring Your on Device).

Key security failures in such cases are [9]:

- 1) inconsistent security policies;
- 2) leakage in shared media;
- 3) minimal device management;
- 4) readable data stays in disposed devices; and
- 5) inter-application data leakage.

A security lifecycle, as show in Figure 3.7, can help address the above and other common security challenges and reduce risks at various stages of the mobile device during its participation in the network. Key stages of a mobile addressed by the security



Figure 3.7 Mobile security lifecycle functions.

lifecycle are mobile provisioning, configuration, management and monitoring. In the sections below, we will cover the security lifecycle functions for mobile systems and networks in detail.

3.3.1 Secure Device Management

While MSP (Mobile Service Provider) have limited control on mobile device security, they do offer basic authentication and authorization of mobile devices for network access. However, enterprises using BYOD to offer corporate access on personal mobile devices requires advance management capabilities and uses tools such as MDM (Mobile Device Management) to centrally manage and protect mobile devices used by their employees. MDM helps organizations enforce security policies, protect against malicious threats and restrict unauthorized access to mobile devices. Mobile devices need to register or enroll with the centralized enterprise MDM before they download the security policies, configurations and controls to protect that device. Such policies can enforce protection mechanisms for mobile devices, such as applying complex alphanumeric passcodes, defining mobile auto-lock settings or accessing selected apps.

3.3.2 Mobile OS and App Patch Management

The mobile OS (Operating System) vendor regularly releases patches and updates to close known vulnerabilities and loop holes in their software that can possibly cause device compromise or in some cases major security breaches. Attackers often look for opportunities to get hold on these vulnerabilities and exploit them to gain unauthorized access to the mobile devices. It is critical for users to regularly patch their mobile device operating. Mobile service providers often release advisories for their users to update certain patches and upgrades to avoid a security flaw.

Similarly, app developers also send regular updates to their mobile apps to fill in for any insecure features or code residing in their applications. Application patches and updates are independent of OS updates and need to be handled separately by the mobile users. Mobile OS vendor often make this process simple through their app stores, as for custom application installations, it is challenging to keep those apps up to date.

3.3.3 Security Threat Analysis and Assessment

As we know, 5G networks will rely heavily on IP-based network communication and protocol and will be leveraging the new software defined mobile networks (SDMN). 5G is planning to leverage the benefits of SDMN to separate the management plane, control plane and data plane of mobile networks to simplify the networks and offer new and improved services using new programmable networks. SDMN with its capabilities offers new security challenges and can cause vulnerabilities on different planes (management, control and data) of the network. Threat vectors for 5G- and SDMN-based networks are complex and can introduce network flaws dynamically at different points of mobile networks.

Traditional security assessments and threat analysis techniques are based on the static and simple nature of traditional mobile networks and do not address the challenges introduced with 5G- and SDMN-based dynamic network behaviors. Security assessment for SDMN networks need to cover and address all the components and layers (planes) of the mobile network. New security assessment approaches are recommended that focus around the dynamic nature of SDMNs. Such new assessments should use the attack graphs to cover all SDMN factors and levels, and also need to use an Analytic Hierarchy Process (AHP) to build the structure [10].

3.3.4 Security Monitoring

With the evolution of mobile networks from LTE to LTE Adv. and now 5G, mobile RAN and core networks technologies have also evolved and are superseded by new technologies such as Cloud RAN, NFV and SDMN. It is critical for mobile operators to have a comprehensive visibility and knowledge for their network operators in a real-time fashion, not only to offer better service assurance but also protect their critical network infrastructures from security threats. Legacy mobile security monitoring solutions were not designed to protect SDN- or NFV-based networks and had no or limited capability to integrate with the modern technological components of the mobile network and so need to be replaced with security monitoring solutions that offer higher performance, scalability and the capability to integrate and work with these new mobile technologies.

Security monitoring solutions for LTE and 5G networks should offer a capability to monitor and inspect both signaling and data traffic at multiple network points, starting from UE to RAN and all the way to LTE/5G core network components. The solution should be able to not only inspect the IPv4 and IPv6, but also offer visibility to other protocols such as TCP, UDP, GRE, etc. Instead of traditional packet-based inspection, 5G networks could also leverage SDN control and data plane separation and perform centralized network flow traffic monitoring for a deeper visibility and correlation of traffic traversing inside the network.

Security monitoring for the mobile network needs to offer some advance security services such as:

- vulnerability tests;
- regular security health check for entire network;
- flow-based network visibility;
- security alert management system; and
- traffic monitoring and inspection.

3.4 Conclusion

In this chapter, we covered the evolving threat landscape of mobile networks, starting from 1G when mobile offered voice only services and when the security threats mostly revolved around financial gains by cloning the mobile phones. We then discussed how the security threats evolved and targeted the mobile network infrastructure, such as base stations, causing wider impact. With the introduction of data services and mobile devices replaced by handheld computing devices called smartphones, mobile networks built over an IP core, security threats also widened remarkably. Attacks once targeted at computers were leveraged and reused for mobile devices, the aim now beyond financial gains, as new threats such as espionage and DDoS attacks were introduced. Mobile devices were a major source of carrying the spywares, worms and malwares, and every day a new vulnerability was found in mobile OSes, as billions of devices gained access to insecure apps, causing unconceivable impact. 5G networks carrying forward the legacy of previous networks are also planned to serve as a critical infrastructure for business and offers new capabilities such as cloud, IoT and ultra-broadband, which will face unique challenges and become the target of sophisticated security threats such as ransomware and Botnets.

We also discussed a security lifecycle approach to offer effective and comprehensive protection mechanisms for modern mobile devices and networks. Various stages of mobile devices and its participation in the network were discussed in detail. Following an end-to-end multi-stage lifecycle approach can help build a security wall around evolving networks, such as LTE and 5G.

References

- 1 White paper, 5G Vision, by 5G PPP supported by the European Commission. Available at: <https://5g-ppp.eu/wp-content/uploads/2015/02/5G-Vision-Brochure-v1.pdf>
- 2 Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020, White Paper. Available at: <http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html>; http://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp1_security.pdf
- 3 Yahoo Discloses New Breach of 1 Billion User Accounts. Available at: <http://www.wsj.com/articles/yahoo-discloses-new-breach-of-1-billion-user-accounts-1481753131>

- 4 AMPS: Cloning. Available at: https://en.wikipedia.org/wiki/Advanced_Mobile_Phone_System
- 5 SonicWall Annual Threat Report (2016) Available at: <https://www.sonicwall.com/whitepaper/2016-dell-security-annual-threat-report8107907>
- 6 NowSecure Mobile Security Report (2016) Available at: <https://info.nowsecure.com/rs/201-XEW-873/images/2016-NowSecure-mobile-security-report.pdf>
- 7 Rodriguez, J. (ed.) (2015) *Fundamentals of 5G Mobile Networks*. West Sussex, UK: John Wiley & Sons, Ltd.
- 8 Robles, R.J., Choi, M-K., Cho, E-S., Kim, S-S., Park, G-C. and Lee, J-H. (2008) Common threats and vulnerabilities of critical infrastructures. *International Journal of Control & Automation*, 1(1), 17–22.
- 9 Zahadat, N., Blessner, P., Blackburn, T. and Olson, B.A. (2015) BYOD Security Engineering: A framework and its analysis. *Computers & Security*, 55, 81–99.
- 10 Luo, S., Dong, M., Ota, K., Wu, J. and Li, J. (2015) A security assessment mechanism for software-defined networking-based mobile networks. *Sensors*, 15(12), 31843–31858.

4

Design Principles for 5G Security

Ijaz Ahmad¹, Madhusanka Liyanage¹, Shahriar Shahabuddin¹, Mika Ylianttila¹, and Andrei Gurtov²

¹ Centre for Wireless Communications (CWC), University of Oulu, Finland

² Department of Computer and Information Science, Linköping University, Linköping, Sweden

4.1 Introduction

The vision of the 5G wireless networks lies in providing very high data rates, higher coverage through dense base station deployment with increased capacity, significantly better Quality of Service (QoS), and extremely low latency [1]. 5G is considered to provide broadband access everywhere, entertain higher user mobility, enable connectivity of a massive number of devices (e.g. IoT), and the connectivity will be ultra-reliable and affordable [2]. The development towards an all-IP-based communication, for example in 4G, has already helped develop new business opportunities, provide new online services and connect industrial machines, home appliances and business units. However, with this development, the security challenges and threat vectors have also increased.

Wireless communication systems were prone to security vulnerabilities from the very beginning. In the first generation (1G) wireless networks, mobile phones and wireless channels were targeted for illegal cloning and masquerading. In the second generation (2G) of wireless networks, message spamming became common, not only by pervasive attacks but also by injecting false information or broadcasting unwanted marketing information. In the third generation (3G) wireless networks, IP-based communication enabled the migration of Internet security vulnerabilities and challenges in the wireless domains. With the increased necessity of IP-based communication, the fourth Generation (4G) wireless networks enabled the proliferation of smart devices, multimedia traffic, and new services into the mobile domain. This development lead to a more complicated and dynamic threat landscape. With the advent of the fifth generation (5G) wireless networks, the security threat vectors will be bigger than even before, with greater concerns for privacy [3].

One haunting fact that has always stayed alive during the development towards the 5G is that the IP-based communication not only increased the variety of services and network traffic but opened new doors to develop new cracking and hacking mechanisms for wireless networks and mobile devices. Wireless networks and user equipment, however, had difficulty in keeping up the pace with the increasing security

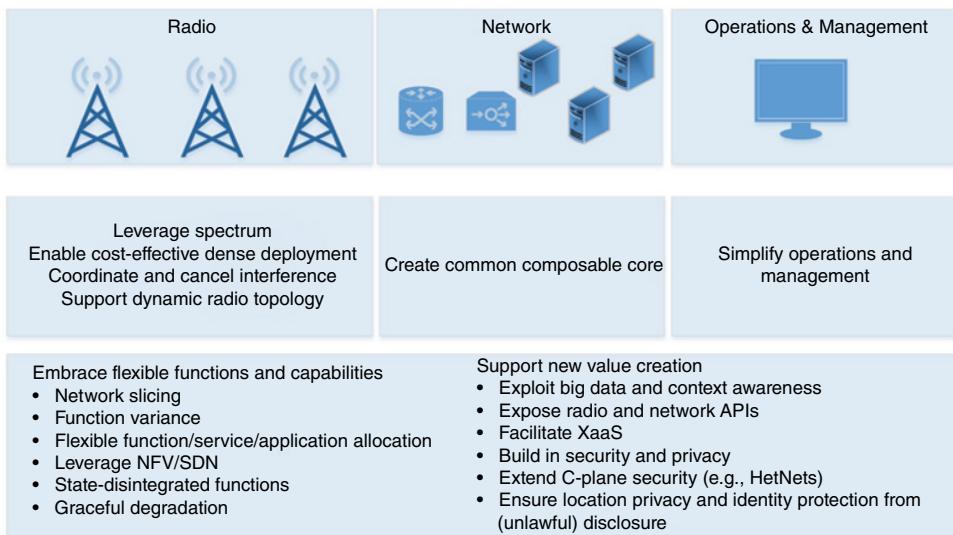


Figure 4.1 5G design principles.

challenges surfacing due to IP connectivity. Hence, new solutions and technologies have always been sought to protect the network, user traffic and services. In this chapter, we will provide an overview of the new types of security threats, and then present the solutions proposed for those threats.

5G will connect a critical infrastructure that will require more security to ensure safety of not only the critical infrastructure but safety of society as a whole. For example, a security breach in the power supply systems can be catastrophic for all the electrical and electronic systems on which the society depends upon. Similarly, data is critical in decision making and the data will be carried by the 5G network. Hence, adequate measures are required to safeguard the data. Similarly, it is envisioned that public safety systems will also be connected through 5G networks, hence it is more critical to develop proper measures to secure not only the network but also the services using 5G networks as a communication medium. Therefore, it is suggested by the Next Generation Mobile Networks (NGMN) [4] that 5G should provide more than hop-by-hop and radio bearer security. The 5G design principles elaborated in Figure 4.1 outlines the need for highly elastic and robust systems. Such architectures must support the deployment and placement of security functions whenever required in any network perimeter.

4.2 Overviews of Security Recommendations and Challenges

The security threat vectors in 5G will be multi-dimensional, right from the physical interfaces up to the application interfaces, services in the clouds, and user information. 5G networks will connect critical infrastructures, interconnect societies and

industries, provide anything as-a-service, and integrate new models of service delivery. The 5G ecosystem cannot be fully visualized at this moment, due to the rapid development and integration of new devices and services. However, the main attractions of 5G beyond extended connectivity, higher data rates and lower latencies will be the easy placement and utilization of new services and functions. This will complicate the security landscape as well. To make the security landscape easy to comprehend, we provide a discussion on security in two domains. First, the security of access networks, for example Radio Access Networks (RAN) that can be a composite of multiple access technologies such as cellular networks RAN comprising small and large base stations and WiFi, etc. Second, the security of the core network in which the network control resides with operator and vendor specific services. The International Telecommunication Union's Telecommunication sector (ITU-T) has proposed dimensions of security for telecommunication networks that address all the aspects of security [6]. Hence, first we will provide a brief introduction to these recommendations and then we will discuss different security challenges in different areas of 5G networks.

4.2.1 Security Recommendations by ITU-T

Security dimensions are proposed by ITU-T in its security recommendation [6] to address almost all the aspects of network security. The security dimensions include a set of security measures that can be used to protect the users and network against all major security threats. These dimensions are:

- *Access Control*: security measures that ensure only authorized personnel or devices access the network resources.
- *Authentication*: security mechanisms that ensure identities of the communicating parties and that a user or device is not attempting a masquerade or unauthorized replay of previous communications.
- *Non-Repudiation*: ensure that a particular action has been performed by a specific user or device is non-repudiation. Proper identities are used to ensure that authentic user or device can access particular services and resources.
- *Data Confidentiality*: security mechanisms to protect the data from unauthorized access. Encryption, access control mechanisms and file permissions are used to ensure data confidentiality.
- *Communication Security*: ensure that the data flows between the authorized end-points and is not diverted or intercepted in between.
- *Data Integrity*: ensures the correctness or accuracy of data in transmission and protects it from unauthorized modification, deletion, creation and replication.
- *Availability*: ensures that there is no denial of authorized access to network resources and applications. Events impacting the network, such as system failures or disasters, scalability and security compromise, must not limit access to authorized users and devices.
- *Privacy*: mechanisms that ensure protection of information, which might be derived from observing network activities.

4.2.2 Security Threats and Recommendations by NGMN

The Next Generation Mobile Networks (NGMN) [7] provides recommendations for 5G based on the current network architectures, and the lacking security measures that are either not implemented or available. The recommendation highlights the cautionary notes. These include the infancy of 5G with many uncertainties, lack of defined design concepts and the unknown end-to-end (E2E) and subsystem architectures. The recommendation highlights the limitations in the access networks and cyber-attacks against the network infrastructure. The details of the security limitations and recommendations can be found in [7]. Below we highlight the key points in the recommendations:

- *Flash network traffic:* It is known that the number of end user devices will grow exponentially in 5G, thus the large-scale events may cause significant changes in the network traffic patterns that could be either accidental or malicious. Therefore, it is recommended that the 5G systems must minimize large swings in traffic usage and provide resilience whenever such surges occur, while maintaining an acceptable level of performance.
- *Security of radio interface keys:* In the previous generations, even in 4G, keys for the radio interface encryption are generated in the home network and sent to the visited network over unsecure links causing a clear point of exposure of the keys. It is recommended that the keys are either not sent over those links, such as SS7/Diameter, or properly secured.
- *User plane integrity:* 3G and 4G do not provide cryptographic integrity protection for the user data plane, though these provide protection to some signaling messages. It is recommended to provide the protection at the transport or application layer that terminates beyond the mobile network. The exception to this could be network level security for resource constrained IoT or latency sensitive 5G devices and services. Application level E2E security may involve too much overhead for data transmission in the packet headers and handshakes.
- *Mandated security in the network:* There are service-driven constraints on the security architecture leading to optional use of security measures. Unfortunately, these constraints undermine system-level security assumptions and cannot be completely eliminated. The challenge increases in multi-operator scenarios where one operator suffers due to inadequate measures by others. Therefore it is highly recommended that some level, if not all, must be mandated in 5G after proper investigation to recognize the most critical security challenges.
- *Consistency in subscriber level security policies:* There is a need that the user-security parameters are not changed due to roaming from one operator network to the other. In the case of highly mobile users, it is highly possible that all the security services are not updated frequently and per-user basis as the user moves from place to place or from one operator network to another in the case of roaming. When a user from one operator moves to another and is using latency sensitive services, the services might be provided through the edge of the visited operator network such as in Mobile Edge Computing (MEC). So will the security or the security of the service being used be automatically offered or configured at the new location? This needs security policy sharing among network operators on a much faster scale to secure user traffic with roaming. The recommendation discusses the possibility of using virtualization

techniques in such situations that can enable per-user slice configuration to keep the security policies and services intact whenever and wherever the user moves.

- *DoS attacks on the Infrastructure:* DoS and Distributed DoS (DDoS) attacks might circumvent the operation of devices controlling the critical infrastructure such as energy, health, transportation, and telecommunications, causing life threatening consequences with tremendous human and capital losses. DoS attacks are designed such that they exhaust the physical and logical resources of the targeted devices. The challenge will be more threatening due to the possibility of attacks from machines that are geographically dispersed in locations and in huge numbers. The network must be capable of servicing the increasing number of connections caused by the increasing proliferation of connected devices (e.g. IoT) with different operating capabilities and limitations.

4.2.3 Other Security Challenges

We can classify the security challenges on a high level into three domains, that is, security challenges in the access network, DoS Attacks, and security challenges in the core network. Below we briefly describe each of them.

4.2.3.1 Security Challenges in the Access Network

Network access security provides secure access to the network and services with protection from vulnerabilities in the radio. For example, the user must be ensured security from malicious network activities and the network must be secured from malicious access. 5G will utilize a variety of access technologies and integrate different types of access networks for extended coverage, higher throughput and lower latencies. To keep the network working, 5G must improve the system robustness against jamming attacks of the radio signals and channels. Furthermore, the security of small cell nodes must be improved due to their geographical distribution and ease of access.

One of the key challenges in 5G will be the excessive nodes sending data and receiving data simultaneously, practically jamming the radio interfaces. The challenge can be exacerbated by malicious nodes sending excessive signaling traffic, causing availability challenges or, in other words, leading to Denial of Service (DoS) attacks. Such signaling traffic or attacks must be recognized early and stopped before the jamming the network. 3G and 4G provided cryptographic integrity protection of some signaling messages but the user data plane was still not protected.

From 2G to 4G, the radio interface encryption keys are computed in the home core network and are transmitted to the visited radio network over SS7 or Diameter signaling links. These keys can be leaked, thus creating a clear point of exposure in the network [7]. Therefore, well designed key management protocols should be in place for 5G to reduce the threats. The basic techniques include improving the SS7 and Diameter security by introducing firewalls [7]. However, other approaches can be applied, such as using different secure control channels for distributing the keys. Some of these approaches are described in Chapter 10. Security of the physical layer is described in Chapter 6.

4.2.3.2 DoS Attacks

DoS and DDoS attacks originating from large sets of connected devices will very likely pose a real threat to 5G networks. These attacks can be either against the network infrastructure or the end user devices. Attacks against the infrastructure are designed to

deplete the resources of the network operator infrastructure that serves the users and devices. Though the original target is the operator network, the subscribers are indirectly affected. Attacks against users/devices are designed to deplete the resources of the users and devices. In this case, the subscribers and devices are directly targeted, but this impacts the network operator indirectly. Compromised user devices can also be used to cause the attacks against the network infrastructure.

DoS attacks on 5G network infrastructure would likely target the resources that are related to connectivity and bandwidth at promised levels of service. Hence, the focus can be against the following areas:

- 1) the signaling plane needed for authentication, connectivity and bandwidth assignment, and mobility of 5G users;
- 2) user plane needed to support two-way communication of devices;
- 3) management plane that supports the configuration of network elements that support signaling and user planes;
- 4) support systems that performs user/devices billing;
- 5) radio resources providing access to user devices; and
- 6) physical and logical resources supporting network clouds.

DoS attacks against the user devices will target the physical resources of the user devices such as memory, battery, processing units, radios, and sensors, etc. These attacks can also target the logical resources such as operating systems, applications, configuration data, and user data, etc.

4.2.3.3 Security Challenges in the Control Layer or Core Network

The massive penetration of IP protocols in the control and user planes in all network functions make the 5G core network highly vulnerable. Hence the network must be capable of ensuring availability with improved resilience against signaling-based threats. Specific security features must be incorporated for latency sensitive applications and use cases. The network must also incorporate the security requirements defined by the 3GPP. Furthermore, 5G networks should ensure communication in emergency situations such as when part of the network is either inaccessible or destroyed.

Overloading the signaling plane with huge number of infected IoT or M2M devices, either as an attempt of DoS attack or to gain access to the network, will be another pressing challenge in 5G networks [7]. IoT devices, in billions [46], will be resource constrained, thus making two kinds of requests. First, due to limited capabilities, these devices will require the resources in the clouds to perform processing, storing or sharing of information. Second, also due to their limited capabilities, these will be an easy target to masquerade or operate in a compromised environment for attacks on the network in the form of DoS attacks. Hence, the increasing number of connected devices will be a huge challenge for the signaling plane or core network of 5G networks. An example of this is the authentication and authorization requests to the HSS, which can potentially make the HSS inaccessible to legitimate users, in other words, compromise the centralized entity through a DoS attack or saturation attack [48].

The increasing range of communication services and devices leads to high traffic volumes for signaling purposes, such as authentication and bearer activation. Such traffic bursts bring about a signaling storm and may crash the core network [1]. Similarly, the signaling procedures occur at the NAS layer of 3GPP protocols that

include attach/detach, bearer activation, location update, and authentication. These form the NAS signaling storms [1]. This can be more challenging in 5G, where billions of devices will be connected to the same core network. Nokia Siemens Networks published [15] that signaling traffic is increasing 50% faster than the data traffic. Small cells with a vast number of connected devices being mobile will increase mobility handovers, thus increasing the signaling traffic. This will not only increase the signaling load on Mobility Management Entity (MME), but on other control entities such as HSS, public data network gateway (P-GW) and Serving Gateways (S-GW), to maintain the Quality of Service (QoS). Furthermore, the NAS layer of 3GPP protocols for UE attach or detach functions, bearer activation, location update, and authentication can cause signaling storms [16]. 3GPP recommends the use of IPsec encryption for LTE interfaces, such as X2, S1-MME, S5 and S6, etc. Thus, each eNB is required to support hundreds of IPsec tunnels, while the backhaul has to support thousands of tunnels. Such a massive tunnel establishment not only complicates the security establishment but massively increases the signaling load in the network. Furthermore, the tunnel establishment is static in nature and predefined by the administrator that further complicates the process of ensuring security of the control traffic. The tunnels, statically established, might not be in use but still sending periodic control information also increases the signaling load.

Therefore, using the currently deployed security architectures in 5G will cause major scalability and availability challenges, thus paving the way for DoS and DDoS attacks. Novel security architectures are needed for 5G to ensure the security of users and protect the network from malicious attacks.

4.3 Novel Technologies for 5G Security

Since 5G is not an incremental improvement in 4G, security systems should also be re-designed according to the design and architectural requirements of 5G. The vision for secure 5G systems outlined by NGMN is based on three principles:

- 1) Flexible security mechanisms;
- 2) Supreme built-in security; and
- 3) Automation.

The vision is that 5G should provide highly robust security systems against cyber-attacks, with enhanced privacy and security assurance. The security mechanisms must be flexible to incorporate novel technologies, for example for authentication and identification. The flexibility should enable the option of using encryption for the user plane and per network slice security parameters adjustment. The security systems must also be able to be automated to adjust and adapt itself intelligently according to the environment, threats or security controls. A holistic security orchestration and management will be highly required [4].

Since 5G has higher flexibility and agility, the two concepts that are most prominent to play a vital role in 5G are virtual network functions (VNFs) and software-based network control. These features are foreseen to be enabled by Network Functions Virtualization (NFV) and Software Defined Networking (SDN). NFV enables vendors to implement network function in software called VNFs and deploy them on high-end servers or cloud

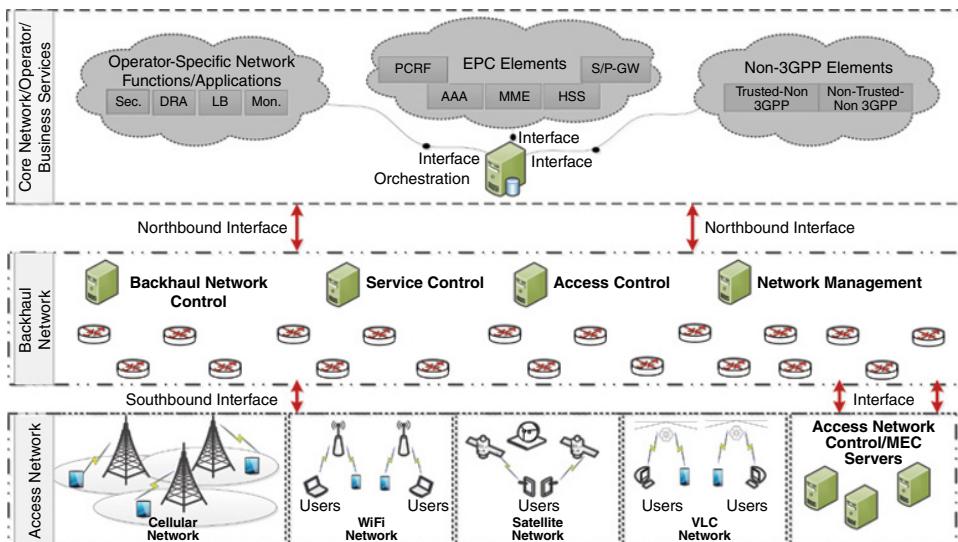


Figure 4.2 High-Level 5G architecture integrating multiple access technologies.

platforms instead of specialized function-specific hardware. SDN, on the other hand, separates the network control plane from the data forwarding plane to enable innovation in network control systems. Therefore, NFV and SDN will be vital in providing foolproof security in 5G systems. The basic diagram of the 5G network that provides connectivity to all devices at all times, covering almost every aspect of the society, is presented in Figure 4.2. The high level architectural diagram presented in the figure leverages the concepts of NFV and SDN to dynamically provide security in various network parameters, as the need arises. Below we describe the security challenges existing in both technologies and how these technologies can be used to increase the overall network security.

4.3.1 5G Security Leveraging NFV

Virtualization is used to decouple a system's service model from its physical realization and has been used in networking, for example for creating virtual links (tunnels) and broadcast domains (VLANs). Through virtualization, logical instances of a physical hardware can be used for different tasks, where the physical and logical instances are mapped through a network hypervisor [8]. The concept has led to the development of Mobile Virtual Network Operators (MVNOs). To minimize the investment in the infrastructure, MVNOs lease virtual resources including network resources from Mobile Network Operators (MNOs), thus a physical mobile network can host several MVNOs. MVNOs have their own operating and support systems and can offer independent services from the MNO. Therefore, NFV has a vital role, not only in the MVNO and MNO ecosystem, but also in the overall networking ecosystem to utilize the hardware resources in the most efficient manner possible.

NFV will enable function placement in different network perimeters without requiring function specific hardware but based on the need of the function or service at that location and time. Telecom networks will expose Application Programming Interfaces (APIs) on their hardware platform to users and third-party software providers to deploy their

services on the same hardware to reduce costs. However, decoupling the software and hardware will require new security models for the whole system, since the platform-specific security will not suffice for a shared hardware platform. There will be a demand for strong isolation mechanisms to secure each service running on the same hardware.

Virtualization enables multiple tenants or network users to share the same physical network resources that can create security vulnerabilities. A literature study on security implications of virtualization [9] shows that it has a positive effect on availability but has threatening security challenges related to confidentiality, integrity, authenticity and non-repudiation. Virtual machines can be created, deleted and moved around a network easily, hence tracking a malicious virtual machine would be much more complex. Similarly, if a hypervisor is hijacked, the whole system can be compromised [10]. Another major security challenge of NFV is to ensure trust among new elements such as hypervisors, virtual machines and management modules [54]. For instance, VNFs can store and fetch executable code from any server anywhere in the world. Therefore, a trusted mechanism is needed between the operator and cloud provider to ensure that the code is safe and correct.

On the other hand, NFV can highly improve network and user security. For example, secured network slicing can separate the communication of different parties, thus alienating malicious traffic from the remainder. Similarly, distributed can be deployed to resolve DoS and DDoS attacks, and with further intelligence, these VNFs can substantially improve self-protection of 5G networks [14]. Network hypervisor, a program that provides an abstraction layer for the network hardware, enables network engineers to create virtual networks that are completely decoupled from the network hardware. In this section, we outlined the importance of NFV and its security implications at a high level to show its importance for future networks. More detailed analysis of the security threat vectors and counter measures for VNFs and MVNOs is presented in Chapters 14 and 15.

4.3.2 Network Security Leveraging SDN

SDN separates the network control from the forwarding hardware and centralizes the network control into software-based controller platforms. The software-based control function will be centralized in high-end servers. This will accelerate novelty in network feature development, enhancement and rapid deployment. Therefore, the SDN-based wireless network has been a hot research topic and there are many proposals for SDN-based wireless networks [5]. The SDN architecture is vertically separated into three functional layers with interfaces between the layers, as shown in Figure 4.3. OpenFlow is the first viable implementation of SDN and also follows the three-tier architecture of SDN with OpenFlow applications, OpenFlow controller and OpenFlow switches.

The three logical planes are described as:

- 1) *Application plane*: consists of applications for various network functionalities such as network management, QoS management and security services, etc.
- 2) *Control plane*: is the logically centralized network control platform running the Network Operating System (NOS), having a global view of the network resources and stats, and provides hardware abstractions to the applications in the application plane.
- 3) *Infrastructure plane*: also called the data plane that consists the data forwarding elements that act on the instructions of the control plane for dealing with the data packets or traffic flows.

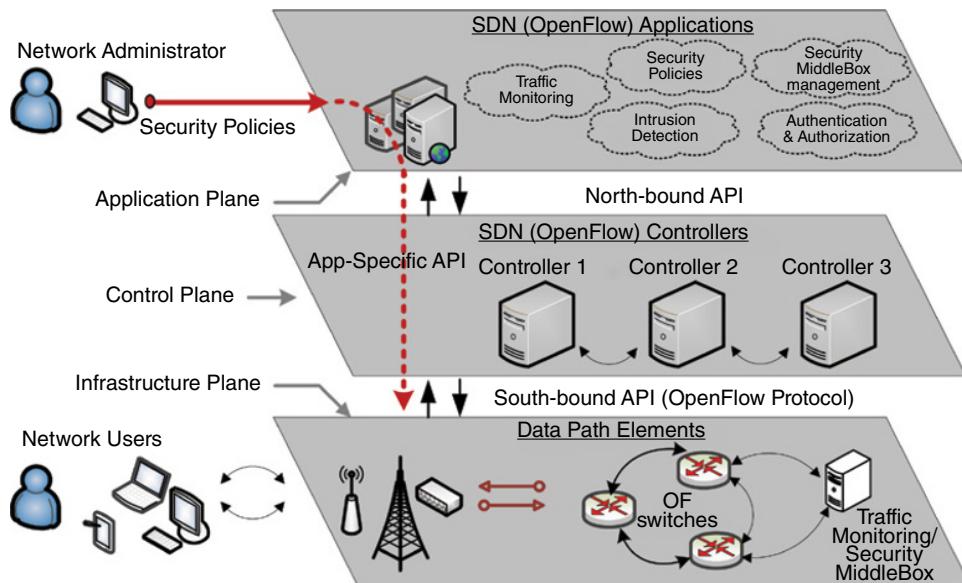


Figure 4.3 An overview of the SDN architecture.

In SDN, network security functions can be implemented as applications deployed in the SDN application plane. The applications gather traffic or network stats information through the control plane from the data forwarding plane using the north-bound interface (applications-control plane API). For example a security application, such as an intrusion detection application, can gather packet samples to perform analysis and then direct the data forwarding plane through the control plane to either drop the packets or forward the packets to a specific port. The port can be either towards the end user or a security middle box for further analysis. This makes the security systems highly flexible. When coupled with NFV, SDN would enable run-time network security function placement at any network perimeter as the need arises. The other main benefit is the decoupling of the network security functions from vendor-specific hardware. This decoupling would allow the network operators to change the security functions whenever deemed necessary, irrespective of the hardware specifications, or changes in the firmware of various hardware used for security purposes.

However, centralizing the network control and softwarizing network function opens new security challenges. For example, the centralized control will be a favorable choice for Denial of Service (DoS) attacks, and exposing the critical APIs to unintended software can render the whole network down. Some of the main threats are highlighted in Table 4.1. Therefore, SDN-based networks need novel security architectures right from the beginning. Below, we highlight the main security threats in SDN-based networks.

4.3.3 Security Challenges in SDN

4.3.3.1 Application Layer

SDN has two principle properties which form the foundation of networking innovation on one hand and the basis of security challenges on the other. First, the ability to control a network by software, and second, centralization of network intelligence in

Table 4.1 Security challenges in SDN.

| SDN Layer | Type of Threat | Threat Description |
|----------------|-------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Application | Lack of authentication and authorization | There are no compelling mechanisms for authentication and authorization of applications, and is more threatening in the case of a large number of third-party applications |
| | Fraudulent rules insertion | Malicious applications can generate false flow rules |
| | Lack of access control and accountability | A problem for the management plane and for illegal usage of network resources |
| Control | DoS, DDoS attack | Due to the visible nature of the control plane |
| | Unauthorized controller access | No compelling mechanisms to obligate access control for applications |
| | Scalability or availability | Centralizing intelligence in one entity will most likely present scalability and availability challenges |
| Data Plane | Fraudulent flow rules | Data plane is dumb and hence more susceptible to fraudulent flow rules |
| | Flooding attacks | Flow tables of OpenFlow switches can store a finite or limited number of flow rules |
| | Controller hijacking or compromise | Data Plane is dependent on the control plane, making its security dependent on controller security |
| Ctrl-Data Int. | TCP-Level attacks | TLS is vulnerable to TCP-level attacks |
| | Man-in-the middle attack | Optional use of TLS and complex configuration of TLS |

network controllers [13]. Since most of the network functions can be implemented as SDN applications, malicious applications, if not stopped early enough, can spread havoc across a network. The main security challenges that applications can pose to the network will be due to the availability of open APIs in network equipment, trust relationship between the controller and the applications (mainly third-party applications) and authentication and authorization of applications to change or modify the network behavior [13]. In 5G most of the functionalities will be implemented as applications due to the ease in modifications, making updates, and deployment. NFV will be the key enabler of application-based services and will take application-based services into the networking domains. Therefore, securing the network from anomalies generated by applications will be highly important.

4.3.3.2 Controller Layer

In SDN the control plane (e.g. OpenFlow controller) is a centralized decision-making entity. Hence, the controller can be highly targeted for compromising the network or carrying out malicious activities in the network due to its pivotal role. The same reason is valid for DoS and DDoS attacks. Furthermore, malicious applications can acquire network information from the controller if there are no compelling authentication and authorization mechanisms in place in the controller. The visible nature of the controller

makes it a favorite choice for DoS attacks. Since the SDN controller modifies flow rules in the data path, the controller traffic can be easily identified, thus making the controller a visible entity in the network. Scalability of the controller is another challenge that can be targeted to make the controller a bottleneck for the whole network. If the number of controllers is less or the controller capabilities are not good enough to respond to the queries of the data path elements, the controller can easily become a bottleneck [17].

4.3.3.3 Infrastructure Layer

The SDN switches have flow tables used by the controller to install flow rules for each flow. If the number of flows increases in the switch, there is a high chance that the flow tables will be exhausted. Thus, malicious users can send flows with different field headers making the flow tables to exhaust to cause saturation attacks. In this case, legitimate flows will be discarded due to the limited capability of the switch to buffer legitimate TCP/UDP flows. Since the switches are dumb by taking intelligence to the control plane, it will not be possible for the switches to differentiate genuine flows from the malicious ones. Therefore, the switch can be used for attacks against other switches and the controller. Furthermore, the data plane is dependent on the security of the control plane. If the security of the controller is compromised so that it does not provide instructions for the incoming flows, the data plane will be practically offline. This also makes the controller-data plane link a favorable choice for attacks. Transport Layer Security (TLS) and Datagram Transport Layer Security (DTLS) are specified for the controller-switch communication. However, the use of TLS and DTLS are left optional mainly due to its configuration complexity. This leaves the controller-switch communication open to attacks, thus increasing the vulnerability of the data and control planes.

4.3.4 Security Solutions for SDN

The logically centralized control plane of SDN provides a global view of the network and enables run-time configuration of the network elements. As a result, the SDN architecture supports highly reactive and proactive security monitoring, traffic analysis and response systems to facilitate network forensics, alteration of security policies and security service insertion [18]. SDN facilitates quick threat identification through a cycle of harvesting intelligence from the network resources, states and flows. The SDN architecture supports traffic redirection through flow-tables modification to analyze the data, update the policy, and reprogram the network accordingly. The programmability achieved by SDNs facilitates dynamic security policy alteration without the need of individual hardware configuration. The automation, thus achieved, would reduce the chances of misconfiguration and policy conflicts across different networks. Consistent network security policies can be deployed across the network due to the global network visibility, whereas security services such as firewalls and Intrusion Detection Systems (IDS) can be deployed on specified traffic according to globally defined security policies.

Below, we define the security of each plane or layer of SDN.

4.3.4.1 Application Plane Security

The SDN control plane works between the network hardware and applications to hide the network complexity from applications. Hence, the centralized control architecture makes it easy to use applications by providing them with the network statistics and

packet characteristics to implement new security services. Therefore, various solutions are proposed to enable the applications work in its functional boundaries with controlled access to network resources. The PermOF is a fine-grained permission system that provides controlled access of data and control planes to the SDN applications. The design of PermOF provides read, notification, write and system permissions to various applications to enforce permission control.

The NGMN security recommendation advises application level data integrity protection for battery constrained IoT devices or low latency 5G devices for user plane data integrity. This will enable data protection beyond the mobile network, thus minimizing the chances of vulnerability of data due to compromises in the network. Thus, SDN enables such applications to implement end-to-end security for constrained devices beyond the security implications of the network.

4.3.4.2 Control Plane Security

Since the security of the control plane is pivotal to the whole network, there have been many proposals and approaches for securing the control plane. The Security-Enhanced (SE) Floodlight controller [19] is an extended and secure version of the original floodlight controller [20]. By securing the SDN control layer, the SE-Floodlight controller provides mechanisms for privilege separation by adding a secure programmable north-bound API to the controller and operates as a mediator between the application and data planes. It verifies flow rules generated by applications and attempts to resolve flow rules conflicts between applications.

To mitigate the risks of controller failure due to scalability, or the chances of DoS attacks due to its centralized role, controller resilience strategies have been proposed. The strategies include controller resilience through redundancy, maximizing its storage and processing capabilities, and distributing controller functionalities among multiple control points in the network. The OpenFlow variant of SDN supports wildcard rules so that the controller sends an aggregate of client requests to server replicas. By default, microflow requests are handled by the controller that can create potential scalability challenges and increase the chances of failures due to DoS attacks. Normally reactive controllers are used that act on a flow request when it arrives at the controller. Proactive controllers would install the flow rules in advance, thus minimizing the flow request queue in the controller. Similarly, various load balancing techniques are suggested that will balance the load among multiple controllers in a network.

4.3.4.3 Data Plane Security Solutions

The data plane that transports the actual packets also requires proper security mechanisms. The data plane must be secured from unauthorized applications. Applications can install, change or modify flow rules in the data plane, therefore security mechanisms such as authentication and authorization are used for applications that can change the flow rules in the data plane. FortNox [21] enables the controller to check contradictions in flow rules generated by applications. FlowCchecker [22] identifies inconsistencies in the flow rules in the data plane switches. Multiple controllers proposed for the controller resilience also help the data plane elements work if one controller fails to provide flow rules for newly arrived traffic flows.

4.4 Security in SDN-based Mobile Networks

The current version of SDN, that is the OpenFlow, operates on traffic flows. A flow can be a number of packets with the same characteristics, for example same TCP connection, or packets with a particular MAC or IP address. Operating on flows has been shown to be much more feasible in terms of control and granularity. The basic operation on flows is such that OpenFlow has three main entities as explained for the concept of SDN. These are:

- 1) *OpenFlow applications*: SDN application plane;
- 2) *OpenFlow controllers*: the SDN control plane; and
- 3) *OpenFlow Switches*: the SDN data plane.

The OpenFlow switches are dumb data path elements that forward packets between ports based on the instructions installed in their flow tables by the controller. The OpenFlow switch has three basic elements:

- 1) a flow table with actions associated with each flow;
- 2) a secure channel to the controller; using
- 3) an OpenFlow protocol that provides an open and standard mechanism for the controller to communicate with the switch [22,23].

When a new flow arrives, the switch checks its flow table for a matching entry. If there is no matching entry, the switch forwards it to the controller. The controller installs a matching entry in the switch flow table. Henceforth, when flows arrive at the switch, the switch checks its flow tables and acts accordingly. The flow tables have basically three types of actions for the packets. First, forward the flow to a given port as enlisted in the matching flow entry in the table. Second, encapsulate and forward the flow to the controller. Third, drop the flow's packets. This makes security services rather simple in SDN and forms the basis of security in future technologies:

- *Flow sampling*: is the selection of packets or packet header fields through various algorithms for analysis. Selected samples can be sent to security applications or systems to analyze the content of the flow and verify security threats or vulnerabilities. Basic analysis targets can be the content of the flow packets or header fields, frequency of particular types of packets, and inter-arrival times of packets with different characteristics. In SDNs, flow sampling can be as easy as changing the output port numbers and counters in the flow tables of the switch. The destination on that port can be a security system and the counter can show the number of packets to be sent to that destination.

In the following sections, we elaborate how the concepts of SDN can be used to provide robust security for mobile networks.

4.4.1 Data Link Security

Data link security is necessary to ensure that the data flows between the authorized endpoints and is not diverted or intercepted while in transit. The previous generations, that is, 3G and 4G, did not provide cryptographic integrity to user plane communication. In 5G, it will be a major security concern and will expose private communication not only

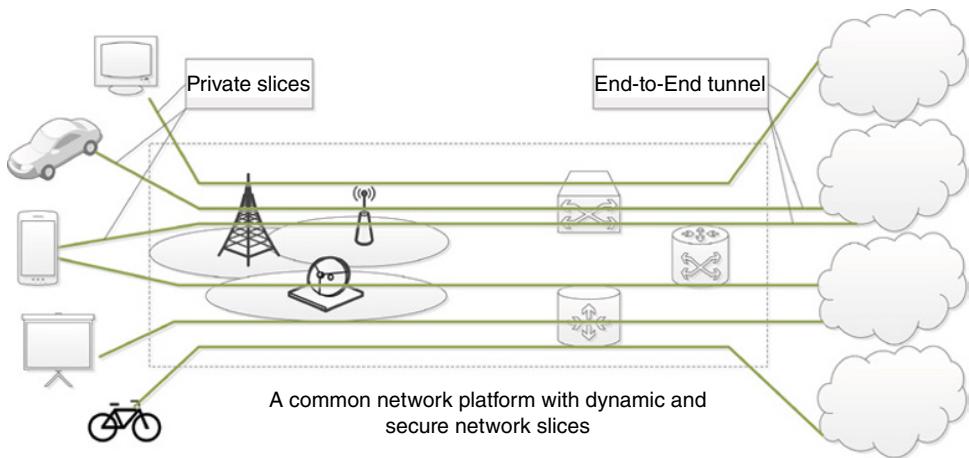


Figure 4.4 Secure network slices for different services.

between users but between devices carrying sensitive information such as data of health care systems and other critical infrastructures. Therefore, new mechanisms are needed to secure the data communication between users and devices. The OpenFlow protocol supports Transport Layer Security (TLS) and Datagram Transport Layer Security (DTLS). TLS is used to provide privacy and data integrity for the communication between users. DTLS is used to secure data between communicating applications, mainly UDP traffic. These technologies use symmetric cryptography for data encryption. The TLS protocol is composed of two layers, that is, the TLS record protocol and the TLS handshake protocol. The Record Protocol guarantees connection privacy and reliability by means of data encryption. The TLS Handshake protocol authenticates the communicating parties with each other and negotiates the encryption algorithm and cryptographic keys before transmitting the first packet of an application.

Besides the use of TLS and DTLS, virtual networking or network slicing can be used to provide private communication channels for both data and control information, as shown in Figure 4.4. Slicing can also provide isolation-based data integrity and privacy. Slices of individual users can be separated by a networking hypervisor such as the FlowVisor [11]. Traffic isolation can be used to protect one type of traffic from another to strengthen the confidentiality and integrity of user traffic [43]. Hence, the Open vSwitch platform provides isolation in multi-tenant environments and during mobility across multiple subnets [44]. The OpenFlow Random Host Mutation (OF-RHM) [45] technique is proposed to avoid scanning attacks on end-hosts. Using the moving target defense (MTD) technique, the OF-RHM mutates IP addresses of end-hosts to avoid scanning attacks. The VAVE [14] platform validates the source addresses of all incoming packets to prevent data from being spoofed or forged through the OpenFlow interface attached to legacy devices.

4.4.2 Control Channels Security

Control channels carry the important control information between user and network, and among network entities.

Mutual authentication and key agreement between the UE and the network is important in many aspects, the most important being the identity insurance of the UE. In LTE, the UE and the network, or its entities such as the Mobility Management Entity (MME), perform mutual authentication through the Evolved Packet System (EPS) Authentication and Key Agreement (AKA), known as the EPS AKA. The EPS AKA is secure enough and has no visible vulnerabilities demonstrated so far [47]. When a UE connects to the EPC through the non-3GPP access network, the UE is authenticated through the AAA server. For trusted non-3GPP access networks, the UE and AAA server use Extensible Authentication Protocol-AKA (EAP-AKA) or improved EAP-AKA for authentication. For mistrusted non-3GPP access networks, the UE uses the evolved packet data gateway (ePDG) IPsec tunnel establishment to connect to the EPC [49]. Such control channels, besides being secure, have the following benefits [47]:

- The messages are short compared to other authentication protocols.
- It requires only one handshake between the UE and serving network, and between the serving and home networks.
- The HSS is updated through the serving network, thus is capable of handling many requests.
- The symmetric-key-based protocol makes the computations required in the authentication center (part of the HSS), and in the USIM (Universal Subscriber Identity Module) very efficient compared to public-key-based mechanisms. However, the advantages of the use of public-key based authentication and key agreement schemes could include that the home network does not need to be contacted for each authentication.

It is expected that in 5G there will be multiple control points in a network, which will require security of the control channels among those control points. For example, the concepts of SDN will be used for the benefits described in the previous sections. Thus, multiple controllers will be used for higher availability and scalability. Therefore, the control channels among the controllers must be secured. Similarly, the control channel between SDN controller and SDN switches must also be secured. The OpenFlow variant of SDN uses TLS, in which identification certificates are properly checked in either direction and allow encrypting the control channel in order to secure it and prevent it from eavesdropping. Furthermore, multiple control channels (associations) between switches and controllers are suggested to avoid the chances of services outages due to connection failures. The latest OpenFlow specifications support multiple connections between switches and controllers to improve network resilience in case of link failures. Therefore, fast link restoration mechanisms and backup entries with different priorities in the OpenFlow switches have been proposed and demonstrated in [49]. The backup links are computed by the controller and the traffic is switched to the backup link upon failure of the existing link. Similarly, flow entry migration techniques are proposed in [51] to reinstate a flow within 36 ms. This mechanism fulfills the carrier grade recovery requirement of 50 ms. Furthermore, HIP-based [52] secure control channels between the switches and the controllers are also proposed [53].

Moreover, IPsec is the most commonly used security protocol to secure the communication channels in current telecommunication networks such as 4G-LTE [55]. Thus, novel IPsec-based communication architectures were designed to secure control and data channels of 5G [56]. The proposed architecture use distributed Security

Gateways (SecGWs) to secure the controller and IPsec Encapsulating Security Payload (ESP) Bounded-End-to-End-Tunnel (BEET) mode tunnels to secure the control and data channels communication. Moreover, the Identity-Based Cryptography (IBC) protocol-based security mechanism is also proposed to secure the inter-controller and control channel traffic in a general multi-controller SDN networks [58].

4.4.3 Traffic Monitoring

Traffic monitoring can be used to detect intrusions and prevent them. Intrusion Detection Systems (IDS) and Intrusion Prevention Systems (IPS) observe network traffic according to different security policies, find vulnerabilities, threats and attacks, and use countermeasures to secure the network. In traditional networks, the control planes of network elements are loosely coupled with limited communication between their control planes. Hence, IDS and IPS technologies in each network domain are independently configured to deal with the challenges in its domain. This makes the current systems hard to update with newly identified types of attacks and threat vectors. SDN takes a different path by enabling applications to retrieve switch statistics or extract samples of packets from flows for security analysis. After security analysis, the applications can direct the controller to either drop the packets or forward them to security systems or middle boxes for further investigation. Therefore, SDN makes traffic monitoring rather simple by enabling global visibility of the network traffic behavior and network programmability.

4.4.4 Access Control

Traditional access control mechanisms use firewalls deployed on network boundaries to examine the incoming or outgoing packets to prevent attacks and unauthorized access. Therefore, insiders are considered as trusted partners. This can lead to serious security breaches, since in-zone users could launch attacks or circumvent the security mechanisms. Furthermore, the changes in network policies and traffic conditions require complex configuration of the firewalls, that make it further complex to keep the security in place. SDN enables automation through programmability and centralizes the network control that achieves global visibility of the network traffic behavior. Thus, traffic coming from the outside and traffic originated inside the network can be easily monitored. For example, the network ingress ports in OpenFlow switches can be dynamically configured from the controller to forward packets to a firewall in a separate middle box or an SDN firewall application. Similarly, traffic originating within the network can also be easily checked by updating the flow tables of the switch, in the same way as traffic coming from outside of the network. A number of firewall applications are already developed for SDNs, such as FLOWGUARD [24] and OpenFlow firewall [25].

4.4.5 Network Resilience

Network resilience mechanisms help the network to operate in the presence of diverse challenges, such as cyber-attacks, wrong configurations, operational overload, or equipment failures. The network must be capable to provide services to users when such challenges occur. Basic resilience strategies include Defend, Detect,

Remediate, Recover, Diagnose and Refine ($D^2R^2 + DR$) [26]. Through global visibility and a cycle of harvesting intelligence from the underlying network, the SDN control plane stays updated of the network situation. With programmable APIs in network equipment, fast reaction to failures in network equipment and miss-configuration is achieved. SDN-based resilience frameworks are developed that provide policy-controlled management with policy-based network configuration according to resilience strategies. An OpenFlow application is presented in [27] to enable interaction between multiple resilience mechanisms. The framework described in [27] also enables translation of high-level policies to device level configuration to act promptly to various failures. Other approaches include using multiple controllers to increase resilience of the SDNs.

4.4.6 Security Systems and Firewalls

To explain the possibilities of anomalies due to applications, consider the already existing example. Cellular network applications and middleboxes are independently managed by cellular operators and application developers. Application developers are unaware of the middlebox policies enforced by the operators. The operators have less knowledge of the application behavior and requirements. Such a mismatch or lack of understanding can create potential security challenges. For example, an operator can set an aggressive timeout value to quickly release the resources occupied by inactive TCP connections in the firewall. This could cause frequent disruptions in important application sessions [28].

Normally traffic is routed to various middleboxes to perform network security evaluation or check the traffic behavior and legality. However, the traditional middleboxes have a number of challenges regarding its placement, scalability and security policy alteration. These challenges mostly occur due to complex manual configurations, the need of path-specific middlebox placement, and non-flexibility of the existing network architectures [29]. SDN makes the deployment of middleboxes simple and elegant through network programmability and centralized network control. In [30], it is proposed to integrate the processing of middleboxes into the network itself, by using the concepts of SDN for policy consistency and higher visibility of the behavior of middleboxes through a centralized control. The simplicity of deploying and managing diverse and complex middleboxes, due to the above-mentioned characteristics of SDN, is presented in [31].

4.4.7 Network Security Automation

Automation is the process of minimizing human-machine interaction by delegating complex control functions to machines for reliability and accuracy. The main purpose of machine execution of complex functions, called automation, is accuracy and reliability through:

- i) information acquisition;
- ii) information analysis;
- iii) decision and action selection; and
- iv) action implementation [37].

However, automation has been used in a limited variety of networks even though human errors cause many network security and traffic management problems [35]. In today's multi-vendor networks, 62% of network downtime comes from human errors and 80% of corporations' IT budget is spent on maintenance and operations [36]. Stable and robust security policy deployment requires global analysis of policy configuration of all the networked elements to avoid conflicts and inconsistency in the security procedures and to diminish the chances of serious security breaches and network vulnerabilities [33]. As a security concern, a small oversight can lead to a global security problem such as placing a significant functionality on an unreliable system [34]. Therefore, automation of network and user security is highly important to avoid these challenges.

However, there are many challenges that make it difficult to automate and deploy automated security systems in today's communication networks. For example, most of the network systems used today are hardwired with specific control logic that require manual configuration of individual boxes. Such independent control systems in communication networks make it difficult to deploy consistent network-wide security policies throughout large networks that comprise a mix and match of control systems for different functionalities. Therefore, there are many proposals for the redesign of the communication systems and architectures.

Network security is an important and integral part of the network management that must be considered from planning to the deployment and use of the network [32]. Similarly, consistent policies over the network are highly important to avoid policy collusion that lead to security lapses. Among the proposals for such networks, SDN enables consistent network-wide policies through global visibility of the overall network systems and the policies implemented in each. By enabling programmability, and abstracting away the low-level configurations from individual boxes, SDN provides designing languages and network controllers that are capable of automatically reacting to the changing network state [38,39]. By abolishing the need of individual node configuration, taking the intelligence out of the networking components used to forward data and abstracting the control from the networking nodes, SDN paves the way for network security automation [40]. SDN enhances the automation of many processes and procedures, including physical and virtual network management and reconfiguration, and introduces the possibility of deploying new automated services. As a result, there are already several proposals for network automation using the concepts of SDN.

Proceria [41] is a network control framework for operators, which implements flexible policies based on the network view. Proceria maps high-level event driven policies to low-level network configuration driven policies, thus abolishing the need for manual configurations. The OpenFlow Management Infrastructure (OMNI) [42] simplifies OpenFlow management and provides mechanisms for a responsive autonomic control platform. Among the set of tools provided by OMNI, a web interface for the tools and a multi-agent system to autonomously control the network, OMNI tools for collecting statistics of flows and another that probes the network to obtain the physical topology, can be used for synchronizing the network traffic with the network security policies. The flows can be migrated to different physical paths according to the QoS and security requirement and without packet loss or security compromises. Using new technologies such as mentioned above, security systems can be automated in 5G.

Since the number of devices connected in 5G will be huge (the latency requirements will be strict that will require quick threat detection and faster response) security automation will be the need. Global visibility of the entire network behavior will enable quick threat detection and network programmability will enable faster threat mitigation. Since these features will be brought about by SDN, SDN-based automation will be highly valuable in 5G networks.

4.5 Conclusions and Future Directions

5G should be designed to provide more options of security beyond the currently used node-by-node security systems. It must provide more security than 4G and enable specific security designs for use cases that have specific requirements such as low latency, small cell sizes, and radio constraints. Sound security technologies and solutions must be built into the architecture of 5G from the beginning. This requires proper analysis of existing and future security threats to develop futuristic security solutions for 5G. New technologies will be integrated into the 5G ecosystem that requires new solutions for security as well. Flexible and agile technologies should be the core of network designs that enable adaptation not only to the service requirements but its security also.

The architectural limitations of current networks must not propagate to the future networks. The main limitation is the inflexibility of current networks due to its proprietary and closed nature. Such limitations make it difficult to deploy and use novel mechanisms and technologies when the need arises. These needs may be due to changes in user or business requirements, new service models or the limitations in the technology itself growing beyond its meaningful use. The concepts of SDN and NFV enable quick network updates, smoothen deployment of new technologies and services, and enable parallel deployment of old and new technologies and services. Being programmable in nature, these technologies also open the networking arena for innovation. This is the reason that new security systems and services are already developed by the industry, academia, and individuals working on open source projects.

However, enabling programmability of networking components is not entirely risk free in environments such as envisioned by 5G. For example, the critical infrastructure connected by 5G networks must be secured from malicious programs and access to programmable APIs in critical infrastructure for users or third-party developers must be constrained.

Network security has been rarely researched in parallel to network load balancing. It is extremely important in SDN-based mobile networks to develop load balancing architectures that work according to network security policies and vice-versa. For example, load balancing technologies can be used to avoid the saturation attacks. Similarly, security lapses of the controller can introduce delays in setting flow rules in the switches, leading to congestion in switches with unsolicited traffic flows. Therefore, it is necessary to consider network security in parallel with network traffic load balancing technologies in SDN-based mobile networks.

Most of the IoT devices will be constrained by capacity. Therefore, various types of wireless networking technologies are proposed having different cell sizes, differing architectures and heterogeneous infrastructures to perform functions on behalf of

constrained IoT devices. However, these devices can also easily become potential weak points for the networks. It is important to keep this limitation in mind while integrating networks of vulnerable IoT devices to the mainstream cellular or communication networks. Since these devices can easily be compromised to launch a DoS attack, it is necessary to either segregate them or use virtualization technologies so that they do not induce security vulnerabilities into the main network.

4.6 Acknowledgement

This work has been carried out under the projects SECURE-Connect (Secure Connectivity of Future Cyber-Physical Systems) and the Naked Approach Project (The Naked Approach Nordic perspective to gadget-free hyper connected environments).

References

- 1 Agiwal, M., Roy, A. and Saxena, N. (2016) Next generation 5G wireless networks: A comprehensive survey. In: *IEEE Communications Surveys & Tutorials*, 18(3), 1617–1655.
- 2 Kutscher, D. (2016) I's the network: Towards better security and transport performance in 5G. *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, San Francisco, CA, pp. 656–661.
- 3 Ahmad, I., Kumar, T., Liyanage, M., Okwuibe, J., Ylianttila, M. and Gurtov, A. (2017) 5G security: Analysis of threats and solutions. *2017 IEEE Conference on Standards for Communications and Networking (CSCN)*, Helsinki, pp. 193–199.
- 4 Alliance, N.G.M.N. (2015) 5G white paper. *Next Generation Mobile Networks*.
- 5 Liyanage, M., Ylianttila, M. and Gurtov, A. (eds) (2015) *Software Defined Mobile Networks (SDMN): Beyond LTE Network Architecture*. John Wiley & Sons, USA.
- 6 ITU Telecommunication Standardization Sector (2003) *Security Architecture for Systems Providing End-to-end Communications*. Geneva, Switzerland.
- 7 Alliance, N.G.M.N. (2016) 5G security recommendations Package, White paper.
- 8 Casado, M., Koponen, T., Ramanathan, R. and Shenker, S. (2010) Virtualizing the network forwarding plane. *Proceedings of the Workshop Programme*. Routers Extensible Serv. Tomorrow, p. 8.
- 9 van Cleeff, A., Pieters, W. and Wieringa, R. (2009) Security implications of virtualization: A literature study. *Proceedings of the International Conference CSE*, 3, 353–358.
- 10 Vaughan-Nichols, S. (2008) Virtualization sparks security concerns. *Computer*, 41(8), pp. 13–15.
- 11 Sherwood, R. et al. (2009) Flowvisor: A network virtualization layer. *OpenFlow Switch Consortium*, Technical Report OPENFLOW-TR-2009-1, Stanford University, Stanford, CA.
- 12 Chung, C.-J., Khatkar, P., Xing, T., Lee, J. and Huang, D. (2013) NICE: Network intrusion detection and countermeasure selection in virtual network systems. *IEEE Trans. Dependable Secure Computing*, 10(4), 198–211.
- 13 Kreutz, D., Ramos, F. and Verissimo, P. (2013) Towards secure and dependable software-defined networks. *Proceedings of the 2nd ACM SIGCOMM Workshop. Hot Topics Software Defined Networks*, pp. 55–60.

- 14 Iwamura, M. (2015) NGMN view on 5G architecture. In: *Vehicular Technology Conference (VTC Spring)*, 81st Proceedings of the IEEE, pp. 1–5.
- 15 Nokia (2016) Signaling is growing 50% faster than data traffic [Online]. Available at: <https://blog.networks.nokia.com/mobile-networks/2012/12/05/a-signaling-storm-is-gathering-is-your-packet-core-ready/> [accessed June 2017]. Published December 2012.
- 16 Zhou, X., Zhao, Z., Li, R. et al. (2014) Toward 5G: When explosive bursts meet soft cloud. *Network, IEEE*, 28(6), 12–17.
- 17 Ahmad, I., Namal, S., Ylianttila, M. and Gurtov, A. (2015) Security in software defined networks: a survey. *IEEE Communications Surveys & Tutorials*, 17(4), 2317–2346.
- 18 Sezer, S. et al. (2013) Are we ready for SDN? Implementation challenges for software-defined networks. *IEEE Communications Magazine*, 51(7), pp. 36–43.
- 19 Security-Enhanced Floodlight (2013) SDx Central, Sunnyvale, CA. [Online]. Available at: <http://www.sdncentral.com/education/towardsecure-sdn-control-layer/2013/10/>
- 20 Switch, B. (2012) Developing floodlight modules. Floodlight OpenFlow controller. [Online]. Available at: <http://www.projectfloodlight.org/floodlight/>
- 21 Porras, P., Shin, S., Yegneswaran, V., Fong, M., Tyson, M. and Gu, G. (2012) A security enforcement kernel for OpenFlow networks. *Proceedings of the 1st ACM Workshop on Hot Topics in Software Defined Networks*, pp. 121–126.
- 22 Al-Shaer, E. and Al-Haj, S. (2010) FlowChecker: Configuration analysis and verification of federated OpenFlow infrastructures. *Proceedings of the 3rd ACM Workshop on Safe Configuration*, pp. 37–44.
- 23 McKeown, N., Anderson, T., Balakrishnan, H., Parulkar, G., Peterson, L. et al. (2008) OpenFlow: Enabling innovation in campus networks. *ACM SIGCOMM Computer Communication Review*, 38(2), 69–74.
- 24 Hu, H., Han, W., Ahn, G.J. and Zhao, Z. (2014) FLOWGUARD: building robust firewalls for software-defined networks. *Proceedings of the 3rd ACM Workshop on Hot Topics in Software Defined Networking*, pp. 97–102.
- 25 OpenFlow Firewall: A Floodlight Module. [Online]. Available at: <http://www.openflowhub.org/display/floodlightcontroller>
- 26 Sterbenz, J.P. et al. (2010) Resilience and survivability in communication networks: Strategies, principles, and survey of disciplines. *Computer Networks*, 54(8), 1245–1265.
- 27 Smith, P., Schaeffer-Filho, A., Hutchison, D. and Mauthe, A. (2014) Management patterns: SDN-enabled network resilience management. *Proceedings of the IEEE NOMS*, pp. 1–9.
- 28 Wang, Z., Qian, Z., Xu, Q., Mao, Z. and Zhang, M. (2011) An untold story of middleboxes in cellular networks. *Proceedings of the ACM SIGCOMM Conference (SIGCOMM 2011)*. ACM, New York, pp. 374–385.
- 29 Joseph, D.A., Tavakoli, A. and Stoica, I. (2008) A policy-aware switching layer for data centers. *Proceedings of the ACM SIGCOMM*, New York, pp. 51–62.
- 30 Lee, J., Tourrilhes, J., Sharma, P. and Banerjee, S. (2010) No more middlebox: Integrate processing into network. *ACM SIGCOMM Computer Communication Review*, 40(4), 459–460.
- 31 Gember, A., Prabhu, P., Ghadiyali, Z. and Akella, A. (2012) Toward software-defined middlebox networking. *Proceedings of the 11th ACM Workshop HotNets-XI*, pp. 7–12.
- 32 Casado, M., Freedman, M.J., Pettit, J., Luo, J., McKeown, N. and Shenker, S. (2007) Ethane: Taking control of the enterprise. *ACM SIGCOMM Computer Communication Review*, 37(4), 1–12.

- 33 Hamed, H. and Al-Shaer, E. (2006) Taxonomy of conflicts in network security policies. *IEEE Communications Magazine*, 44(3), 134–141.
- 34 Creery, A. and Byres, E. (2005) Industrial cybersecurity for power system and SCADA networks. *Petroleum and Chemical Industry Conference. Proceedings of the IEEE 52nd Annual Industry Applications Society*, pp. 303–309.
- 35 Luo, J., Pettit, J., Casado, M., Lockwood, J. and McKeown, N. (2007) Prototyping fast, simple, secure switches for ethane. *Proceedings of the 15th Annual IEEE Symposium on High-Performance Interconnects HOTI*, pp. 73–82.
- 36 Casado, M., Freedman, M.J., Pettit, J., Luo, J., Gude, N. et al. (2009) Rethinking enterprise network control. *IEEE/ACM Transactions on Networking (TON)*, 17(4), 1270–1283.
- 37 Parasuraman, R., Sheridan, T.B. and Wickens, C.D. (2000) A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 30(3), 286–297.
- 38 Kim, H. and Feamster, N. (2013) Improving network management with software defined networking. *IEEE Communications Magazine*, 51(2), 114–119.
- 39 Shenker, S. et al. (2011) The future of networking, and the past of protocols. *Open Networking Summit*.
- 40 Ortiz Jr., S. (2013) Software-defined networking: On the verge of a breakthrough? *Computer*, 46(7), 10–12.
- 41 Voellmy, A., Kim, H. and Feamster, N. (2012) Procera: a language for high-level reactive network control. *Proceedings of the ACM 1st Workshop on Hot Topics in Software Defined Networks*, pp. 43–48.
- 42 Mattos, D.M., Fernandes, N.C., da Costa, V.T., Cardoso, L.P., Campista, M.E.M. et al. (2011) Omni: OpenFlow management infrastructure. *International Conference of the IEEE on Network of the Future (NOF)*, pp. 52–56.
- 43 Gutz, S., Story, A., Schlesinger, C. and Foster, N. (2012) Splendid isolation: A slice abstraction for software-defined networks. *Proceedings of the 1st Workshop HotSDN*, pp. 79–84.
- 44 Pfaff, B., Pettit, J., Amidon, K., Casado, M., Koponen, T. and Shenker, S. (2009) Extending networking into the virtualization layer. *Proceedings of Hotnets*, pp. 1–6.
- 45 Jafarian, J.H., Al-Shaer, E. and Duan, Q. (2012) OpenFlow random host mutation: transparent moving target defense using software defined networking. *Proceedings of the 1st Workshop Hot Topics Software Defined Networks*, pp. 127–132.
- 46 Wang, C.X., Haider, X., Gao, A., You, X.H., Yang, E. et al. (2014) Cellular architecture and key technologies for 5g wireless communication networks. *IEEE Communications Magazine*, 52(2), 122–130.
- 47 Schneider, P. and Horn, G. (2015) Towards 5G Security. *IEEE Trustcom/BigDataSE/ISPA*, Helsinki, pp. 1165–1170.
- 48 Piqueras Jover, R. (2013) Security attacks against the availability of LTE mobility networks: Overview and research directions. *Proceedings of the 16th International Symposium on Wireless Personal Multimedia Communications (WPMC)*. Atlantic City, NJ, pp. 1–9.
- 49 Cao, J., Ma, M., Li, H., Zhang, Y. and Luo, Z. (2014) A survey on security aspects for LTE and LTE-A Networks. *IEEE Communications Surveys & Tutorials*, 16(1), 283–302.
- 50 Sgambelluri, A., Giorgetti, A., Cugini, F., Paolucci, F. and Castoldi, P. (2013) Effective flow protection in OpenFlow rings. *Proceedings of the OFC/NFOEC*, pp. 1–3.

- 51 Li, J., Hyun, J., Yoo, J.-H., Baik, S. and Hong, J.-K. (2014) Scalable failover method for data center networks using OpenFlow. *Proceedings of the IEEE NOMS*, pp. 1–6.
- 52 Nikander, P., Gurtov, A. and Henderson, T (2010) Host Identity Protocol (HIP): Connectivity, mobility, multi-homing, security, and privacy over IPv4 and IPv6 networks. *IEEE Communications Surveys & Tutorials*, 12(2), 186–204.
- 53 Namal, S., Ahmad, I., Gurtov, A. and Ylianttila, Y. (2013) Enabling secure mobility with OpenFlow. *IEEE SDN for Future Networks and Services (SDN4FNS)*, Trento, pp. 1–5.
- 54 Liyanage, M., Abro, A.B., Ylianttila, M. and Gurtov, A. (2016) Opportunities and challenges of software-defined mobile networks in network security perspective. *IEEE Security and Privacy*, 14(4), 34–44.
- 55 Bikos, A.N. and Sklavos, N. (2013) LTE/SAE security issues on 4G wireless networks. *IEEE Security and Privacy*, 11(2), 55–62.
- 56 Liyanage, M., Braeken, A., Jurcut, A.D., Ylianttila, M. and Gurtov, A. (2017) Secure communication channel architecture for software defined mobile networks. *Elsevier Journal on Computer Networks (COMNET)*, 114, 32–50. Available at: <http://dx.doi.org/10.1016/j.comnet.2017.01.007>. (<http://www.sciencedirect.com/science/article/pii/S1389128617300075>)
- 57 Liyanage, M., Ylianttila, M. and Gurtov, A. (2014) Securing the control channel of software-defined mobile networks. *Proceedings of the IEEE 15th International Symposium on World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, Sydney, Australia.
- 58 Lam, J-H., et al. (2015) Securing distributed SDN with IBC. *Seventh International Conference on Ubiquitous and Future Networks*. IEEE

5

Cyber Security Business Models in 5G

Julius Francis Gomes¹, Marika Iivari¹, Petri Ahokangas¹, Lauri Isotalo², Bengt Sahlin³, and Jan Melén³

¹ Martti Ahtisaari Institute of Global Business & Economics, Oulu Business School, University of Oulu, Finland

² Elisa Corporation, Finland

³ Oy LM Ericsson AB

5.1 Introduction

We live in the era of 4G, and some parts of the globe are still mainly entertained by 3G or 2G technologies. Yet, we are already heading towards 5G and with all the hyperconnectivity and high-tech innovations and services, we are increasingly exposed to serious threats of cyber-attacks. Dobrian [1] remarks that cyber security has been a major issue for sectors such as financial services, defence, healthcare, media and online social media. From a technical point of view, significant research effort has been made so far into standardizing 5G and cyber security [2]. For instance, 5GPPP, a public-private partnership body funded by the EU and several other research projects are working in areas such as the physical layer, overall architecture, network management and software networks [3]. Although one of 5GPPP's recent publications mentions that 5G is business driven in a European context, research results from a business and management perspective still remain negligible [4].

Moreover, there are also gaps in the knowledge as to how 5G is expected to impact the management organization at the security level. These issues raise some important questions. In a world flooded with 5G innovations, increased amounts of information, ultra high speed and low latency, how is the security scene expected to improve? Is it about alternative ways of organizing existing techniques that facilitate security delivery? Or will 5G facilitate new kinds of innovation for cyber security? Furthermore, since 5G is deemed to be business driven, how to monetize security as a service in 5G is a significant question. It can be argued that the more monetizable the security is, the more business entities will be interested to deliver secure services and products by investing more in it.

We utilize the concept of business models as a boundary-spanning unit of analysis [5], in order to construct a founded understanding of 5G security business. The business model as a concept links abstract strategies to the practical level of decisions and actions within the uncertainties of the modern business context [6–10]. Business models are

fundamental for many technological businesses, as even a mediocre technology can succeed in the markets if it has a well-designed business model, but perfectly designed technology may fail because of a weak business model [11]. This is even more crucial for the cyber security business where failing in business may cause serious risks beyond economical risks. The business model as a concept in the literature has been defined as an architectural [12] system of interdependent activities [13] and an interrelated set of core logic and strategic decision variables [10,14], explaining transaction content, transaction governance and transaction relationship structures [15,16] for maximized value creation and value capturing [13]. Therefore, in this chapter we aim to display the possible business impacts of cyber security in 5G, in order to increase awareness of how and why business model thinking matters.

The rest of the chapter is structured as follows: first, we discuss the context of cyber security business in 5G by explaining the types and costs of cyber threats. In the next part, we delve deeper into the business aspect of cyber security by opening up theoretical discussions on business models. We elaborate on how the business model approach helps to identify more tenable business opportunities through the 4C framework. We also present an overview of the business case for cyber security in 5G. We display four scenarios for 5G security provisioning for the future, which leads us to draw the technical landscape from service and user perspectives. Furthermore, as the main contribution of this chapter, we present a business model framework for identifying avenues for cyber security business in 5G and the relevant business model options.

5.2 The Context of Cyber Security Businesses

Cyber security is usually delivered by specific third-party providers as a service, as a product or as a combination of a service and product, which helps organizations and individuals to protect their digital assets. From this perspective, there are two distinct types of organization whose business is dependent on security. First are the organizations who are directly or indirectly involved in the delivery of security solutions. These companies either sell the security as a service or as a bundle with some other service, product or infrastructure. Second, there are all the other business organizations who have remarkable digital footprints and face the potential threats of cyber-attacks. In the imperfect world we live in, any cyber-attack will have a direct or indirect impact on either or both of these types of organizations' business. In this chapter, we show different ways to organize security offerings in the 5G era and how to monetize that.

In this section, we first provide the readers with a general perspective on the economic impact of the cyber security phenomenon and present a brief discussion on the types of cyber threat and the cost of cyber-attacks. Cyber threats can be defined as events or the deliberate exploitation of vulnerabilities by threat agents or attack vectors leading to the disruption of an organization's operations, or the loss or takeover of an organization's assets [17]. The threat to an organization's assets (like information and IT infrastructure) may arise from a natural occurrence such as an earthquake, equipment failure or the unintentional actions of employees. However, a deliberate, planned attack – which poses the highest risk and is able to wreak incalculable loss to organizations – is of interest in this chapter.

5.2.1 Types of Cyber Threat

Cyber threats are classified differently in the literature, but the one most used is the model that classifies “cyber threats” on five levels, based on the motivational factors of the threat agents [17,18]. These are cyberactivism, cybercrime, cyber-espionage, cyberterrorism and cyberwarfare, as summarized in Table 5.1.

Cyberactivism is the first level of threat and entails cyber-vandalism, so-called hacktivism, and hacking. The intent of the actors may not be to cause any damage, as they may be using the attack to embarrass an organization or send a political message [18,19]. “Hacktivists” are individuals or groups who hack into publicly available websites and overload email servers in order to send a politically motivated message or use it to convey a protest message, for example, against limiting civil liberties [17,19].

The second level of threat is cybercrime. It involves the use of information systems and networks by adversaries for the commission of crime against a victim’s IT infrastructure. This act can be perpetuated by individuals, loosely organized groups, terrorists, insiders or spammers. The motive for such an attack could be to steal vital information or to disrupt the functions or operations of organizations for financial gain or an ideological cause [19,20]. According to Lehto [17], cybercrime can be categorized into three groups: the use of ICT to commit traditional crimes like fraud and forgery; the publishing of illegal material using electronic media; and attacks directed at the electronic network.

Cyber-espionage is the next level of threat. It is the use of illegal means on the Internet, networks, programs or computers to gain secret information from individuals, organizations, competitors and governments for political, military or monetary gain [17,21]. Cyber-espionage is carried out by professional intelligence agents, individuals or groups who exploit the vulnerabilities in their adversary’s system in order to obtain high-value information. It is a tactic employed by nation states and their militaries to gather intelligence on their perceived or real enemies [18,19]. Cyber-espionage is not limited to political

Table 5.1 The basic characteristics of different types of cyber-attack.

| Cyberactivism | Cybercrime | Cyber-espionage | Cyberterrorism | Cyberwarfare |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"> • Hacking • Might not cause any damage • Might intend to embarrass an organization or convey a political message • Examples: Anonymous, TeamPoison, Chaos Computing Club | <ul style="list-style-type: none"> • Usually criminal activities • Can be done by individuals, loosely organized groups, terrorists, insiders • Motives: to steal vital information, disrupt functions for financial gain or ideological purposes | <ul style="list-style-type: none"> • Illegal access to networks or computers to get secret information • Can be employed by nation states in some cases • Besides political espionage, business players can initiate such acts | <ul style="list-style-type: none"> • Cyber-attacks targeted at intimidating governments or causing fear and panic • Usually initiated by sophisticated terrorist groups • Attackers use offensive IT weaponry | <ul style="list-style-type: none"> • Warfare in cyberspace • Threat agents: nation-state military, intelligence services, organized insurgent groups, terrorists • Employed as a combination of multiple attack strategies in practice |

espionage, but also extends to economic domains. Professional organized crime groups or agents of an organization's competitors can hack into a business entity's system and steal their proprietary information, such as intellectual property or trade secrets.

The fourth level of threat is cyberterrorism. This is the use of cyber-attacks targeted at IT systems or the critical infrastructure of government and private organizations, with the intent of intimidating the government or causing fear and panic among the civilian population [19]. This type of attack is perpetuated by sophisticated terrorist groups whose aim is to grab national or international attention [22]. They utilize offensive IT weaponry, either in isolation or in combination with other means of attack [17]. For example, in 2011, the Canadian Government reported a major cyber-attack against its agencies, including Defense Research and Development Canada. The attack forced the Finance Department and Treasury Board, Canada's main economic agencies, to disconnect from the Internet [23].

Cyberwarfare is the fifth level of cyber threat and involves the conduct of warfare in the virtual world or cyberspace [17]. The typical threat agents are nation states' militaries and intelligence services, organized insurgent groups or terrorists. The action aims at immobilizing the information system or destroying the critical infrastructure of the enemy through the use of weapons such as computer viruses, worms or denial-of-service (DOS) attacks. Cyberwarfare is not a stand-alone strategy but is used with other strategies (e.g. "kinetic" warfare) in an offensive or defensive operation [18]. For example, in 2007, the Estonian government suffered some serious cyber-attacks against its websites and some banks' websites, leading to a halt in online banking transactions. This incident arose when the government decided to relocate a WWII Soviet Union memorial [24].

The forgoing discussion has centred on the various types, levels of severity and complexity of cyber threats. The threat can emanate from various sources such as nation states, organizations, organized crime groups, individuals, terrorists, insurgent groups and competitors. The motive of these actors may be to enhance their ego or have some bragging rights, to advance a political or ideological cause, monetary gain, to gain access to sensitive information for a future course of action, to cause fear and panic among people or to force a government to take or abandon a certain cause. They can be used as a strategy in conflicts and warfare. The severity of these attacks may differ from one to another and the intent may be to cause minimal or collateral damage. Nevertheless, in all instances, a cyber-attack results in some form of loss, such as financial loss, infrastructure or equipment damage, or loss of reputation. In the next sub-section, we examine the costs of cyber-attacks to nation states and businesses.

5.2.2 The Cost of Cyber-Attacks

Incidents of cyber-attacks are increasing with a concomitant increase in cost to governments and businesses. The actual cost of these attacks is difficult to quantify; however, numerous studies have churned out estimated costs [25]. The US Chamber of Commerce estimated that the losses to the US resulting from cybercrime alone ranges from between US\$24 billion and US\$ 120 billion and the global cost is reported to be US\$1 trillion [26]. Also, the Intellectual Property Commission estimates that the US loses around US\$ 300 billion annually through intellectual property theft. In a recent study of 58 benchmarked US organizations, the Ponemon Institute found that the average cost of cybercrime to these organizations was US\$ 15 million. This showed an increase of 19% in the 2014 survey figure [27].

The cost related to cyber-attack can be broadly divided into two costs: the preventive cost and the post-attack cost [25]. The preventive cost is investing in infrastructure and systems, for example, to reinforce the perimeter defence of an organization to prevent an intrusion or reduce the impact of a breach. Although the cost of preventive solutions may be high, it may not be as expensive as the post-attack remedies. The post-attack cost involves both the actual cost (e.g. the amount stolen and extortion) and the cost of measures to remedy the consequences of the attack. This includes the cost of the replacement or repair of damaged infrastructure or systems; the cost of repairing lost business, such as customer acquisition activities and image rebuilding efforts; and the cost of detection and reporting, such as forensic and investigative activities, audits of installations and systems, crisis management and communication to stakeholders [27].

5.3 The Business Model Approach

Lying at the intersection of entrepreneurship and strategy, the business model concept can be seen as a bridge between abstract strategies and the practical level of decisions and actions amidst the uncertainties of the modern business context [6–10]. For instance, Zott and Amit [13] conceptualize the business model as a “boundary-spanning” set of activities aimed at creating and appropriating value. Morris *et al.* [14] viewed the concept of the business model as a set of decisions related to the venture strategy, architecture and economics of a firm (value creation and capture) that need to be addressed to create sustainable competitive advantage in the chosen markets.

Zott and Amit [13] further argue that a business model functions to explore and exploit a business opportunity. The business model as a concept thus covers a variety of elements, and there are myriads of conceptualizations available. The key issues these conceptualizations cover can be summarized accordingly, as seen in Figure 5.1. When built around opportunity, the business model can claim to consist of the elements of when, what, how, why and where the firm is acting to create and capture value when exploring and exploiting opportunities [28]. This indicates the applicability of the

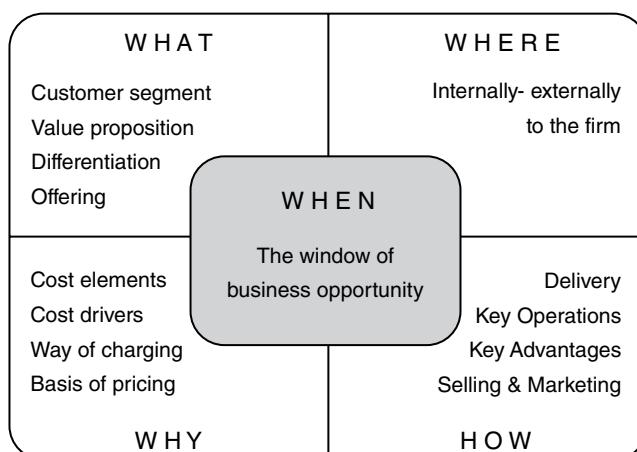


Figure 5.1 The business model's definition (adapted from [28]).

business model in finding ways to create value for customers and in return in finding ways to capture available value from the market, thus successfully exploiting a business opportunity.

The business model approach entails looking at organizational issues through the lenses of different business model elements. The most frequently mentioned business elements in the literature include value creation, value proposition/offering, value capture, partners/actors/suppliers/value networks, customers, customer relationships, processes/activities/value chain, revenue stream, cost structure/cost drivers/fixed costs, differentiation/cost leadership/pricing, competencies/capabilities/resources/assets and technology infrastructure [5,10,14,29].

The business model, seen as a boundary-spanning unit of analysis [5], connects an organization with its business environment, other organizations, customers and individuals, as well as to society at large [30]. Thus, since the list of business model elements cover a wide stream of external organizational elements, it usually comes in handy in analyzing different situations. When pondering bridging business models and cyber security, there are two main issues. First, since almost all of the entities operating within the digital sphere face multifaceted cyber threats, how can the business model approach help organizations to respond to such situations? Second, how can the business model approach help to identify opportunities to monetize security in future 5G in particular?

5.3.1 The 4C Typology of the ICT Business Model

Along with the rise of the mobile telecommunications industry, business models have increasingly been discussed in connection with shifting organizational boundaries through the vertical and horizontal integration of the industry and complex provision of new services [31]. This integration in the ICT sector resulted in value-creation focused vertical business models, applied mainly by infrastructure and technology providers, and value-capture focused horizontal business models, applied mainly by service providers [31,32].

Wirtz *et al.* [33] discussed four business models for classifying Internet-based business models in particular. They divided Web 2.0 business models based on connection, content, context and commerce. Yrjölä *et al.* [34] interpreted these business models as a *layered 4C model*, where the “lower”, more technically focused level is required for the “higher” one to exist. These layers are the domains where opportunities for value creation and capture in the industry can be identified, highlighting simultaneous value creation and capture. This 4C model can be applied either to examine a single-layer player or a player that is active in all four layers [34]. Thus, the 4C model (Figure 5.2) can be used to describe the structure and interaction in the ICT industry from the business model perspective

The first layer is concerned with a *connection*-related business model, where a stakeholder provides connection-related services [28]. The second layer is the business model, focusing on monetizing *content*. In the content layer, all sorts of online content services (e.g. mobile video streaming) are classified (i.e. relevant, up-to-date or interesting) and are conveniently accessible for the end user. The content might be peer-to-peer/ user-oriented content (i.e. the exchange of personal content), web browsing content (i.e. information storage), or online collected and selected educational and entertainment content (audio, video, text, etc.). The key is to understand who owns and can monetize

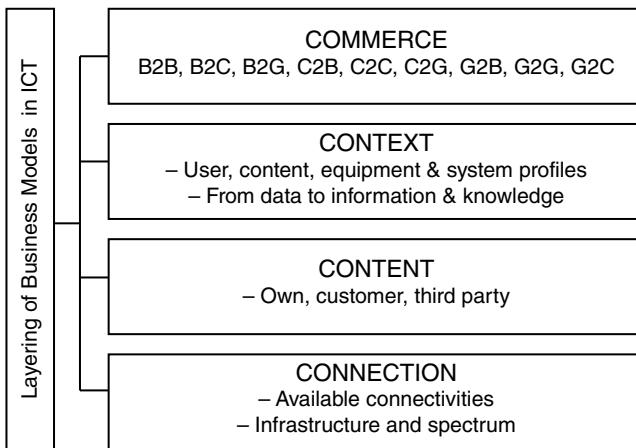


Figure 5.2 The 4C business model for ICT business (adapted from [28]).

content or whether it is freely accessible to end users, as in the case of advertisement content.

The third layer, *context*, concerns the ability to create and monetize user, content, equipment/user device and system profiles and turn (big) data into meaningful information and knowledge. In this layer, the information that already exists on the Internet can also be structured and aggregated to create less complex services through search and navigation (e.g. Google) and provide a relevant and useful context for the content. It helps the users to identify the content that they need and leads to a more transparent market. Through 5G networks, all of the businesses would be available on the user's mobile and there would be a variety of services based on the user's profiles, time, place and history data. The context stakeholders will provide information about the network, service, device and user's profile.

The fourth layer concerns *commerce*, the ability to monetize any or all of the connection, content- or context-specific resources, actors or activities related to the ongoing communications. In this layer, we can recognize business (B), consumer (C) and public/ government (G) types of communication [35]. Thus, B2B (business-to-business), B2C and B2G communication – as well as C2B, C2C and C2G or G2B, G2C and G2G communication – may be monetized in this layer.

5.3.2 Business Models in the Context of Cyber Preparedness

On a general note, it is assumed that businesses are nowadays aware of their vulnerabilities to potential cyber-attacks. Still, some expose themselves and only react when they are attacked [36]. Others that appear proactive leave the task of protecting their systems and IT infrastructure to their security managers [37]. This ad hoc approach to managing cyber risk exposes businesses to cyber-attack and its attendant consequences [38].

Cyber preparedness is the deliberate institution of cyber security preventive measures targeted at the dynamic vulnerabilities and cyber threats businesses constantly face. It is a strategic and policy issue that needs to evolve from an organization's leadership and cascade down to the whole organization. Bodeau *et al.* [18] initially proposed a

cyber-preparedness framework, which adopts a four-stage methodology in designing such strategies. Additionally, Gomes *et al.* [39] elaborated on the framework with a more organizational, business and management perspective.

In the first stage, the leadership examines the organization's threat levels, the existing security risk management frameworks that guide strategy formulation. In the next step, leadership critically assesses its cyber preparedness level in relation to the threat levels identified. The cyber preparedness levels are *perimeter defence*, *critical information protection*, *response awareness*, *architectural resilience* and *pervasive agility* [18]. In the third stage, the organization evaluates the nature of the threat it faces and the technical, operational and process capabilities at its disposal to mitigate the threat it faces. Having developed the cyber preparedness policy, the final step entails drawing a roadmap to integrate the policy into the overarching strategic plan.

Gomes *et al.* [39] combined the concept of the business model with different management dimensions and in order to facilitate the formulation of an organization's cyber-preparedness strategies. The extant literature of business models addresses elements such as customers, channels, customer relationships, partners, competitors, complementors and resources (human, non-human, technical and non-technical), among many other elements. The key notion is that the business model as a unit of analysis can help managers look at potentially cyber-risky areas by considering the what, how, when and why elements. These elements can be located either within the internal boundaries of the organization or at the external boundaries.

Pinto [36] stated that excellent security technologies are available, but the bigger issue is to comprehend cyber security as a competitive advantage and a revenue-generating advantage. The business model is conventionally used as a concept to find new revenue streams in most cases, but in the case of cyber security, for most digital companies it can enhance their business competitiveness, as they utilize the business model as a boundary-spanning unit of analysis. Pinto [36] added that organizations need to consider cyber-security costs as the costs of upfront product development instead of securing after the event. In this way, cyber preparedness strategies can secure the system and can also offer differentiated competitive advantages and business opportunities.

5.4 The Business Case of Cyber Security in the Era of 5G

IHS Markit speculates that 5G will not only be a catalyst for advancing mobile communication technology, but also for various other sectors from technology and economic perspectives. It is assumed that 5G will be a revolutionary step in technology development, much like the printing press, the Internet, electricity, and the steam engine. The common thread among these technologies is that they have been catalysts for the disruptive transformation of global practices. They are often titled general-purpose technologies (GPTs). According to IHS's projection, 5G will enable 12.3 trillion dollars worth of global economic production in the year 2035. Furthermore, in 2035, the global 5G value chain alone shall generate an output of 3.5 trillion dollars while supporting 22 million jobs. As a GPT, 5G will directly affect businesses and operations in all industries that are digitally intensive. Thus, there should be a significant opportunity for cyber security to operate in the realm of 5G in the long term [40].

As previously mentioned, we observe cyber security as a phenomenon having significant impact on two broad types of organization. The ones who directly or indirectly sell or provide security-related solutions and all the other organizations that have significant digital footprint, are thus vulnerable to cyber threats.

The companies providing different cyber security related products/services have two broad customer groups. The more successful customer group so far has been that of organizational entities, which are widely referred to as B2B (business-to-business) customers or B2G (business-to-government) customers in business literature. This group of customers has shown growing interest in investing in cyber security in recent years for the purpose of their own business/organizational sustainability. Since business organizations or government agencies are more cautious about the security of their digital assets, there has been a steady business opportunity in this segment. However, the second broad group of customers is individual consumers, who are referred to as B2C (business-to-customer) customers. For security providers, penetrating this customer group has been comparatively difficult up until now. The difficulty is triggered because until now the value of individual people's private data/information on the Internet and other networks was perhaps seen to be dispensable by individuals themselves. Though, in the era of 5G, where trillions of Internet of Things (IoT) devices are going to flood the home environment, this scene is likely to change.

The second type of companies whose business is affected by cyber security are those who utilize cyberspace as a vital element for their businesses. This group can comprise of almost all types of organization in the modern world. Among these, different organizations are exposed to different levels of cyber threat. Unfortunately, even in today's globalized world, we still have organizations who are unable to evaluate the value of the digital information at their disposal and some fail to invest in it at all at times or do not invest enough in security [39].

Traditionally, organizations invest in cyber security solutions based on the need for security when improving cyber preparedness against potential attacks [39]. This strategy for investing in cyber security products/services has been beneficial for many organizations, but failing to estimate the need for security has resulted in fatality many times too. We believe that the lacking interest to invest an adequate amount in cyber security solutions arises from the failure to generate new revenue.

In Figure 5.3, we plot our understanding of the relationship between the level of security and new revenue generation. To make any digital solution (physical or virtual) with a higher security level, organizations need to invest an additional sum. The main intent of any business organization is to be profitable by creating new revenues against investment. Hence, if organizations find ways to generate new revenue for the improved security level of the solution, they would eventually be more interested in investing the required sum (although we argue that additional investments in security may not always straightforwardly result in new revenue when there is no need for additional security). So, in practice, organizations should analyze the level of their security need smartly and improve the security level accordingly. In the case where the need for security and the level of security is at an optimum balance, organizations can then push for new revenue generation.

Figure 5.3 thus reflects the case of new revenue generation and security level for different technological eras. In this figure, we plot four different presumptive cases for different technological eras as A, B, C and D. In the early days of the digital era, there were many organizations who created handsome new revenues without having highly

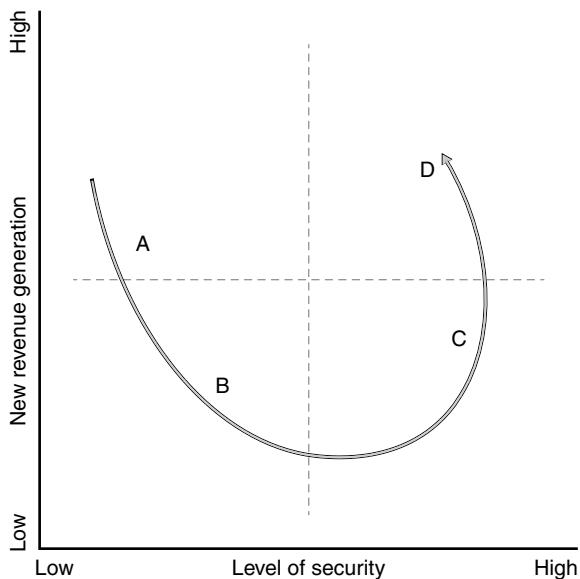


Figure 5.3 The relationship between the level of security and new revenue generation.

secure solutions; this is denoted as case A in the figure. However, as time passed by, customers became more aware of the basic security features of digital solutions and the new revenue generation took a dive downwards; this is case B. But, during the last few years, technology solution providers have started to see the need to secure solutions and thus act to improve the security level. This is denoted as case C; where it shows both improved security level and an upward trend for new revenue generation. Finally, we project case D for the future era of 5G, in which organizations need to provide highly secured services and solutions in order to be able to generate new revenue.

In the following subsections, we briefly explain the overall business context of cyber security in 5G by opening up different scenarios. We also take different security provisioning into consideration, in order to map different business model options for cyber security in 5G.

5.4.1 The Users and Issues of Cyber Security in 5G

Global society is flooded with hyperconnected devices generating a mass data flow over networks. In 2016, the number of IoT devices alone almost tripled that of the global population and is expected to exceed 50 billion by 2020 [41]. When looking at cyber security in 5G, it is important to identify the users and what is the impact of the security level of different solutions and vice versa in this respect. Broadly looking at the user base in the 5G era, we identify two major groups: human users, and other devices and machines. Major issues in the 5G era could be identification, the user experience of human users/the quality of the services of devices and machines, and privacy and safety/security.

To clarify this conceptualization, if we consider a “secured solution” made for human users, then it needs to have secured personal identification (Figure 5.4). The human user experience also needs to be competitive with other “unsecured solutions”. Finally,

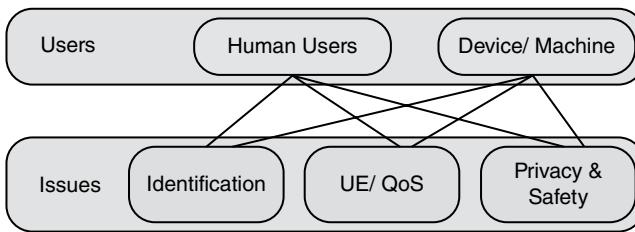


Figure 5.4 The users and issues in 5G security.

these secured solutions need to ensure personal privacy and safety, and have other security features. Similarly, if we consider a “secured solution” made for machine users, device-unique identification needs to be constantly secured. The quality of service also needs to be competitive with other commercial devices. Additionally, these secured solutions for device/machines should ensure relevant security features applicable to their contextual purposes. In our opinion, cyber security in 5G should solve security and privacy related issues beyond the conventional data security. By doing so, more and more business entities will be interested in offering secured solutions and finding new ways to deliver value to their customers and monetize that.

5.4.2 Scenarios for 5G Security Provisioning

To look at the overall cyber security provisioning for future 5G, it is amenable to being understood as the interplay between the industry and different types of security solutions and services. In doing so we draw Figure 5.5, where we identify four major security drivers for creating new businesses in 5G in the future.

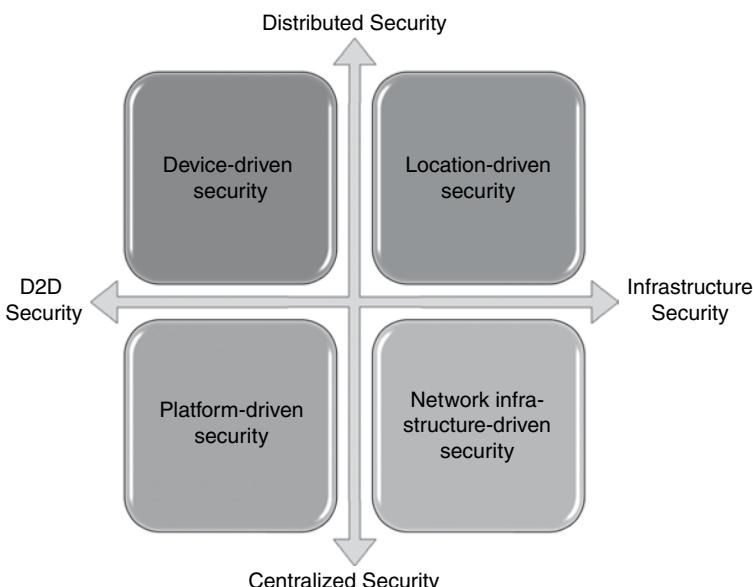


Figure 5.5 Scenarios for 5G security provisioning.

To formulate these scenarios, we first identified four broader schemes of security provisioning: device-to-device (D2D) security, infrastructure security, distributed security and centralized security. We observe the fundamental distinction between distributed and centralized security, hence putting them at polar opposites of one axis. On the other hand, we also see the vital difference between D2D security and infrastructure security and put them at opposite ends of the other axis.

As a result, we find four major drivers of security that will potentially come with new business opportunities in the 5G era. Device-driven security comprises distributed and D2D security techniques. Platform-driven security will focus on centralized and D2D security techniques. Network infrastructure-driven security should focus on centralized and infrastructure security methods. Lastly, location-driven security should harness distributed and infrastructure security techniques.

5.4.3 Delivering Cyber Security in 5G

So far in this chapter, we have taken a business-oriented standpoint. At this point we look at the technical aspects of security delivery from the service and user perspectives, in order to validate our scenarios. Figure 5.6 depicts a landscape of security arrangements in the context of 5G. Though we are still in the process of 5G standardization, this diagram can be considered to show where and how security features are needed in practice. There are different technologies and entities mentioned in the diagram, such as home carrier, roaming carrier, interconnect, signalling gateway, IT cloud, EDGE, NR (new radio), narrowband IoT, LTE, WiFi, radio access, radio terminals, M2M gateways and IoT sensors. The following gives a brief characterization of these terms:

- **Home Carrier:** An operator that holds user subscription and user data; a primary party, charging the user for network services.

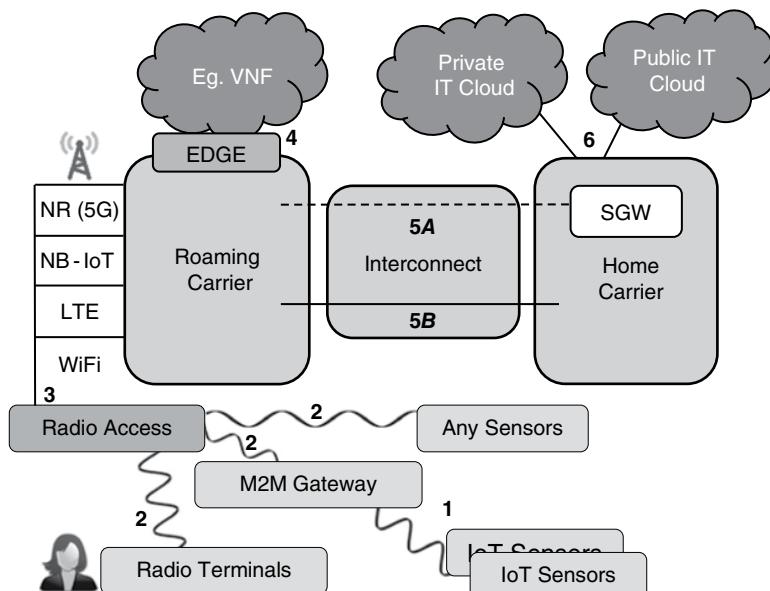


Figure 5.6 The 5G security arrangement landscape.

- *Roaming Carrier*: An operator that serves the user when outside the user's home carrier network.
- *Edge Network*: Infrastructure to support the user with data and services, when user data from the IT cloud needs to be processed locally.
- *Radio Access*: The 5G radio infrastructure that provides radio coverage and connectivity to the carrier network; different type of radio access technologies in 5G will include LTE, WiFi, narrowband IoT and new radio.
- *M2M Gateway*: A 5G device that connects with 5G radio infrastructure, and provides aggregation functions for IoT sensors over low-power, low-distance radio.
- *Radio Terminals*: A 5G device with which a human user connects with using 5G radio infrastructure.
- *IoT Sensors*: A low-power device that connects with the M2M gateway and provides connectivity for IoT devices/machines.
- *Cloud*: A computing entity that delivers hosted services over the network to users, both human and device/machine users.
- *Interconnect*: A technical structure and the related processes that allow carriers to transmit/receive traffic to/from other carriers.

Additionally, in Figure 5.6 we use numbers to denote different communication locations where security will be needed. Number 1 denotes the need for security from IoT sensors to the M2M gateway. Number 2 denotes secure connectivity from user radio terminals or the M2M gateway, or an individual sensor having the capacity of radio access. In number 3, we consider the need for security between radio access and the operator. Number 4 denotes the security need for the Mobile EDGE network. Numbers 5A and 5B denote secure connectivity between multiple carriers through an interconnection. Number 5A specifically marks security for signalling and 5B specifies the need for payload transferral. Lastly, number 6 marks secured communication between carriers and IT clouds. We included both a public IT cloud and private IT cloud, due to the fact that in practice there can be a few aspects of security that needs to be stressed when choosing either a private or public cloud.

Combining the security provisioning scenarios in Figure 5.5 with the security arrangements in Figure 5.6, gives us an idea of which sort of security provisioning is required and where security features are needed. This ideation can help identifying different business model options for business entities in the future. Utilizing this conceptualization, we can go forward to identify different business model options for cyber security in 5G.

Table 5.2 The security needs for different scenarios.

| Provisioning scenario | Security need location |
|----------------------------------------|------------------------|
| Device-driven security | 1 + 2 |
| Location-driven security | 3 + 4 |
| Platform-driven security | 1 + 2 + 6 |
| Network infrastructure-driven security | 5A + 5B + 6 |

5.5 Business Model Options in 5G Cyber Security

In the previous part of this chapter, we presented the concept of business models and the 4C layers (connectivity, content, context, and commerce) of business models for the ICT industry. In this section, we use this approach in combination with the four scenarios built. In order to identify different business model options, we create a 4×4 matrix with the 4C layers on one side and the four security provisioning scenarios on the other. We realize that for each scenario there could be possible business models in multiple layers of 4C. However, we also comprehend that in reality there will be some layers of 4C, which will not be covered in every scenario.

Figure 5.7 depicts our identification of the different business opportunities and business model options for cyber security in 5G. According to our analysis, the dark boxes in Figure 5.7 present the key business opportunities. Additionally, we have identified further business models for each security driver.

From the network infrastructure point of view, we observe that the most relevant business opportunities will arise for mobile network operators (MNOs), network infrastructure vendors (NIVs) and other carriers offering secure 5G connectivity in the connectivity layer. In doing so, we also see different technologies at play in the future-like network slicing, software-defined mobile networking, network functions virtualization, etc., where additional security features could generate new revenues. Furthermore, for infrastructure-driven security, there could be new business opportunities created by offering secure connectivity using mobile/moveable infrastructure for mass events such as special sports events, fairs, outdoor concerts, etc.

In the case of platform-driven security, key business opportunities will arise from securing content on digital platforms. Business models that could monetize in this area

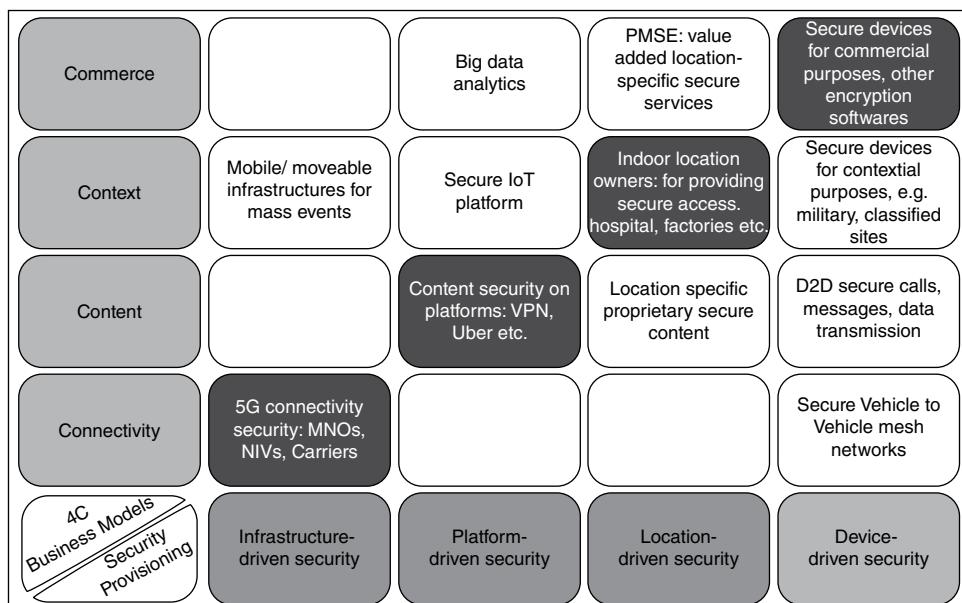


Figure 5.7 The business model options for cyber security in 5G.

come from secure digital platforms or VPN providers who secure data transmission for users by altering location data. Additionally, we see the potential of platforms with intra-platform payment options and the payments need to be secured as well. We give Uber as an example here; in reality there are many more examples where the platform needs to be secured for intra-platform payments in addition to secure payment gateways. A secure IoT platform is a potential business model for the context layer. Finally, these secure platforms can make new revenue by offering big data analytics from the data they create and collect in the commerce layer.

When we look at location-driven security, the most opportune business models should rise in the context layer. In our analysis, indoor location owners can exploit one of the challenges of 5G small-cell base stations: getting seamless connectivity in indoor spaces. Indoor locations, such as hospitals, universities, production factories, sports arenas or shopping malls, can then take the responsibility to operate as a local operator. We call this role that of the micro operator. On top of standard connectivity, these micro operators can add a layer of value-added location-specific service to a package, often referred to as PMSE in the commerce layer. Ahokangas *et al.* [28] defined micro operators as entities that offer “1) *mobile connectivity combined/locked with specific, local services*, that is 2) *spatially confined* to either its premises or to a defined (but narrow) area of operation, and is 3) *dependent on appropriate available spectrum resources*”. However, micro operators’ major business partners will be MNOs and NIVs for spectrum licensing and network infrastructure reasons. From the security perspective, micro operators can consider procuring services from MNOs or other security providers. For the content layer, there can be business around location-specific proprietary secure content, which is only accessible by authorized personnel within the specific location. One example could be offering video content replays during a sports event in a stadium.

Finally, for device-driven security, we observe the commerce layer to offer key business opportunities. These opportunities can be monetized by offering secure devices for commercial purposes. Devices can range from secure handheld mobile devices to secure IoT devices. Additionally, examples for businesses in the commerce layer can be different types of encryption software, applications and VPN providers, which secure devices and content in various ways. We consider secure devices to create opportunities in the context layer as well if the secure devices are targeted to secure specific contexts, such as military contexts or a specific classified location. We also see the potential for some businesses in the connectivity layer from the device-driven security perspective for the self-driving vehicles of the future, for which vehicle-to-vehicle communication will be key. Device-driven security techniques can be deployed in this context too, for secure communication in such dynamic-mesh networks between multiple vehicles.

This listing of business model options for cyber security in 5G should not be considered as exhaustive, but rather as suggestive and a starting point from which to look for more business models for new revenue generation. In order to create this framework, we have used the business model as a boundary-spanning unit of analysis [5,13,15] in creating the security provisioning approach. These scenarios enabled us to draw the technical landscape from both a service and user perspective. The technical landscape also validates the scenarios we built from a technological point of view.

In our opinion, given that an organization identifies a viable business opportunity and is able to invest sufficient sums to establish the required level of security, they can also

utilize this kind of business model framework for further identifying new revenue-generating business models. The business model options framework combines the business model perspective through the 4C model [33,34] and security provisioning from the scenarios. From a broader perspective, this framework provides us with some ideas about what, when, how and why elements of business models [28] for different kinds of organization in different industries. Additionally, the 4C typology of ICT business models helps us to create a structured formation of all the players involved in the market and also hints about where the opportunity for more players is. From a business opportunity creation/identification/discovery point of view, this kind of framework can be useful for identifying a viable opportunity and subsequently designing an organization-specific business model.

This chapter discussed various aspects in relation to business for cyber security in 5G. We particularly elaborated the business model approach to 5G security provisioning and provided an integrated framework for illustrating the types of services that we see emerging from the context of 5G. Security issues not only relate to businesses offering such services, as cyber threats are a risk to any user or device connected to the Internet. With the development of ICT and IoT technologies, hyperconnectivity is ubiquitous and present in situations we previously could not even imagine. However, 5G security is not just a technical issue, but also involves what kind of business might exist when 5G hits “full force”. The business model approach can help companies to identify potential avenues of business opportunities and know how to actually monetize from 5G.

5.6 Acknowledgement

This study has been supported by the DIMECC Cyber Trust program.

References

- 1 Dobrian, J. (2015) Are you sitting on a cyber security bombshell? *Journal of Property Management*, 80, 8–12.
- 2 Gomes, J.F., Ahokangas, P. and Moqaddamerad, S. (2016) Business modeling options for distributed network functions virtualization: Operator perspective. In: *Conference Proceedings from European Wireless 2016*, pp. 37–42.
- 3 The 5G Infrastructure Public Private Partnership (2016) 5G PPP architecture working group: view on 5G architecture. 5GPPP, Brussels.
- 4 The 5G Infrastructure Public Private Partnership (2016) 5G empowering vertical industries: Roadmap paper. 5GPPP, Brussels.
- 5 Zott, C., Amit, R. and Massa, A.L. (2011) The business model: recent developments and future research. *Long Range Planning*, 37, 1019–1042.
- 6 Afuah, A. (2004) *Business Models: A Strategic Management Approach*. New York: McGraw-Hill/Irwin.
- 7 Alt, R. and Zimmermann, H. (2001) Introduction to special section – business models. *Electronic Markets*, 11, 3–9.
- 8 Chesbrough, H. and Rosenbloom, R.S. (2002) The role of the business model in capturing value from innovation: evidence from Xerox Corporation’s technology spin-off companies. *Industrial and Corporate Change*, 11, 529–555.

- 9 Richardson, J. (2008) The business model: an integrative framework for strategy execution. *Strategic Change*, 17, 133–144.
- 10 Shafer, S.M., Smith, H.J. and Linder, J.C. (2005) The power of business models. *Business Horizons*, 48, 199–207.
- 11 Chesbrough, H. (2010) Business model innovation: opportunities and barriers. *Long Range Planning*, 43(2), 354–363.
- 12 Timmers, P. (1998) Business models for electronic markets. *Electronic Markets*, 8, 3–8.
- 13 Zott, C. and Amit, R. (2010) Business model design: an activity system perspective. *Long Range Planning*, 43, 216–226.
- 14 Morris, M., Schindehutte, M. and Allen, J. (2005) The entrepreneur's business model: toward a unified perspective, *Journal of Business Research*, 58, 726–735.
- 15 Amit, R. and Zott, C. (2001) Value creation in e-business. *Strategic Management Journal*, 22, 493–520.
- 16 Weill, P. and Vitale, M. (2002) What IT infrastructure capabilities are needed to implement e-business models? *MIS Quarterly*, 1, 17–34.
- 17 Lehto, M. (2015) Phenomena in the cyber world. In: *Cyber Security: Analytics, Technology and Automation*. Springer International Publishing, USA, pp. 3–29.
- 18 Bodeau, D.J., Graubart, R. and Fabius-Greene, J. (2010) Improving cyber security and mission assurance via cyber preparedness (cyber prep) levels. *2010 IEEE Second International Conference on Social Computing*, pp. 1147–1152.
- 19 Vatis, M. (2002) Cyber attacks: Protecting America's security against digital threats. *Discussion Panel*, USA (online). Available at: URL=<http://nsarchive2.gwu.edu//NSAEBB/NSAEBB424/docs/Cyber-015.pdf>
- 20 McCusker, R. (2006) Transnational organised cyber crime: distinguishing threat from reality. *Crime, Law and Social Change*, 46, 257–273.
- 21 Liaropoulos, A. (2012) War and ethics in cyberspace: cyber-conflict and just war theory. In: *Leading Issues in Information Warfare and Security Research*, 2nd edition, Good News Digital Books, USA, pp. 121–134 (online). Available at: https://play.google.com/store/books/details/Julie_Ryan_Leading_Issues_in_Cyber_Warfare_and_Sec?id=JZFmCwAAQBAJ
- 22 Beggs, C. (2006) Proposed risk minimization measures for cyber-terrorism and SCADA networks in Australia. In: *Proceedings of the 5th European Conference on Information Warfare and Security*, pp. 9–18.
- 23 NATO (2013) *The History of Cyber Attacks – a timeline*. Available at: <http://www.nato.int/docu/review/2013/cyber/timeline/EN/index.htm>
- 24 Herzog, S. (2011) Revisiting the Estonian cyber attacks: digital threats and multinational responses. *Journal of Strategic Security*, 4, 49–60.
- 25 Ulsch, N.M. (2014) *Cyber Threat! How to Manage the Growing Risk of Cyber Attacks*. John Wiley & Sons, Hoboken, NJ.
- 26 US Chamber of Commerce (2015) The case for enhanced protection of trade secrets in the trans-Pacific partnership agreement. US Chamber.
- 27 Ponemon Institute (2015) 2015 cost of data breach study: Global analysis. Ponemon Institute LLC.
- 28 Ahokangas, P., Moqaddamerad, S., Matinmikko, M., Abouzeid, A., Atkova, I. et al. (2016) Future micro operators business models in 5G. *The Business & Management Review*, 7(5), 143–149
- 29 Onetti, A., Zucchella, A., Jones, M.V. and McDougall-Covin, P.P. (2012) Internationalization, innovation and entrepreneurship: business models for new technology-based firms. *Journal of Management & Governance*, pp. 337–368.

- 30 Teece, D.J. (2010) Business models, business strategy and innovation. *Long Range Planning*, 43(2), 172–194.
- 31 Ballon, P. (2007) Business modelling revisited: the configuration of control and value. *Info*, 9, 6–19.
- 32 Ahokangas, P. *Vertical, Horizontal and Oblique Business Models (Blog)*. Available at: <http://techbusstratintfutures.blogspot.fi/>
- 33 Wirtz, B.W., Schilke, O. and Ullrich, S. (2010) Strategic development of business models: implications of the Web 2.0 for creating value on the Internet. *Long Range Planning*, 43(2), 272–290.
- 34 Yrjölä, S., Ahokangas, P. and Matinmikko, M. (2015) Evaluation of recent spectrum sharing concepts from business model scalability point of view. *2015 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, pp. 241–250.
- 35 Mitola, J., Guerci, J., Reed, L.J., Yao, Y.D., Y. Chen, Y. et al. (2014) Accelerating 5G QoE via public-private spectrum sharing. *IEEE Communications Magazine*, 52, pp. 77–85.
- 36 Pinto, Z. (2013) *Cyber Security – product of service?* Available at: <http://www.automationworld.com/security/cyber-security-product-or-service>
- 37 Scully, T. (2014) The cyber security threat stops in the boardroom. *Journal of Business Continuity & Emergency Planning*, 7, 138–148.
- 38 Ernst & Young (2016) *Cyber Preparedness: the Next Step for Boards*. Available at: <http://www.ey.com/gl/en/issues/governance-and-reporting/ey-cyber-preparedness-the-next-step-for-boards>
- 39 Gomes, J.F., Ahokangas, P. and Owusu, K.A. (2016) Business modeling facilitated cyber preparedness. *The Business Management Review*, 7(4), 1–12 (online). Available at: URL=<https://search.proquest.com/docview/1799575982?pq-origsite=gscholar>
- 40 Campbell, K., Diffley, J., Flanagan, B., Morelli, B., O'Neil, B. and F. Sideco, F. (2017) The 5G economy: How 5G technology will contribute to the global economy. In: *Economic Impact analysis: IHS economics & IHS Technology*, pp. 1–34.
- 41 Statista (2016). *Internet of Things (IoT): Number of Connected Devices Worldwide from 2012 to 2020 (in billions)*. Available at: <https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/>

Part II

5G Network Security

6

Physical Layer Security

Simone Soderi^{1,2}, Lorenzo Mucchi³, Matti Hämäläinen¹, Alessandro Piva³, and Jari Linatti¹

¹ Centre for Wireless Communications, University of Oulu, Oulu, Finland

² Alstom Ferroviaria SpA, Florence, Italy

³ Department of Information Engineering, University of Florence, Florence, Italy

6.1 Introduction

Along with the rapid deployment of wireless communication networks, wireless security has become a critical concern. Unfortunately, security risks are inherent in any wireless technology. Some of these risks are similar to those of wired networks, some are exacerbated by wireless connectivity and some other are completely new. The most significant source of risks in wireless networks is that the communications medium is open to intruders. Mobile and handheld wireless devices are resource constrained (e.g. battery life) and hence such devices have limited transmission power and may use weaker cryptographic mechanisms for saving power, thereby making them easy targets for powerful adversaries. Self-configuring heterogeneous networks may use very different levels of security, the lower secured links being a potential breach for the whole system. A direct consequence of these risks is the loss of data confidentiality and integrity and the threat of denial of service (DoS) attacks to wireless communications. Unauthorized users may gain access to system and information, corrupt the data, consume network bandwidth, degrade network performance, launch attacks that prevent authorized users from accessing the network, or use resources to launch attacks on other networks.

The security of radio interfaces of wireless networks is nowadays crucial for many applications: broadband internet, e-commerce, radio-terminal payments, bank services, machine-to-machine communications, health/hospital remote services, etc. In addition, the approaching sensing procedures of future radio access technologies such as white space and cognitive networks will result in numerous transmissions of precious geo-referenced radio engineering data, whose integrity and confidentiality must be well secured.

Most commonly-used security methods rely on cryptographic techniques, either asymmetric (based on public and private keys) or symmetric (based on a shared secret

not known by others), which are located at the upper layers of a wireless network. Encryption does not protect from the undesired demodulation of the information by eavesdroppers, but only from the interpretation of the data as meaningful words. The cryptographic protocols base their security on the fact that, statistically, the amount of time for performing a decrypting analysis is enormous. The time to break a codeword is related to the computational power of the attacker, that is, cryptography intrinsically assumes that the eavesdropper has a limited amount of computational capability. Recent efforts of academia and industries to power up the amount of operations per second of the digital processors make this assumption weaker and weaker. Physical layer security does not make any assumption on the computational power of the attackers. Moreover, the standard practice of adding authentication and encryption to the existing protocols at the various communication layers has led to inefficient aggregations/mixtures of security mechanisms. Since data security is so critically important, it is reasonable to argue that security measures should be implemented at all layers where this can be done in a cost-effective manner. This leads us to point out our attention to the first layer: the physical one.

6.1.1 Physical Layer Security in 5G Networks

Working on the next generation of wireless communications imposed the development of the *security engineering* as a multidisciplinary field. Nowadays, skills required for security range from cryptography and computer science to hardware and embedded systems [15]. Typically, security is implemented through cryptography at upper layers in the open system interconnection (OSI) model. In the past few years, several techniques based on signal processing have been utilized to secure communications at a physical layer and those are promising methods in the applications where standalone security solution is needed [21,23].

In this chapter, the authors focus their attention to the fifth generation (5G) networks' security. 5G is expected to enable the hyper-connected society, supporting the growth and the development in many sectors. These improvements leverage the deployment of new products also in critical infrastructure, such as railway and energy, in which a high degree of availability and dependability is required. On the other hand, this trend leads to an additional exposure of the future mobile communications to cyber-attacks that can undermine the system availability [28]. Security services included in wireless communications are authentication, confidentiality, integrity and availability [15]. The new security idea discussed here addresses countermeasures against the confidentiality attacks.

5G network shall serve as key enabler for future applications that use wireless technologies. Wireless communications beyond 2020 becomes pervasive, introducing the *Internet of everything*, in which many small devices interact with each other and with users. 5G systems will encompass different radio providing ultra-high capacity, energy efficiency and new spectrum management solutions [24–26]. This evolution imposes the development of new security engineering methodologies and the appropriate mitigations, because these systems will have different wireless interfaces and their operating scenarios will include an extensive utilization of wireless links. Due to its nature, wireless communications might be vulnerable to eavesdropping attacks and this chapter propose the utilization of physical layer security techniques to 5G network.

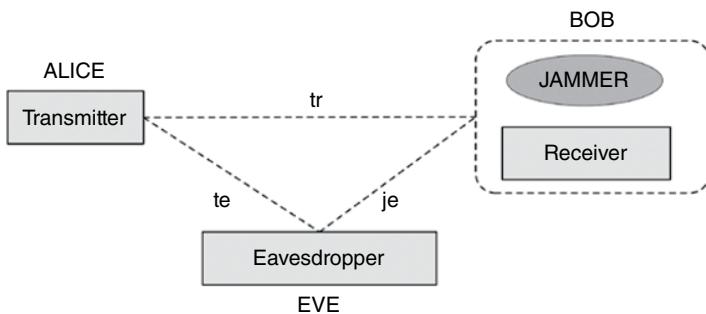


Figure 6.1 Block diagrams of the proposed protocol to analyze physical layer security.

6.1.2 Related Work

In the literature, there are several contributions that deals with a *physical layer security* because, due to their nature, wireless communications might suffer eavesdropping attacks. In 1949, Shannon defined the information theoretic metric for secrecy systems [1] and he proved the perfect secrecy condition where the eavesdropper cannot pull out any information from the transmitted signal. Afterwards, Wyner introduced a wiretap channel model defining a *secrecy capacity* as the maximum transmission rate that is achievable whenever the eavesdropper's channel observations are noiser than the legitimate user's channel [2,17]. Finally, Csiszár *et al.* extended Wyner's results to non-zero secrecy capacity when a non-degraded wiretap channel is utilized [4]. This model includes a transmitter, Alice, a legitimate receiver, Bob, and a passive eavesdropper named Eve. Bob and Eve receive Alice's transmissions through independent channels, as depicted in Figure 6.1, where *tr* indicates transmitter-receiver link, *te* is the transmitter-eavesdropper link, and *je* is the jammer-eavesdropper link. As shown in Figure 6.1, we expanded this model, introducing a receiver with a jammer whose utilization is explained in the rest of this chapter.

In the past few years, researchers exploited jamming as a fundamental part of original ideas for network security. More recently, a channel independent protocol named iJAM has been introduced [20]. The fundamental *iJAM operating principle* is shown in Figure 6.2. Alice, the sender, transmits two times each symbol and Bob, the receiver, randomly jams complementary samples over the two symbols. In this scheme, only the legitimate receiver knows which samples it jammed. Later, Bob is able to get a clean signal by discarding corrupted complementary samples from the original signal and its repetition. In contrast, the eavesdropper cannot remove the interference, because he does not have any information on the jamming characteristics [20].

6.1.3 Motivation

The primary goal of this study is to develop a new transceiver architecture to ensure secure communication combining *watermarking* with *jamming receiver*. As a performance metrics, authors utilize *an outage probability of the secrecy capacity* to evaluate the effectiveness of this secure communication. The proposed scheme is partially based on iJAM's concept and the paper also provides the information theory analysis for the evaluation of this new approach.

Soderi *et al.* proposed the watermark-based blind physical layer security (WBPLSec) as a valuable method to secure communication without neither assumptions on

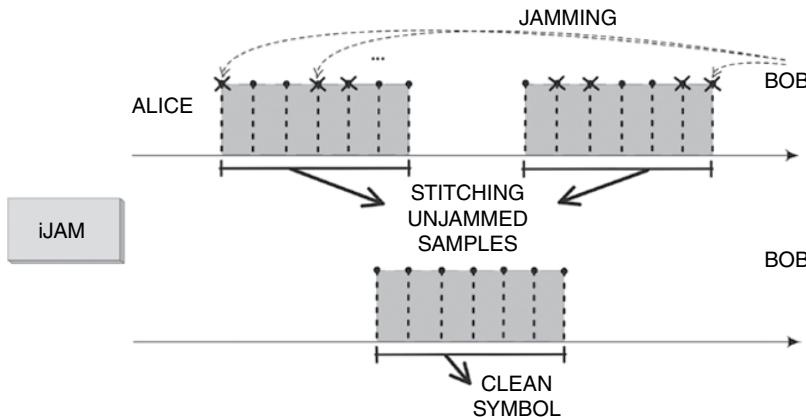


Figure 6.2 iJAM's operating principle.

eavesdropper's channel nor jamming from third-party nodes [27]. In that paper, authors exploit the watermarking concept to increase system performance in terms of outage probability of secrecy capacity.

In the multimedia context, the *digital watermarking* process is utilized to hide or embed a desired signal into another signal, for example into pictures and videos. This process has a lot of similarities with traditional communications. Spread-spectrum (SS) watermarking techniques are frequently utilized to implement physical layer security [22] and we adopt the second paradigm for watermarking described by Cox *et al.* [7], where the information to be embedded is modified prior to insertion, to exploit hidden data.

The truly innovative process for deploying a physical layer security consists of the steps showed in Algorithm 6.1.

Algorithm 6.1 WBPLSec protocol

- 1: **procedure:** PHYSICAL LAYER SECURITY
 - 2: *SS Watermarking (ALICE):*
A message is first modulated with SS and then embedded into the host signal.
 - 3: *Jammer Receiver (BOB):*
The receiver jams N_W samples for each symbol transmitted by Alice.
 - 4: *Watermark Extraction (BOB):*
The receiver extracts the watermark.
 - 5: *Symbol Rebuild (BOB):*
Knowing which samples are jammed, the receiver, i.e. Bob, is able to rebuild a clean symbol using information contained into the watermark.
 - 6: **End procedure.**
-

Note: WBPLSec transmits the information through two independent paths implementing *data decomposition policy*. The information is sent via a narrowband signal and through the SS watermarked signal. The narrowband signal is partially jammed by Bob, but the watermark into the SS signal is utilized to re-compose the entire symbol.

The rest of this chapter is organized as follows: Section 10.2 describes the WBPLSec system model introducing transmitter and receiver architectures. Section 10.3 introduces the outage probability of secrecy capacity of a jamming receiver. Then, in Section 10.4, the application to 5G use case is presented. Finally, the chapter is concluded with Section 10.5.

6.2 WBPLSec System Model

The authors address the general problem of physical layer security presented in [13], in which any secure communications shall handle secrecy to avoid confidentiality attacks. The WBPLSec system model is shown in Figure 6.3, where the jamming receiver together with the watermarking provides secrecy. The selected watermarking technique provides the needed information destroyed with the jamming.

In our study, a modified version of the non-degraded wiretap channel model [4] is used. It includes the so-called *jamming channel* utilized to jam the received signal and the eavesdropper.

The source message $(x_S)^N$ of length N is encoded into code word $(x_{S'})^N$ of length N . In particular, the encoder embeds the watermark $(x_W)^{N_W}$ of length N_W into the host signal $(x_S)^N$. The legitimate user, Alice, transmits $(x_{S'})^N$ to Bob through the *main channel*, which in this case is assumed to be a discrete-time Rayleigh fading channel. The i -th sample of the signal received by Bob is given by

$$y_M(i) = h_M(i)x_{S'}(i) + k_J(i)x_J(i) + n_M(i), \quad (6.1)$$

where $h_M(i)$ and $k_J(i)$ represent the main channel's and the jamming channel's complex Gaussian fading coefficients, $n_M(i)$ is the complex zero-mean Gaussian noise, and $x_J(i)$ denotes the jamming signal, which is generated by Bob.

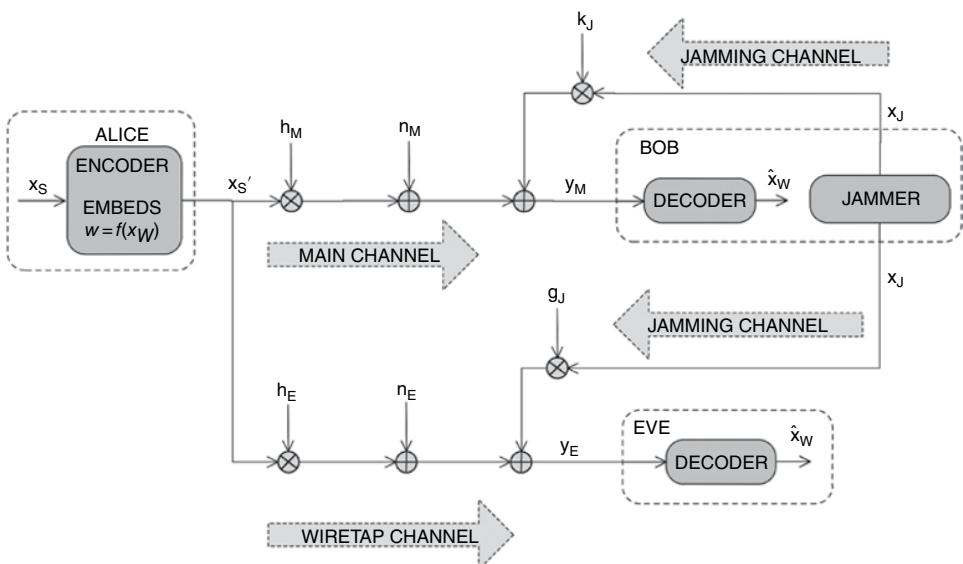


Figure 6.3 Non-degraded wiretap channel model with jamming receiver.

Figure 6.3 shows how the eavesdropper, Eve, is capable of observing Alice's transmission over an independent discrete-time Rayleigh channel, that is, a *non-degraded wiretap channel*. The i -th sample of the signal received by Eve is given by

$$y_E(i) = h_E(i)x_{S'}(i) + g_J(i)x_J(i) + n_E(i), \quad (6.2)$$

where $h_E(i)$ is the wiretap channel's complex Gaussian fading coefficient between Alice and Eve, $n_E(i)$ is the complex zero-mean Gaussian noise, and $g_J(i)$ is the jamming channel complex Gaussian fading coefficient. It is assumed that all channels are quasi-static fading channels, which mean that the channel gain coefficients remain constant during the transmission of a code word: $h_M(i) = h_M$, $h_E(i) = h_E$, $k_J(i) = k_J$ and $g_J(i) = g_J$, $\forall i = 1, \dots, N$.

6.2.1 Transmitter

In accordance with the *data decomposition method* proposed in Section 10.1, Alice conveys the information by means of two independent paths. The information is sent to the legitimate user by means of a narrowband signal. On the other hand, Alice also embeds a SS watermark in the host narrowband signal. The watermark conveys part of the information to the legitimate user, Bob, through a secondary channel.

In accordance with the framework presented by Cox *et al.* [6], the transmitter combines the original modulated signal with an SS watermark, with an embedding rule defined as

$$x_{S'}(i) = x_S(i) + \mu w(i), \quad (6.3)$$

where $x_S(i)$ is the i -th sample of the amplitude shift keying (ASK) transmitted signal, μ is the scaling parameter, and $w(i)$ is SS watermark. Without loss in generality, in the rest of the chapter we use the direct sequence spread spectrum for watermarking. On the other hand, the same mechanism developed in WBPLSec can be implemented throughout orthogonal frequency division multiplexed (OFDM) signals. Corresponding to iJAM, the utilization of OFDM ensures the jammed samples are indistinguishable from the clean samples¹.

The host amplitude shift keying (ASK) modulated signal x_S can be expressed as

$$x_S(i) = \begin{cases} A_a \sqrt{\frac{2}{T_{hs}}} \cdot \cos(2\pi f_{hs}i), & \text{for } 0 \leq i \leq T_{hs}, \\ 0, & \text{elsewhere} \end{cases} \quad (6.4)$$

where A_a is the amplitude, T_{hs} is the symbol time, and f_{hs} is the frequency of the modulated signal. We propose as proof-of-concept the utilization of DSSS signal for watermarking as

$$w(i) = \sum_{k=-\infty}^{+\infty} \sum_{j=0}^{N_c-1} g(i - kT_b - jT_c)(c_W(i))_j (x_W(t))_k, \quad (6.5)$$

¹ OFDM time samples approximate Gaussian distribution and if jamming signal has the same distribution, the overall distribution after jamming does not modify the distribution of an OFDM signal [27].

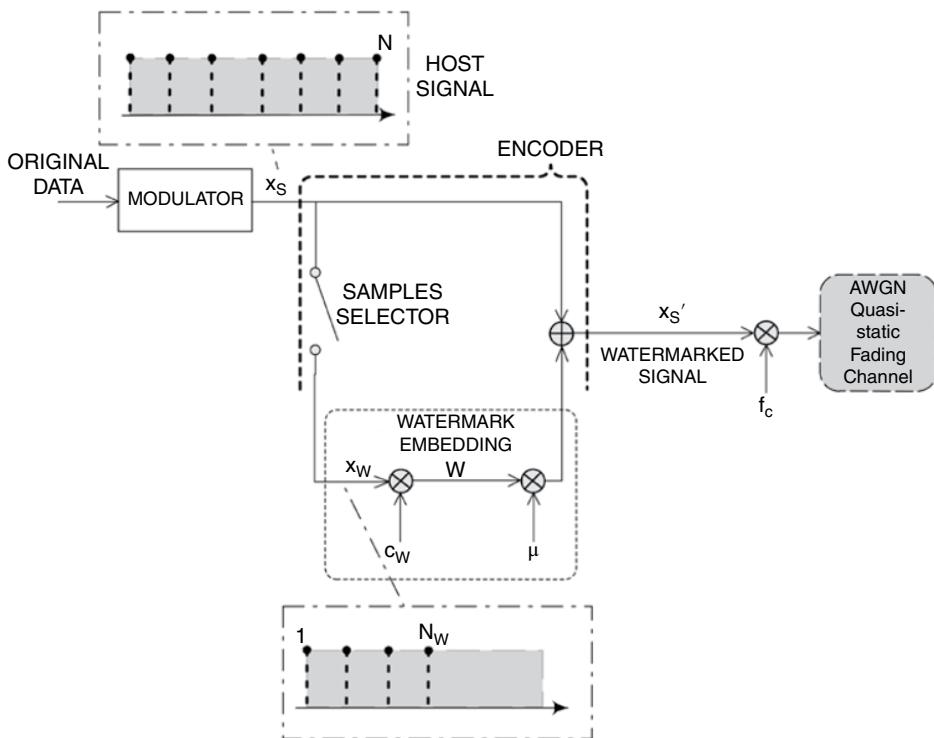


Figure 6.4 Transmitter structure for watermark-based blind physical layer security.

where $(x_W(i))_k$ is the k -th data bit of the watermark signal, $(c_W(i))_j$ represents the j -th chip of the orthogonal pseudo-noise (PN) sequence, $g(i)$ is the pulse waveform, T_c is the chip length, and $T_b = N_c T_c$ is the bit length. The SS watermarking is shown in Figure 6.4, where c_W represents PN code, which spreads the information, that is x_W , that must be inserted in the host signal. With these assumptions, the energy of the watermarked signal is given by

$$E_{S'} = \sum_{i=1}^N |x_{S'}(i)|^2 = \quad (6.6)$$

$$\begin{aligned} &= \sum_{i=1}^N |x_S(i)|^2 + \mu^2 \sum_{i=1}^N |w(i)|^2 + 2\mu \sum_{i=1}^N |x_S(i)w(i)| = \\ &= E_S + \mu^2 E_W, \end{aligned} \quad (6.7)$$

where E_S is the energy of the x_S signal and E_W is the energy of x_W . It is assumed that the host signal and its watermark in Equations (6.4) and (6.5) are uncorrelated.

The signal watermarking is done by utilizing the traditional spread spectrum based approach [9]. The main idea implemented in the watermark embedding phase is that

the transmitter marks, utilizing SS, the host signal x_S utilizing its first N_W over N samples. Then x_W is given by

$$x_W(i) = \begin{cases} x_S(i), & \text{for } 1 \leq i \leq N_W, \\ 0, & \text{elsewhere.} \end{cases} \quad (6.8)$$

Alternatively, the receiver can jam N_W discontinuous samples for each symbol, but even if this randomness requires a wide-band jammer, for example ultra wideband (UWB), the work presented in this chapter is still valid. With $N_W < N$, the energy of the watermark is given by

$$E_W = \frac{N_W}{N} E_S. \quad (6.9)$$

Finally, the signal is mixed to carrier frequency f_c and radiated by the antenna. Figure 6.4 shows the block diagram of the transmitter.

6.2.2 Jamming Receiver

In this chapter, the authors propose a different strategy to implement the jamming receiver's architecture if compared to iJAM [20]. Indeed, the proposed scheme of the receiver works with jammed samples as well as the watermark extraction.

It is assumed that both the jamming signal and the host signal have the same energy over N samples as

$$E_S = \sum_{i=1}^N |x_S(i)|^2 = \sum_{i=1}^N |x_J(i)|^2. \quad (6.10)$$

Assuming N samples for symbol, as Bob jams M samples over N with $M < N$, the energy of the jamming signal is given by

$$E_J = \frac{M}{N} E_S. \quad (6.11)$$

The receiver structure is shown in Figure 6.5. In the WBPLSec, the legitimate receiver can jam at most $M = N_W$ samples, because N_W samples are the information transmitted through the SS watermark. The received signal after the antenna is down-converted to the baseband by the carrier frequency f_c and then processed by the original signal demodulator to recover data exchanged through channel. Due to the jamming, the signal after the low pass filter (LPF), that is \hat{x}_S , is corrupted and unusable alone. To stitch unjammed samples and create a clean symbol, in parallel, the received signal is led to an additional DSSS demodulator used to recover the watermark x_W . Afterwards, as in the iJam protocol [20], the receiver replaces corrupted samples in \hat{x}_S with non-jammed samples, which in our solution are taken from \hat{x}_W . In the end, the clean symbol x_S is achieved and then demodulated.

6.2.3 Secrecy Metrics

In Section 10.1, the authors presented the standard metrics used to measure the secrecy of communications. Regarding the notation used in Figure 6.3, Shannon defined a

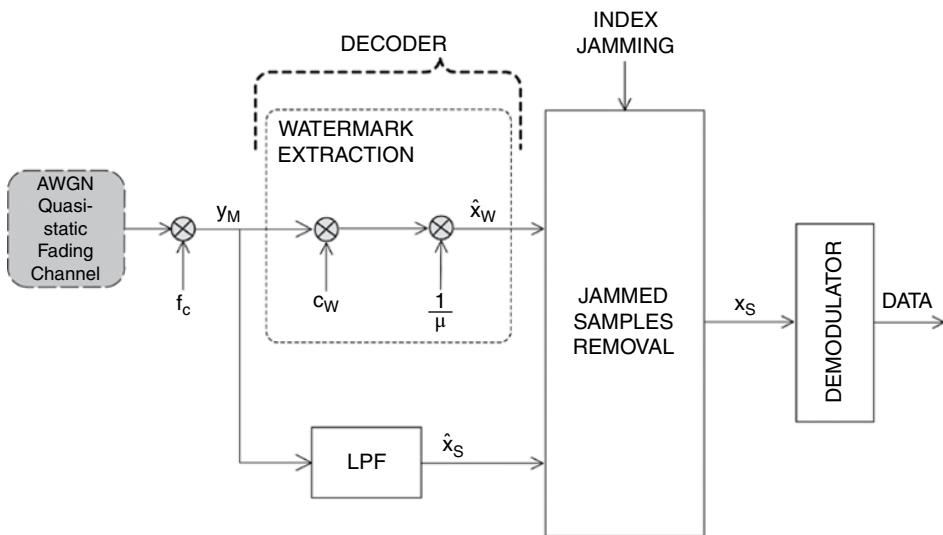


Figure 6.5 Receiver structure for watermark-based blind physical layer security.

system that operates with *perfect secrecy* if the mutual information between the message $(x_S)^N$ and the encoder output $(x_{S'})^N$ is zero [1]. This can be expressed as

$$I((x_S)^N; (x_{S'})^N) = 0. \quad (6.12)$$

Together with the introduction of the wiretap channel, Wyner suggested the utilization of the *weak secrecy*, in which the amount of the information leaked about the message $(x_S)^N$ by the eavesdropper when he observes $(y_E)^N$, is asymptotically zero [2], that is

$$\lim_{N \rightarrow \infty} \frac{1}{N} I((x_S)^N; (y_E)^N) = 0. \quad (6.13)$$

Some applications cannot accept any information leakage and Maurer *et al.* defined the *strong secrecy* as follows [8]:

$$\lim_{N \rightarrow \infty} I((x_S)^N; (y_E)^N) = 0. \quad (6.14)$$

Strong secrecy is hard to design and weak secrecy preserves a practical interest [21]. The authors recall that the secrecy capacity of the legitimate link is defined as the maximum rate that is achievable with strong secrecy [3]. The objective of the physical layer security is to implement a reliable secure communication between Alice and Bob, at a target secure rate, leaking the least possible number of bits. Moreover, when the secrecy capacity is equal to zero, Alice can decide not to transmit, thus avoiding disclosure of any information. Reasonably, the authors selected the outage probability (P_{out}) to describe the secrecy capacity in the modified wiretap channel model depicted in Figure 6.3. P_{out} is defined as the probability that the secrecy capacity is less than a target secrecy rate $R_s > 0$ [11].

6.2.4 Secrecy Capacity of WBPLSec

Win *et al.* [16] utilized a general wireless propagation model to characterize network interference in wireless systems. In accordance with that model, the received power, that is P_{rx} , is $\propto \frac{P_{tx}}{d_n^{2b}}$, where P_{tx} denotes the transmitted power, d_n , the distance between the two nodes and b is the amplitude loss exponent [10].

The power spectra densities of the signals discussed above are illustrated in Figure 6.6. As shown in Figure 6.5, the received signal by Bob is split into two arms. The first part despreads and extracts the watermark. The latter filters the received signal to limit the bandwidth before the signal recovery [5]. The ideal LPF rejects a large fraction of the SS watermark and the magnitude of the residual watermark power density is given by

$$E_{W'} = \frac{B_{hs}}{B_{ss}} E_W = \frac{E_W}{G_p} \quad (6.15)$$

where $B_{hs} = \frac{1}{T_{sa}}$ is the bandwidth of the host signal, T_{sa} is the host signal symbol length, $B_{ss} = \frac{1}{T_c}$ is the bandwidth of SS signal, and $G_p = \frac{T_{sa}}{T_c}$ is the processing gain. $E_{W'}$ interferes with the narrowband demodulator and G_p is defined as the inverse of the E_W reduction factor [5].

Therefore, the instantaneous signal-to-interference-plus-noise ratio (SINR) at the legitimate receiver, that is γ_M , is given by

$$\gamma_M = \frac{\frac{|h_M|^2 E_S|}{d_{tr}^{2b}}}{N_0 + |k_J|^2 E_J} = \frac{\alpha \gamma_{tr}'}{1 + \tilde{\alpha} \gamma_{jr}}, \quad (6.16)$$

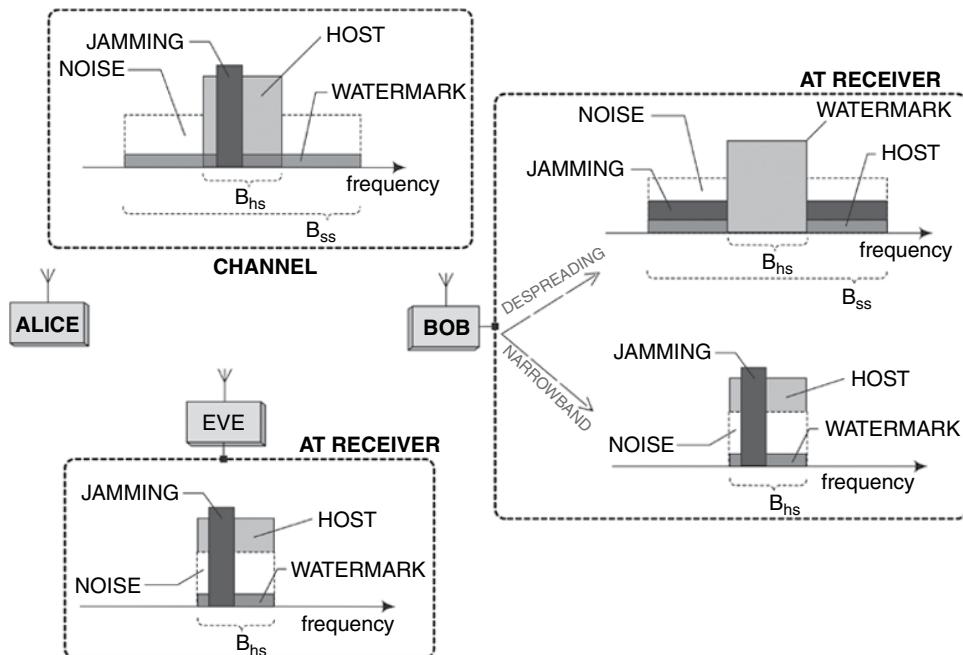


Figure 6.6 Power spectra densities of proposed blind physical layer security.

where both $\alpha = |h_M|^2$ and $\tilde{\alpha} = |k_J|^2$ follow an exponential distribution, $N_0' = N_0 + E_{W'}$, $\gamma_{tr}' = \frac{E_S}{(N_0' d_{tr}^{2b})}$ and $\gamma_{jr}' = \frac{E_J}{N_0'}$. Due to the proposed jamming receiver architecture, the E_J does not undergo any attenuation at the legitimate receiver. Channels are power limited and it is assumed that $P = E_S N$ is the average transmit power, and $P_J = E_J M$ is the average jamming power when Bob jams M samples over N with $M < N$. Moreover, it is assumed that n_M and n_E have the same noise spectral density, that is N_0 .

The instantaneous SINR at eavesdropper, that is γ_E is given by

$$\gamma_E = \frac{\frac{|h_E|^2 E_S}{d_{te}^{2b}}}{N_0' + \frac{|g_J|^2 E_J}{d_{je}^{2b}}} = \frac{\beta \gamma_{te}}{1 + \tilde{\beta} \gamma_{je}}, \quad (6.17)$$

where both $\beta = |h_E|^2$ and $\tilde{\beta} = |g_J|^2$ follow an exponential distribution, $N_0' = N_0 + E_{W'}$, $\gamma_{te} = \frac{E_S}{(N_0' d_{te}^{2b})}$ and $\gamma_{je} = \frac{E_J}{(N_0' d_{je}^{2b})}$.

When Bob has a better channel realization than Eve, that is $\gamma_M > \gamma_E$, the secrecy capacity (C_s) of legitimate link is defined as follows for a non-degraded Gaussian wiretap channel [4]:

$$C_s = \max\{C_M - C_E, 0\}, \quad \text{where} \quad (6.18)$$

$$C_M = \frac{1}{2} \log_2 (1 + \gamma_M) \quad \text{bit/transmission}$$

$$C_E = \frac{1}{2} \log_2 (1 + \gamma_E) \quad \text{bit/transmission}$$

where C_M is the channel capacity from Alice to Bob, that is the main channel, and C_E is the channel capacity from Alice to Eve, that is the wiretap channel exploited by the eavesdropper. Otherwise, if Eve has a better SINR than Bob, C_s is set to 0. In Equation (6.18), the author assumes that the noise plus the interference is still Gaussian.

In the presence of the Rayleigh channel, the secrecy capacity is conditioned to h_M , h_E , k_J , g_J , and without loss in generality in the rest of the chapter we impose $E[h_M^2] = E[h_E^2] = E[k_J^2] = E[g_J^2] = 1$ [18].

The lower bound of the C_s is defined as the secrecy rate (R_s). R_s is given by the difference of the channel capacities from Alice to Bob and from Alice to Eve [2].

6.2.5 Secrecy Capacity of iJAM

In the iJAM, each symbol is transmitted twice. The receiver with jammer randomly jams complementary samples in the original signal and its repetition. The receiver knows which are the corrupted samples and then the clean symbol is achieved by stitching together unjammed samples.

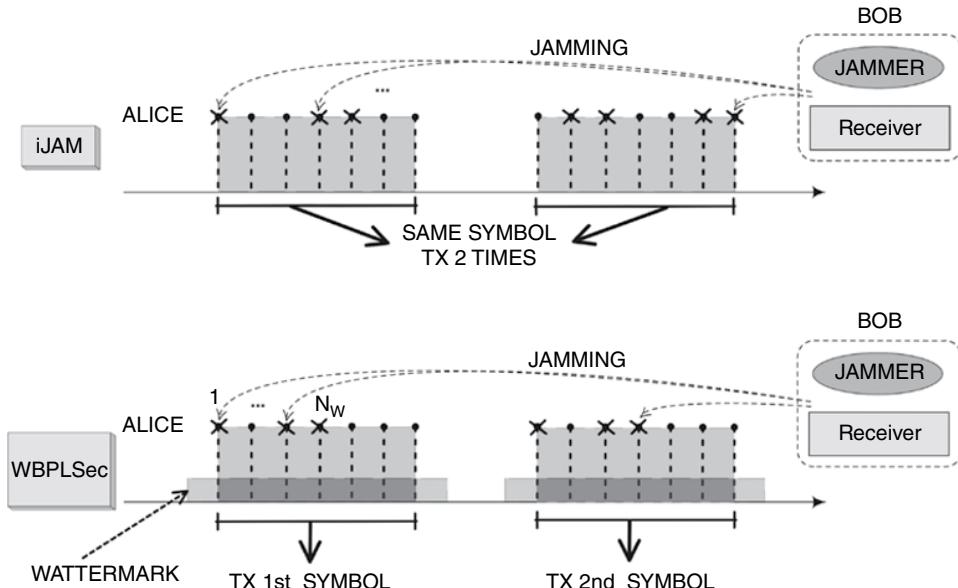


Figure 6.7 Comparison between iJAM and WBPLSec.

The SINR at the legitimate receiver is given by [18]

$$\gamma_M^{iJAM} = \frac{|h_M|^2 E_S'}{d_{tr}^{2b} N_0} = \alpha \gamma_{tr}, \quad (6.19)$$

where, in order to facilitate the comparisons between the two protocols, it is assumed to transmit the same energy, that is E_S' . When iJAM is utilized, the γ_E is still given by [17].

Figure 6.7 shows how in the iJAM the sender repeats its transmission and then halves the data-rate when compared with the WBPLSec proposed here. In particular, iJAM has to transmit twice the same symbol to obtain a clean signal, whereas WBPLSec does not. The authors compared the metrics of iJAM and WBPLSec, assuming the same energy per symbol.

In the scenario of the iJAM and assuming that iJAM and WBPLSec have the same bandwidth, the C_s is given by

$$C_s^{iJAM} = \max\{C_M - C_E, 0\}, \quad \text{where} \quad (6.20)$$

$$C_M = \frac{1}{4} \log_2 (1 + \gamma_M^{iJAM}) \quad \text{bit/transmission}$$

$$C_E = \frac{1}{4} \log_2 (1 + \gamma_E) \quad \text{bit/transmission}$$

As in Equation (6.18), the C_s^{iJAM} is conditioned to the Rayleigh channel's coefficients, that is h_M , h_E , g_J , and without loss in generality in the rest of the chapter, we impose

$E[h_M^2] = E[h_E^2] = E[g_J^2] = 1$ [18]. In Equation (6.20), the author assumes that the noise plus the interference is still Gaussian.

6.3 Outage Probability of Secrecy Capacity of a Jamming Receiver

The outage probability of the secrecy capacity is defined by Bloch *et al.* [13] as

$$\begin{aligned} P_{out} &= P[C_s < R_s] = \\ &= P\left[\frac{1}{2}\log_2\left(\frac{1+\gamma_M}{1+\gamma_E}\right) < R_s\right] = \\ &= P\left[\alpha < p(1+\tilde{\alpha}\gamma_{jr}) + q\beta\left(\frac{1+\tilde{\alpha}\gamma_{jr}}{1+\tilde{\beta}\gamma_{je}}\right)\right] \end{aligned} \quad (6.21)$$

where R_s is the target secrecy rate, $p = \frac{(2^{4R_s}-1)}{\gamma_{tr}}$ and $q = \frac{(2^{4R_s}\gamma_{te})}{\gamma_{tr}}$. Therefore, in the case of WBPLSec, the results follow from simple algebra and can be expressed as [19]

$$\begin{aligned} P_{out} &= 1 - \int_0^{\infty} \int \int e^{-p(1+\tilde{\alpha}\gamma_{jr}) - q\beta\left(\frac{1+\tilde{\alpha}\gamma_{jr}}{1+\tilde{\beta}\gamma_{je}}\right)} \cdot e^{-\tilde{\alpha}} e^{-\beta} d\tilde{\alpha} d\beta d\tilde{\beta} = \\ &= 1 - \frac{1}{(\gamma_{je}\gamma_{jr}p + \gamma_{je} - \gamma_{jr}q)^2} \cdot \\ &\quad e^{-p} (-q\Omega\left(\frac{q+1}{\gamma_{je}}\right) \left(\gamma_{je}(\gamma_{jr}p + \gamma_{jr} + 1) - \gamma_{jr}q\right) - \\ &\quad \Omega\left(\frac{(q+1)(\gamma_{jr}p+1)}{\gamma_{jr}q}\right) \left(\gamma_{je}\gamma_{jr}p - (\gamma_{je} + 1)\gamma_{jr}q + \gamma_{je}\right) \\ &\quad + \gamma_{je}(\gamma_{je}\gamma_{jr}p + \gamma_{je} - \gamma_{jr}q)), \end{aligned} \quad (6.22)$$

where $\Omega(x) = e^x E_1(x)$, $E_1 = \int_0^{\infty} (e^{-t}) dt$ is the exponential integral. It is assumed that the fading channels' coefficients are zero-mean complex Gaussian random variables (RVs). The proof that α , $\tilde{\alpha}$, β and $\tilde{\beta}$ are exponentially distributed is given in Soderi *et al.* [28].

Figure 6.8 shows the outage probability of the C_s versus γ_M for different Eve's positions. The eavesdropper moves along the line that connects Alice with Bob. The selected wireless propagation model accounts for far-field propagation [16]. We considered the near-field region limit at 1 m around Alice and Bob [18], as shown in Figure 6.8. With this model, Eve cannot be closer than 1 m to both Alice and Bob.

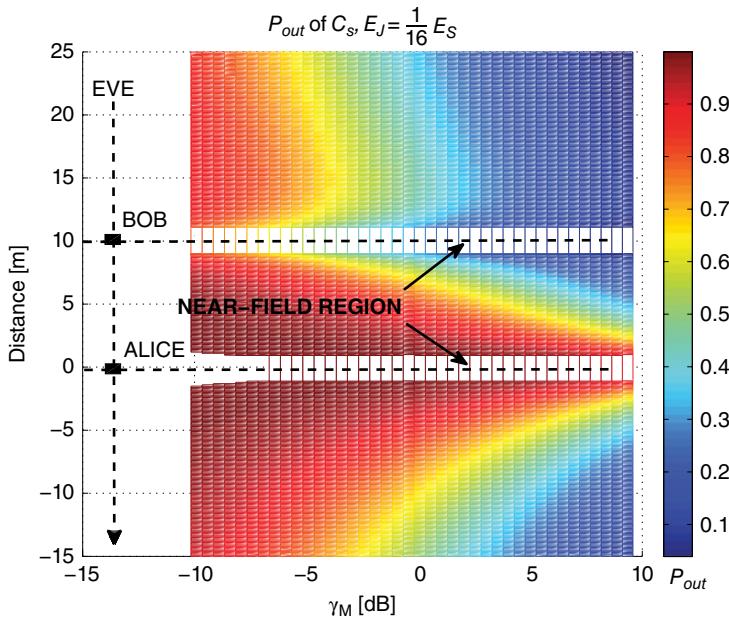


Figure 6.8 Outage probability versus γ_M when Eve moves from Bob to Alice.

In order to compare the proposed protocol against the iJAM, we computed the P_{out}^{iJAM} as

$$\begin{aligned}
 P_{out}^{iJAM} &= 1 - \iint_0^{\left(-\nu - k \frac{\beta}{1 + \beta \gamma_{je}} \right)} e^{-\beta} e^{-\tilde{\beta}} d\tilde{\beta} d\beta = \\
 &= 1 - \frac{e^{-\nu} \left(\gamma_{je} - k \Omega \left(\frac{k+1}{\gamma_{je}} \right) \right)}{\gamma_{je}^2 \gamma_{te}}. \tag{6.23}
 \end{aligned}$$

Figure 6.9 shows the comparison between the WBPLSec and the iJAM with equal energy per symbol, that is E_S . Observe that the proposed protocol has better P_{out} than iJAM. On average, WBPLSec has P_{out} two times better than iJAM, comparing curves in Figure 6.9 with same E_J . Moreover, the higher is the E_J , the lower is P_{out} that yields to increase the performance of the proposed protocol. The scenario depicted in Figure 6.9 assumes Eve in the middle between Alice and Bob.

Due to the jamming strategy implemented in the WBPLSec, Figure 6.10 shows the effect over P_{out} varying the number of jammed samples. Once more, the figure also depicts the P_{out} for the same scenario achieved with iJAM, that is when $M = N_W = 1024$ samples are jammed, that yields to have $E_J = E_W = E_S/4$. As illustrated in Figure 6.10, the more jammed samples per symbol exist, that is higher E_J , the less is the P_{out} . Thus, by controlling the value of E_J , the receiver can control the target secrecy level.

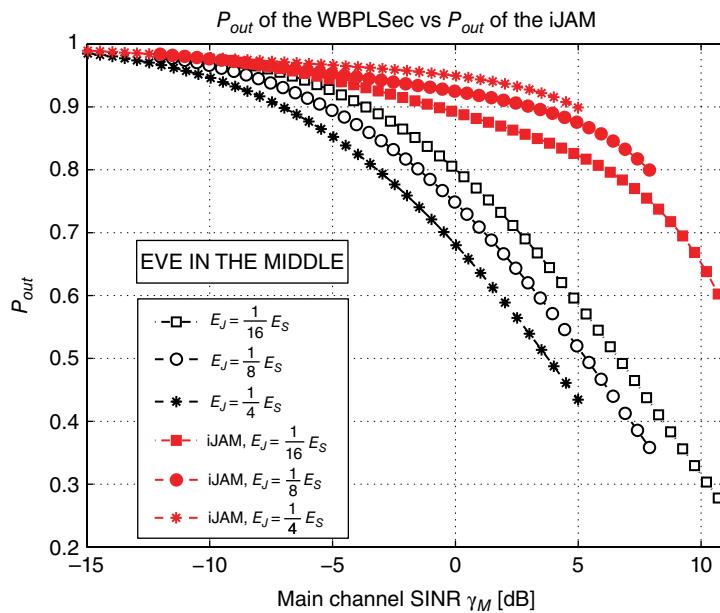


Figure 6.9 Protocol's comparison of P_{out} versus γ_M for a selected Eve's position.

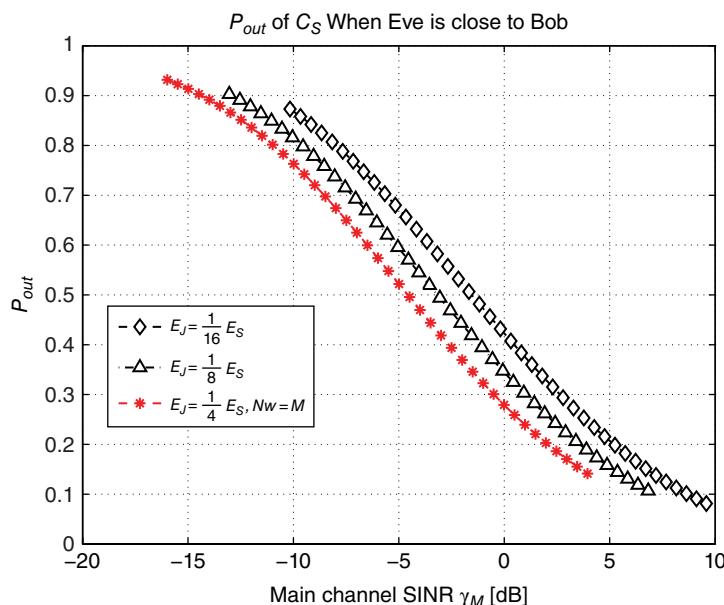


Figure 6.10 Outage probability of the C_S versus γ_M varying the jamming energy when Eve is close to Bob.

6.3.1 Simulation Scenario for Secrecy Capacity

Table 6.1 lists the parameters used for simulations. For each distance of the eavesdropper among the transmitter and the jamming receiver, the C_s was simulated with a different number of jammed samples per symbol. The outage probability of the C_s was calculated, transmitting a watermarked signal with 50 dBJ energy. The watermark varies energy from 20 to 40 dBJ and a scaling parameter until 0.9. All the scenarios simulated refer to free space.

In Figure 6.11, a comparison among three different eavesdropper's positions are shown, that is, 1) Eve is close to Alice; 2) Eve is close to Bob; 3) Eve is in the middle. As illustrated in the figure, the more jammed samples there are per symbol, the less is the effect of the eavesdropper position.

The WBPLSec creates a *security area* around Bob. As shown in Figure 6.12, if Alice and Bob should implement a secure communication with a secrecy outage probability $P_{out} = 0.3$ and $\gamma_M = 10.6$ dB, then Eve should not be close to Alice, that is the unsecured region is 5 m radius around Alice. Legitimate nodes, that is Alice and Bob, might tune E_S and E_J implementing dedicated communication protocol strategies, for example a three-way handshake, and then derive curves of P_{out} useful to define the needed security area. Furthermore, Figure 6.12 shows that with a lower γ_M , the security area is getting worse, because Eve should move away from Alice to achieve the same P_{out} . In Figure 6.12, P_{out} is plotted for two different values of γ_M , and for $N_0 = 3$ dB. It can be seen that the

Table 6.1 C_s scenario parameters.

| Parameter | Value |
|--------------------------------------------------|-------------------------|
| d_{tr} [m] | 10 |
| d_{je} [m] | -15 ÷ 25 ^[3] |
| d_{te} [m] ^[1] | 25 ÷ -15 ^[3] |
| Number of samples ($_N$) per symbol | 4096 |
| Number of jammed samples ($_M$) per symbol | 256, 512, 1024 |
| Number of samples (N_W) per watermark symbol | 1024 |
| $E_{S'}$ [dBJ] | 45 |
| E_W [dBJ] | 20 v 40 |
| Watermarking scaling parameter (μ) | 0.7, 0.9 |
| DSSS Processing Gain (G_p) | 16, 64 |
| AWGN spectral density (N_0) [dBJ] | 3, 9 |
| Amplitude path loss exponent (b) | 1.0 ^[2] |
| Secrecy Rate (R_s) | 0.1 |

[1] $d_{te} = d_{tr} - d_{je}$

[2] $b = 1$ for free-space

[3] Placing Alice at the origin of right-handed coordinate systems and Bob at the distance positive axis, when Eve moves, also negative values occur.

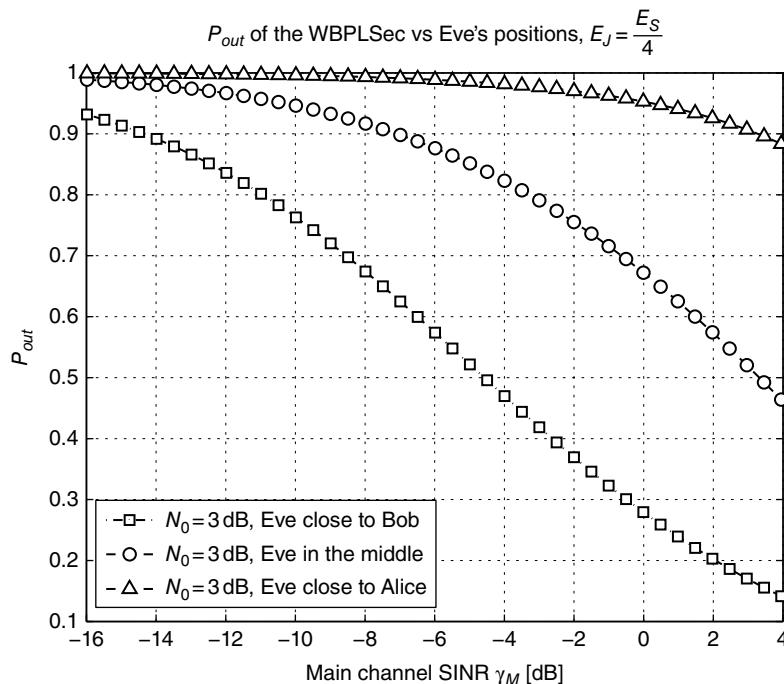


Figure 6.11 Outage probability versus γ_M for different Eve's positions.

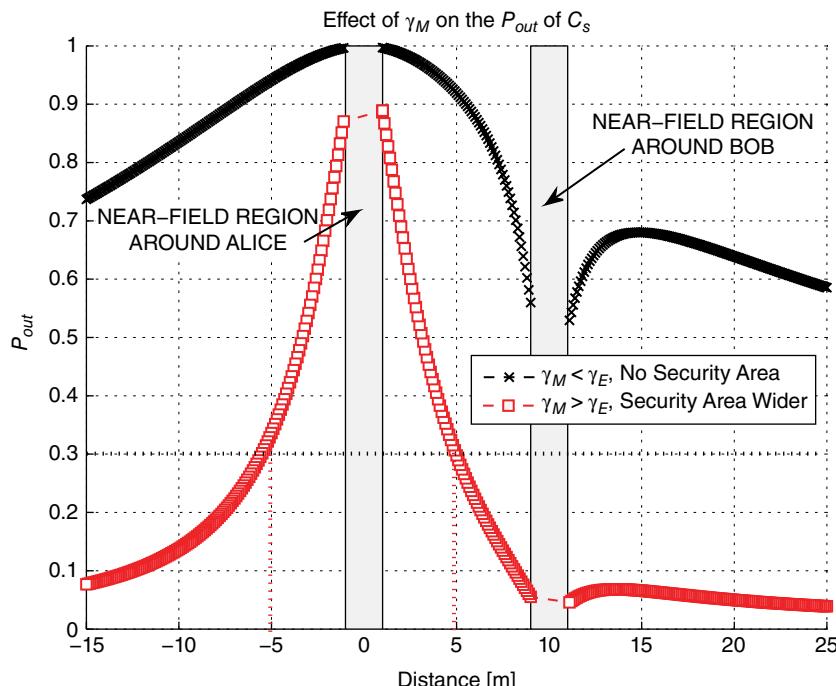


Figure 6.12 Increment of the $M\gamma$ and its effect on the secure region size.

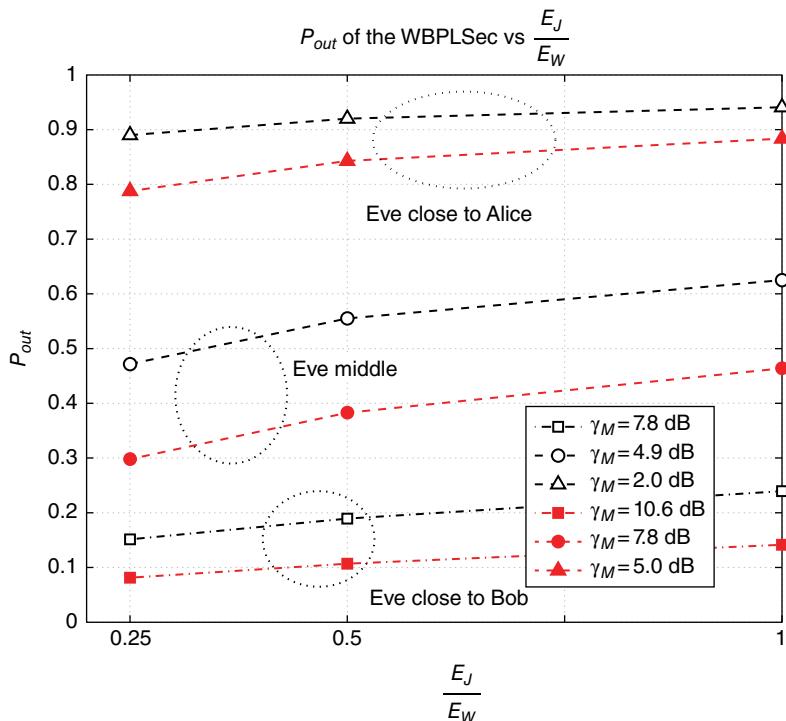


Figure 6.13 Outage probability as function of E_J/E_W .

effect of increasing the jammed samples leads to a lower P_{out} close to Alice. Figure 6.12 also shows the near-field regions around Alice and Bob.

We have already shown that the secrecy outage probability depends on the eavesdropper position and on the number of jammed samples. In Figure 6.13, we have plotted P_{out} as function of the ratio E_J/E_W for three different positions of Eve. Reasoning about the increase of E_J up to $E_J = E_W$, the P_{out} is getting worse.

6.4 WBPLSec Applied to 5G networks

Today, there are two standard practices to secure communications. The first approach adds authentication and encryption to the existing protocols. The second, which is also the approach selected by the authors of this chapter, embeds security technologies at the physical layer. In the context of 5G networks, physical layer security provides advantages when compared with cryptography techniques. The first advantage is that this technique does not depend on the computational complexity. In other words, the security level of the WBPLSec will not be affected, even if a user with high computation capacity would eavesdrop the secure communication. The second deals with the scalability, because any secure communication based on cryptography needs the cryptographic keys distribution and key management. In the case where many devices join and leave the network, that process is very challenging. Instead, the

Figure 6.14 WBPLSec applied to 5G network.

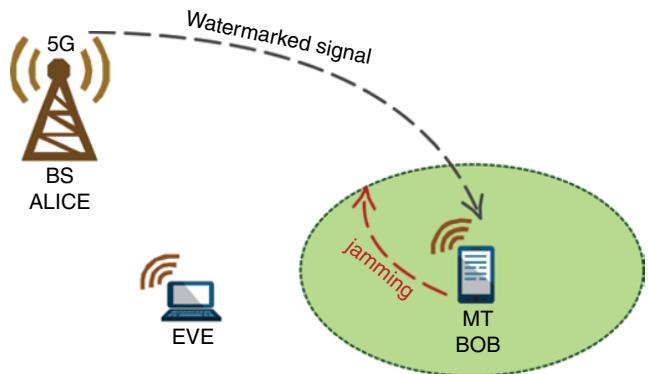
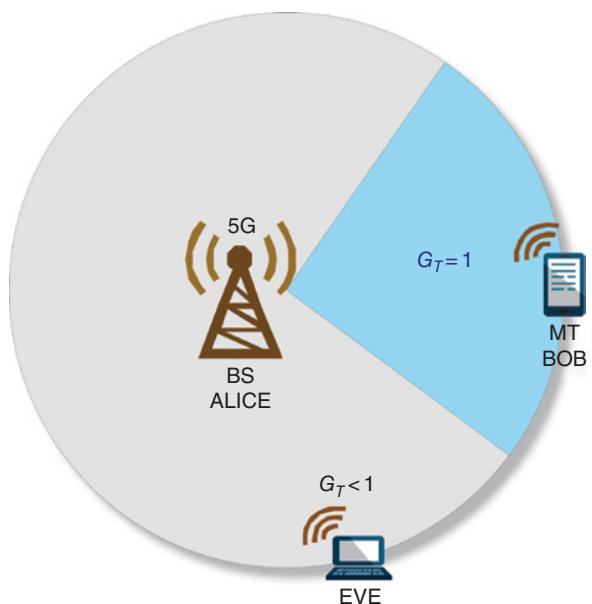


Figure 6.15 Enhancement of the WBPLSec security region in the context of cellular network.



physical layer security can be used either to provide direct secure data communication or to facilitate the distribution of cryptographic keys in the 5G networks [29].

Figure 6.14 depicts how the WBPLSec can be applied to 5G networks. In accordance to the schema already presented in this chapter, the base station (BS), marked as Alice, sends the watermarked signal to the mobile terminal (MT), marked as Bob. We assumed that Bob has multiple radio interfaces to both receive and jam the received signal. The WBPLSec technique creates a secure region around Bob, in which any other MT, such as Eve, cannot eavesdrop the information.

In the previous sections, we assumed isotropic antennas for all the stakeholders involved in the communication. On the other hand, future cellular network will exploit the smart antenna technology to improve the SINR to the legitimate receiver, Bob. Exploiting the smart antennas installed in the BS, Alice can enhance the directivity and the Bob's SINR, as shown in Figure 6.15. This produces an advantage to Bob regarding to Eve and the secure region around will be wider.

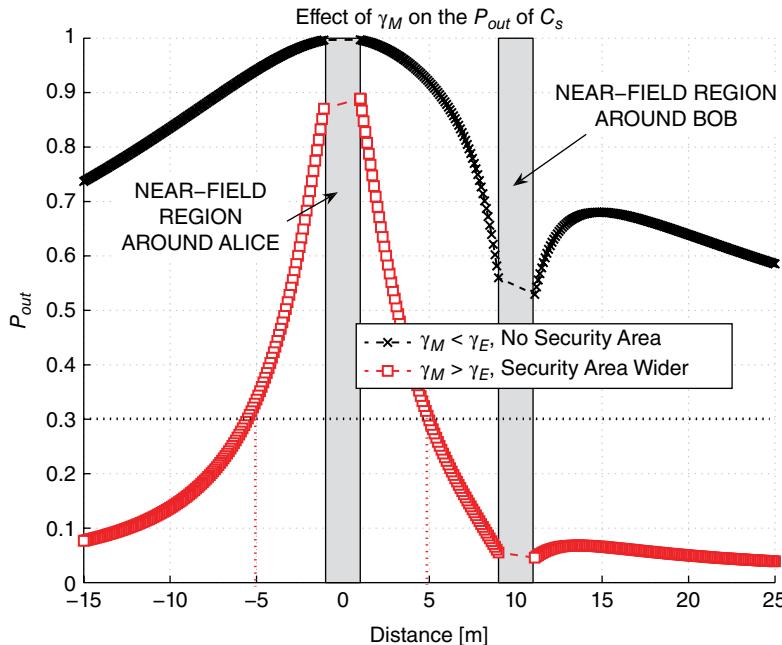


Figure 6.16 Increment of the γ_M and its effect on the secure region size.

Any mechanism used to increment the γ_M to the legitimate 5G user, yields to the positive effect to increase the secure region, as shown in Figure 6.16. On the other hand, when $\gamma_M < \gamma_E$, the security area becomes increasingly smaller up to disappearing. In such case, Alice and Bob cannot implement any secure communication.

The WBPLSec can be successfully applied in those scenarios where mobile devices are equipped with several air interfaces. A definite upward trend in the number of air interfaces for each terminal has two defined possible approaches. At first, multi-modality uses different chip solutions to implement air interfaces diversity. On the other hand, flexible air interfaces implemented via software defined radio (SDR) enables the opportunistic use of spectrum [12,23]. SDR is expected to play an important role of 5G networks, in which those air interfaces are applied to new architectures. The authors plan to leverage security and propose to use WBPLSec with SDR to support physical-layer security solution in 5G devices.

6.5 Conclusions

The authors propose a technique, which is acting at the physical layer level, and not in higher layers, like the symmetric encryption protocol does. Basically, the goal is to improve the communication system compared to those with crypto-protocols, in the same way as any other techniques that can fall into the definition of physical-layer security. Physical layer security can be used together with and not in competition with the conventional cryptographic protocols. In the 5G of mobile networks, in which security shall be built into the system architecture, it is important to develop new

techniques that implement standalone solutions, like the one proposed here. WBPLSec is a reliable physical layer solution against information disclosure attacks, such as eavesdropping. This solution is a trade-off between security and communication reliability, because for fixed symbol energy, E_s , increasing the jamming energy, E_j , a wider security area is achieved with a lower P_{out} . On the other hand, when E_j increases, the watermark extraction is getting worse with a higher P_e [27]. Furthermore, the proposed method exploits the non-degraded wiretap channel without any assumption on Eve's position and channel. One spreading code is utilized to implement SS watermarking. The wide utilization of SS communications makes the sharing of one PN code acceptable for this implementation. The WBPLSec shares the same information in terms of spreading code when compared with a SS communication.

In comparison, with the iJAM, the proposed protocol offers the following advantages:

- it is a full-rate protocol improving the major weakness of iJAM; and
- it has P_{out} two times better than iJAM.

The iJAM is an interesting protocol, but it implements a physical layer security with a split to half the data-rate. The proposed scheme is an advanced version of iJAM. Both protocols utilize SS techniques, and even though the authors implement DSSS for WBPLSec, the same concept can be applied using OFDM, making the jammed samples indistinguishable from the clean samples. The worldwide proliferation of SS communication makes the utilization of a spreading code for physical layer security reasonable for both iJAM and WBPLSec. The utilization of SS watermarking yields WBPLSec to work full rate. Furthermore, the achieved results show how the proposed protocol can be a valuable technique for deploying security, creating a secure region around the legitimate receiver. The physical layer security is one of the most promising techniques for low-power sensor networks and 5G devices. The avoidance of upper layers' cryptography makes the physical layer security attractive as a standalone security solution that can also improve the battery life, because it saves computation when compared to encryption. Soderi *et al.* presented the lower energy consumption of the WBPLSec when compared with iJAM [27].

References

- 1 Shannon, C. (1940) Communication theory of secrecy systems. *The Bell System Technical Journal*, 28(4), 656–715.
- 2 Wyner, A. (1975) The wire-tap channel. *The Bell System Technical Journal*, 54(8), 1355–1387.
- 3 Leung-Yan-Cheong, S. and Hellman, M. (1978) The Gaussian wire-tap channel. *IEEE Transactions on Information Theory*, 24(4), 451–456.
- 4 Csiszar, I. and Korner, J. (1978) Broadcast channels with confidential messages. *IEEE Transactions on Information Theory*, 24(3), 339–348.
- 5 Peterson, R.L., Ziemer, R.E. and Borth, D.E. (1995) *Introduction to Spread Spectrum Communications*. Prentice-Hall: Englewood Cliffs, NJ.
- 6 Cox, I.J., Kilian, J., Leighton, F. and Shamoon, T. (1997) Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12), 1673–1687.
- 7 Cox, I.J., Miller, M. and McKellips, A. (1999) Watermarking as communications with side information. *Proceedings of the IEEE*, 87(7), 1127–1141.

- 8 Maurer, U. and Wolf, S. (2000) From weak to strong information-theoretic key agreement. *Proceedings of the 2000 IEEE International Symposium on Information Theory*, p. 18.
- 9 Malvar, H. and Florencio, D. (2003) Improved spread spectrum: a new modulation technique for robust watermarking. *IEEE Transactions on Signal Processing*, 51(4), 898–905.
- 10 Goldsmith, A. (2005) *Wireless Communications*. Cambridge University Press: New York.
- 11 Barros, J. and Rodrigues, M.R.D. (2006) Secrecy capacity of wireless channels. *Proceedings of the 2006 IEEE International Symposium on Information Theory*, pp. 356–360.
- 12 Fitzek, F. and Katz, M. (eds) (2007) *Cognitive Wireless Networks: Concepts, Methodologies and Visions Inspiring the Age of Enlightenment of Wireless Communications*. Springer, Dordrecht, The Netherlands.
- 13 Bloch, M., Barros, J., Rodrigues, M. and McLaughlin, S. (2008) Wireless information-theoretic security. *IEEE Transactions on Information Theory*, 54(6), 2515–2534.
- 14 Jeon, H., Kim, N., Kim, M., Lee, H. and Ha, J. (2008) Secrecy capacity over correlated ergodic fading channel. *Proceedings of the IEEE Military Communications Conference, MILCOM 2008*, pp. 1–7.
- 15 Anderson, R.J. (2008) *Security Engineering – A Guide to Building Dependable Distributed Systems*, 2nd edition. Wiley, Indianapolis, Indiana, USA.
- 16 Win, M., Pinto, P. and Shepp, L. (2009) A mathematical theory of network interference and its applications. *Proceedings of the IEEE*, 97(2), 205–230.
- 17 Bloch, M. and Barros, J. (2011) *Physical-Layer Security: From Information Theory to Security Engineering*. Cambridge University Press, New York.
- 18 Rabbachin, A., Conti, A. and Win, M. (2011) Intentional network interference for denial of wireless eavesdropping. *Proceedings of the 2011 IEEE Global Telecommunications Conference (GLOBECOM 2011)*, pp. 1–6.
- 19 Vilela, J., Bloch, M., Barros, J. and McLaughlin, S. (2011) Wireless secrecy regions with friendly jamming. *IEEE Transactions on Information Forensics and Security*, 6(2), 256–266.
- 20 Gollakota, S. and Katabi, D. (2011) Physical layer wireless security made fast and channel independent. *2011 Proceedings of the IEEE INFOCOM*, pp. 1125–1133.
- 21 Harrison, W., Almeida, J., Bloch, M., McLaughlin, S. and Barros, J. (2013) Coding for secrecy: an overview of error-control coding techniques for physical-layer security. *IEEE Signal Processing Magazine*, 30(5), 41–50.
- 22 Li, X., Yu, C., Hizlan, M., tae Kim, W. and Park, S. (2013) Physical layer watermarking of direct sequence spread spectrum signals. *Proceedings of the IEEE Military Communications Conference, MILCOM 2013*, pp. 476–481.
- 23 Soderi, S., Dainelli, G., Iinatti, J. and Hamalainen, M. (2014) Signal fingerprinting in cognitive wireless networks. *Proceedings of the 2014 9th International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM)*, Oulu, pp. 266–270.
- 24 Politis, C. (2015) 5G – on the count of three ... paradigm shifts. In: *5G Radio Technology Seminar. Exploring Technical Challenges in the Emerging 5G Ecosystem*, pp. 1–29.

- 25 Zhang, H., Dong, P., Quan, W. and Hu, B. (2015) Promoting efficient communications for highspeed railway using smart collaborative networking. *IEEE Wireless Communications*, 22(6), 92–97.
- 26 I CL, Han, S., Xu, Z., Wang, S., Sun, Q. and Chen, Y. (2016) New paradigm of 5G wireless internet. *IEEE Journal on Selected Areas in Communications*, 34(3), 474–482.
- 27 Soderi, S., Mucchi, L., Hämäläinen, M., Piva, A. and Iinatti, J. (2017) Physical layer security based on spread-spectrum watermarking and jamming receiver. *Transactions of Emerging Tel. Tech.*, 28(7), June 2017.
- 28 Horn, G. and Schneider, P. (2015) Towards 5G security. *Proceedings of the 4th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*.
- 29 Yang, N., Wang, L., Geraci, G., Elkashlan, M., Yuan, J. and Di Renzo, M. (2015) Safeguarding 5G wireless communication networks using physical layer security. *IEEE Communications Magazine*, 53(4), 20–27.

7

5G-WLAN Security

Satish Anamalamudi¹, Abdur Rashid Sangi², Mohammed Alkatheiri³, Fahad T. Bin Muhaya⁴, and Chang Liu⁵

¹ Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huaian, China

² Beijing Huawei Technologies, Beijing, China

³ College of Computing and Information Technology, University of Jeddah, Saudi Arabia

⁴ Department of Management Information Systems, King Saud University, Saudi Arabia

⁵ Information and Communication Engineering, Dalian University of Technology, China

7.1 Chapter Overview

This chapter briefly describes the security considerations of WiFi and LiFi network interconnections with 5G networks. When designing 5G networks with short-range WiFi and LiFi connectivity, architectural considerations must be accompanied with respective security considerations, and such security considerations are expected to influence the architectural decisions. Therefore, this chapter proposes a security-based architectural model for short-range wireless networks (WiFi) and high-speed backbone wireless networks (5G networks). In addition, this chapter explains the security considerations of LiFi interconnection with 5G networks.

7.2 Introduction to WiFi-5G Networks Interoperability

7.2.1 WiFi (Wireless Local Area Network)

WiFi (Wireless Fidelity) is the radio technology for short-range communication that works within licensed spectrum bands (2.4 GHz or 5 GHz ISM bands) through IEEE 802.11 standards. In general, two kinds of access modes, namely “infrastructure mode” and “ad-hoc mode”, are allowed with IEEE 802.11 radio interfaces. Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) is the widely deployed protocol to contend with and access the shared wireless channel in packet switched networks. Access Points (AP) act as a “centralized controller” in the “infrastructure mode”, to enable the nodes to communicate with the wireless network. On the other hand, nodes within the ad-hoc mode need to co-ordinate themselves (no central controller) to communicate with the neighboring nodes within the network.

7.2.2 Interoperability of WiFi with 5G Networks

WiFi Alliance (Wireless Ethernet Compatibility Alliance) is an international, nonprofit organization that focuses on the manufacturing and interoperability of 802.11 wireless LAN devices. The widely deployed WiFi hotspots (IEEE 802.11 a/b/g/n) and wide area cellular networks (2G to 5G networks) open up new opportunities of heterogeneous connectivity through interoperability between WiFi and cellular networks. With interoperability, networked mobile device can dynamically use the multiple network interfaces (WiFi, cellular) to provide seamless network connectivity and user satisfaction, which is clearly shown in Figure 7.1. Ensuring security for interoperable networks (wireless LANs and cellular networks) is a challenging task. This is because mobile devices can connect to multiple interoperable networks concurrently. This challenging task ensures authentication of the users and provides security to the data transmission (confidentiality and integrity) across heterogeneous networks [6,7].

7.2.3 WiFi Security

Fifth Generation (5G) radio access technology will have the capability of supporting high data rates, low latency, ultra-high reliability, energy efficiency and extreme device density. The deployment of 5G networks can be realized by the development of LTE in combination with existing radio-access technologies. In addition, interoperability of LTE with the state-of-the-art radio access technology and high-speed wired backbone networks plays a significant role in proving Internet-of-Things (IoT) and Machine-to-Machine (M2M) communication [1,2]. The wireless connectivity of 5G provides a wide range of new applications, such as smart homes, traffic safety/control, critical infrastructure and industrial processes (time-critical applications). This shows the achievability of ubiquitous connectivity for any device and the kinds of application through 5G networks. 5G networks are highly heterogeneous, with small cells densely deployed in existing circuit switched based cellular networks. Apart from providing extremely high data rates with stringent latency requirements, security provisioning is

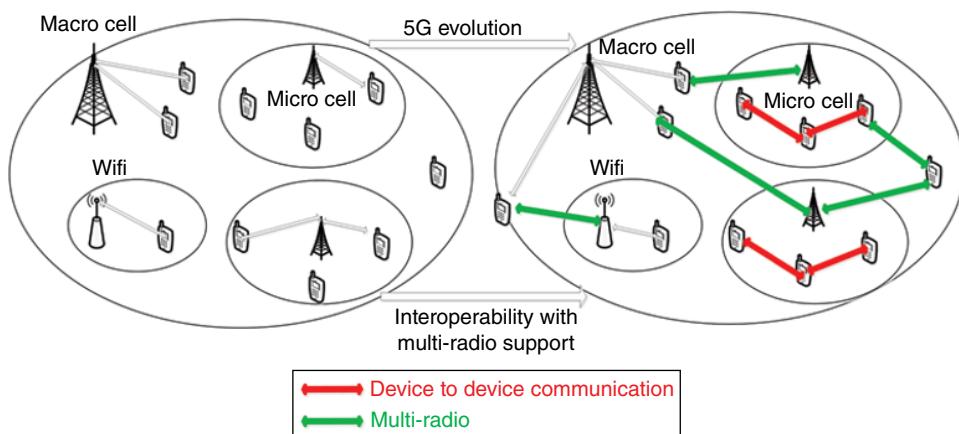


Figure 7.1 Interoperability of WiFi with 5G Networks [3].

becoming most challenging in 5G Heterogeneous Networks. Security management in 5G networks is difficult to handle in micro cells, where users join and leave frequently.

In addition, nodes will have frequent handover to different heterogeneous networks. On the other hand, authentications in small cells also introduce unnecessary latency. Access of 5G networks will be heterogeneous with respect to the properties of link layer, network bandwidth, end-to-end latency and availability, for example, extremely high frequency communication with millimeter (mm) radios. Thus, 5G networks offer super-high throughput with low latency, but only in very small cells. To achieve this, there will be new challenges such as ramping up TCP session sending rates with peers over the Internet. In addition, constrained networks (6LoWPAN) in IoT scenarios are evolving, which need low-bandwidth radios in GAIA.

Tracing out a good network abstraction for heterogeneous networks is significant to achieve the goal of 5G networks. The most crucial challenge is with the physical scarcity of radio spectrum allocated for radio communications. State-of-the-art cellular frequencies are using ultra-high-frequency bands for mobile phones, normally ranging from several 100 MHz to several GHz [3,4]. Hence, these frequency spectrum bands are heavily used, making it difficult for cellular operators to acquire more spectrum bands. Apart from this, another challenge with the deployment of advanced wireless technologies comes at the cost of high node and network energy consumption. The increase of node and network energy consumption in wireless communication systems results in an increase in CO₂ emissions that is considered to be a major threat to the future environment. In addition, based on reports from cellular operators, the energy consumption of base stations (BSs) contributes to over 70% of overall electricity consumption of cellular networks [5]. Even though energy-efficient communication was not the initial requirement of 4G wireless systems, later it was recognized as an important issue to conserve node and network energy consumption. Apart from this, other challenges are average spectral efficiency, high data rate and high mobility, seamless coverage, diverse quality of service (QoS) requirements, and fragmented user experience.

Due to the above-mentioned issues, there is an increased pressure on cellular service providers, who are currently facing continuous increasing demand for high data rates, wider network capacity, increased spectral efficiency, higher energy efficiency and higher mobility required through new wireless applications. On the other hand, currently deployed 4G networks have almost reached the theoretical limit on the data rate with existing technologies. This shows that current 4G networks are not sufficient to accommodate the above challenges. Therefore, it requires ground-breaking wireless technologies to solve the above issues caused by trillions of mobile devices. Researchers and mobile operators have already started to investigate the networks beyond 4G (B4G), which are named as 5G wireless techniques. 5G networks are expected to be standardized around 2020. It is expected and widely agreed that compared to the 4G network, the 5G network should attain 1000 times the system capacity, 10 times the spectral efficiency, energy efficiency and data rate (i.e. peak data rate of 10 Gb/s for low mobility and peak data rate of 1 Gb/s for high mobility), and 25 times the average cell throughput [5].

The objective of 5G networks is to connect to the entire world (anywhere (Ubiquitous), anything (IoT)), and achieve seamless communications between people-to-people, people-to-machine, and machine-to-machine wherever they are (anywhere), whenever they need (anytime), and by whatever electronic devices/services/networks

they wish (anyhow). This clearly depicts that 5G networks should be able to support radio communications for some special scenarios that are not supported by current 4G networks (e.g. for high-speed train users). The basic idea of designing the 5G cellular architecture is to separate outdoor and indoor scenarios, so that signal penetration loss through building walls can be avoided. This will be served by distributed antenna system (DAS) and massive MIMO technology, where geographically distributed antenna arrays with tens or hundreds of antenna elements are deployed [9]. Even though most of the current MIMO systems utilize two to four antennas, the main goal of the massive MIMO system is to exploit the potentially large capacity gains that will arise with larger arrays of antennas.

Outdoor Base Stations (BSs) can be equipped with large antenna arrays, with some antenna elements distributed around the cell and connected to the base station through optical fibers, benefiting from both DAS and massive MIMO technologies. Outdoor mobile users are generally equipped with a limited number of antenna elements. However, they can collaborate with each other to form a virtual large antenna array, which together with base station antenna arrays will construct virtual massive MIMO links. Large antenna arrays can also be installed outside every building to communicate with the outdoor base station or distributed antenna elements of the base station, possibly with line-of-sight (LoS) components. Large antenna arrays will have cables connected to the wireless access points inside the building communicating with indoor users. This can surely increase the infrastructure cost in the short term, while significantly improving the cell average throughput, spectral efficiency, energy efficiency, and data rate of the cellular system over the longer duration.

Security is a prime consideration when planning, designing, deploying and managing a network infrastructure. Design of Wireless LANs with heterogeneous network interoperability present a unique set of challenges to IT and security professionals. In addition, problems like non-secure wireless LANs can expose an organization's network traffic and resources to unauthorized outsiders. Such intruders may capture data and exploit network-based resources in wireless Internet access. Moreover, high-speed wireless access to a backbone network can represent the entry point for various types of attacks, which can crash an entire network and render services unavailable. Security design for interoperable heterogeneous networks is more crucial to provide uninterrupted anywhere (wireless), everywhere (seamless) and anything (IoT) services through 5G networks. When designing 5G networks with short-range WiFi connectivity, architectural considerations must be accompanied by respective security considerations, and such security considerations are expected to influence architectural decisions. This chapter proposes a security-based architectural model for short-range wireless networks (WiFi) and high-speed backbone wireless networks (5G networks).

7.3 Overview of Network Architecture for WiFi-5G Networks Interoperability

Interoperability of wireless networks and cellular networks [8] can be incorporated at multiple levels into the network protocol stack of the mobile device's operating system, as explained below.

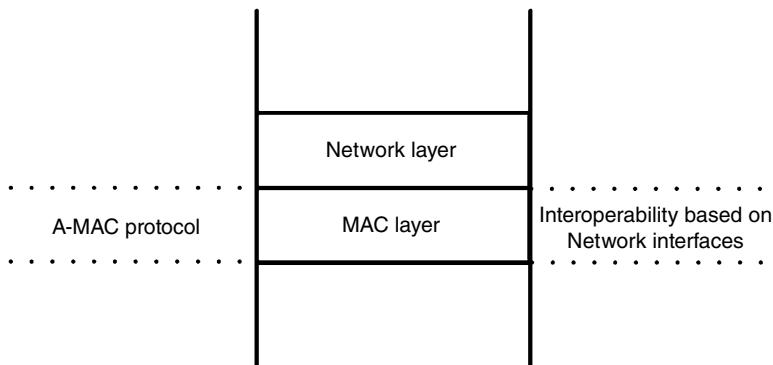


Figure 7.2 5G-WiFi Interoperability at MAC layer.

7.3.1 MAC Layer

A-MAC is proposed as part of the AdaptNet protocol suite (Figure 7.2) [9]. This is a two-layered MAC protocol that can support interoperability for wireless and cellular networks. The master sub-layer of A-MAC protocol is responsible for forwarding the data to the heterogeneous network interfaces based on the virtual cube concept. The decision-making module with the virtual cube concept can be replaced with a dynamic switching module to support the user profile-based interoperability (explained in Section 5.2 [9]).

During packet switching (frame switching) from one communication technology (WiFi) to another (5G), ensuring security for the application data, is of utmost importance to protect the data from intruders. In general, it is easy for intruders to alter the information at the interoperable stage where protocols of different technologies have to be compatible to switch the L2 packet. Hence, it is important for network designers to consider the security issues during the MAC level interoperability to ensure the application data protection at “Network Interface” switching. The state-of-the-art security issues for the MAC level interoperability is clearly described in Section 8.3.2.

7.3.2 Network Layer

Existing networks (wired and wireless) are largely depending on Internet technologies (IP-based networks). However, it is significant to develop a large set of additional gear to support services of non-IP based networks. Let us consider an example of mobility management. With service requirement of “seamless connectivity”, 4G networks employ an anchor point-based mobility approach that is being implemented through tunneling. This can be achieved either with GTP- or proxy-MIP-based solutions. This kind of solution results in centralized design with the usual inefficiency, scalability and security problems. Hence, explorations started to invent new technologies such as Selected Traffic Offload addressing requirements for web access. The most common use case for mobile Internet connectivity, which does not necessarily require seamless IP connectivity, is explained in [9]. The challenge of implementing interoperability at the network layer comes with maintaining a single static IP address (IPv4 or IPv6) for the mobile device across different network interfaces. A single static IP address for a mobile

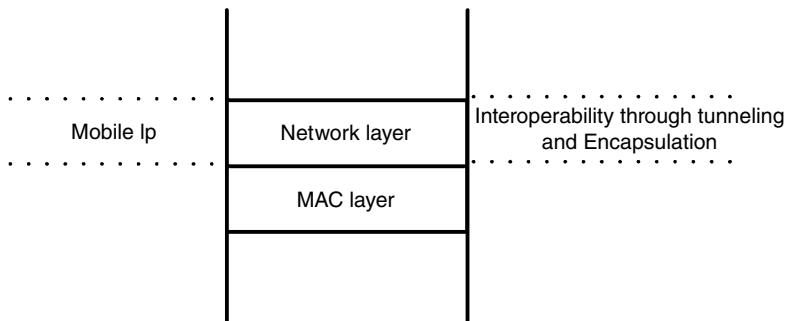


Figure 7.3 5G-WiFi Interoperability at the Network Layer.

device can be achieved through the concept of optimized Mobile IP. The mobile device is considered to have moved away from the home network to a foreign network whenever it switches to a new network interface. Using tunneling and encapsulation, it is easy to ensure that end-to-end communication is provided through a “source IP address”. But this solution requires the WiFi APs and the cellular BSs to support the operation of Mobile IP protocol (tunneling and encapsulation at Foreign Agents).

Ensuring security for the L3 data (IP packets) is prominent for network layer interoperability (Figure 7.3). Mobile IP, a network layer protocol, provides everywhere and anywhere Internet connectivity through IP-in-IP encapsulation and L3-tunneling. State-of-the-art research has come up with an extensive security design to provide the security for the L3 data protection through IP security (IPSec). Mobile IP with IPSec is one way to provide the security for application data (IP packets) at the network layer. Optimized Mobile IP is designed to select the better packet traversal path between source and destination.

7.3.3 Transport Layer

Interoperability implementation at the transport layer involves the major challenge of maintaining end-to-end TCP connections across switching between multiple interfaces (Figure 7.4). A TCP connection consists of 4-tuples (Source IP, Source Port, Destination IP and Destination Port). Since different interfaces have different IP addresses, switching the network interface will break the end-to-end TCP connection. Redirectable Sockets, RedSocks, is one of the solutions to solve such issues. With pTCP [10], smooth interoperability can be attained at the transport layer through bandwidth aggregation, which is achieved by stripping data across the multiple TCP connections.

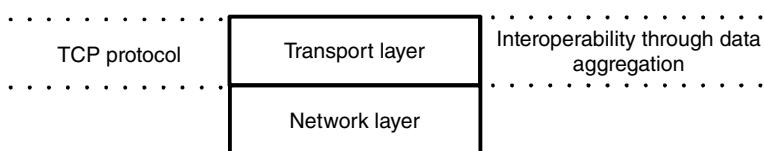


Figure 7.4 5G-WiFi interoperability at Transport layer.

Existing TCP proxies are used by mobile operators to tune the network performance with respect to the desired requirements. With current end-to-end TCP, congestion control does not work perfectly when it has to bridge the heterogeneous networks (cellular and wireless networks). This is because of more end-to-end delays and packet loss. In addition, it is noteworthy that existing TCP-based solutions do not work, especially in mobile networks, which actually design the system for virtual-circuit-like service. Hence, significant buffering, variable latency, no AQM, and no congestion notification will degrade the performance of traditional TCP in 5G networks. One way to overcome this issue is with TCP proxies, which can improve the performance of the cost of interoperability problems. In addition, security at Transport layer interoperability is significant to provide the application data protection for connection-oriented (Transmission Control Protocol) or connection-less (User Datagram Protocol) end-to-end data transmission between source and destination. Ensuring the security for transport layer protocols that provide end-to-end service to different applications is important at Transport layer interoperability. End-to-end encryption at the Transport layer will proliferate rapidly due to the integration of TLS into HTTP/2.

7.3.4 Application Layer

Application-traffic based optimizers are mostly designed for the caching, pacing and transcoding of video traffic. The design of an application-based optimizer can serve other purposes, such as user behavior analytics and statistics. These kinds of systems are implemented to provide a transparent chain of traffic classifiers, load balancers and the functionality of actual application. The traditional TCP/IP solution does not offer a caching on the network/transport layer and explicit HTTP proxies have interoperability problems at the application layer. Hence, HTTP/2 is designed to overcome the existing implementation complexity. However, this will introduce more difficulties due to the requirement of complex encryption. The possible solutions to have interoperability at the “Application layer” is through the new applications, written in such a way that they are aware of the multiple network interfaces, and may incorporate application specific considerations into the switching decisions (Figure 7.5). All calls to open a socket in the application must be modified so as to attach the socket to the interface specified by the switching module rather than to the default interface.

Security provision to application level interoperability is important to protect the data of different applications. In general, applications are going to provide different kinds of confidential service to the end user, which needs to be protected from the intruders. Standardization in the Internet area is designing new application level security protocols to provide end-to-end application level security that is compatible with heterogeneous networking technologies.

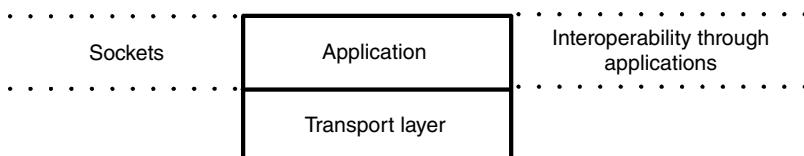


Figure 7.5 5G-WiFi interoperability at application layer.

7.4 5G-WiFi Security Challenges

State-of-the-art security protocols are designed to protect the application data that is being transmitted with single networking technology, and for specific protocol stacks (TCP/IP). This case of designing security protocols is valid for the standard packet switched networks, such as wired technologies (IEEE 802.3) and wireless technologies (IEEE 802.11). However, with the increased demand of network access in wired and wireless technology, a new design of hybrid access technologies through interoperability is being proposed by different Internet standardization organizations. For example, to overcome the shortage of natural available spectrum, Federal Communication Commission (FCC) has come up with a new kind of opportunistic spectrum access through “Software Defined Radio (SDR)” and “Cognitive Radio Network” technologies. At a same time, Internet Engineering Task Force (IETF) and other standardization organizations have proposed to design an IP-based protocol stack for constrained networks that should efficiently utilize the available resources (bandwidth, node transmit power, buffers) for Thing-to-Thing connectivity. In addition, Network Function Virtualization (NFV) is providing software-based Network Functions to provide the interoperability with different network vendors. The objective of 5G networks is to combine the functionalities of software-defined radio (SDR), network function virtualization (NFV) and Internet of Things-protocol suit to provide Internet connectivity through high-speed LTE-based radio access technology. With the deployment of 5G networks, the objective of network connectivity is to provide everywhere, anywhere, anything (Internet of Things) through interoperability of existing packet switched networks (Ethernet, WiFi, LiFi) and circuit switched networks (GSM, UMTS, LTE, etc.). Hence, it is important to plan, design and deploy the security protocols (authentication, confidentiality and integrity) that should protect the data during switching of the packets from one network technology to another through interoperability at any layer of the protocol stack (see Section 7.2 for interoperability at different protocol layers in the TCP/IP protocol suite).

Non-cryptography-based security provisioning methods through terminals or users can provide the fingerprint unique to a specific device. Thus, authentication procedures without complex hardware and additional computation costs are required for non-cryptography based security techniques. When the requirement is based on digital based cryptographic authentication, there will be high risks to the communication network once the security key is spoofed and compromised. Implementation of digital-based cryptographic provisioning or non-cryptographic based security provisioning is application dependent. For applications such as Internet surfing, gaming, non-critical applications and quick single attribute verification, a low security requirement is sufficient, which reduces the entire network complexity and enhances the performance of the 5G network. However, there are applications where improved authentication reliability is required to fulfill high-security requirements. For example, applications like banking or online shopping should be considered as critical secure-context-information (SCI). A new 5G management structure through Software Defined Networking (SDN) is proposed to bring intelligence and programmability to 5G heterogeneous networks for efficient security management.

With Software Defined Networking, the control logic is moved from the underlying infrastructure to a controller in the control layer. With this, software will be implemented upon the central SDN controller to provide the consistent and efficient management over entire 5G networks. Hence, an SDN-based authentication enabled scheme using SCI transfer is designed to provide fast authentication in 5G networks. In addition, seamless authentication during frequent handovers is achieved through SDN-based authentication.

7.4.1 WIFI-5G Security Challenges with Respect to a Large Number of Device Connectivity

In general, there are two classes of security requirements in WiFi or the 5G network, namely the mobile user's privacy and data integrity protection. The most common types of security challenges with increased device connectivity are [11,12]:

- *Impersonation attacks*: with this type of attack, the intruder pretends to impersonate a legitimate user to attack the networks. This case is common at the gateways, where the packet switch interoperates from one network (WiFi) to another network (5G);
- *Network security flaws*: this is where authentication of the network (5G or WiFi) is not provided to the end user. Hence, the intruder can pretend to be a legitimate user and join in any of the networks from where malicious packets can be forwarded and switched to the other networks at WiFi-5G gateways;
- *Anonymity gains*: once the intruder gains the information on the end users habits, then it is easy to compromise the network through calling patterns that can be used against the end user;
- *Attacks against confidentiality*: the intruder makes use of the loop holes in either circuit switched or packet switched network architecture (WiFi or 5G networks). In addition, the flaws in the communication protocols between the networks and the end user can trace out to attack the network by comprising weak confidentiality algorithms. Examples of this type are Brute force attacks, cryptanalysis-based attacks and non-cryptanalysis attacks;
- *Denial of Service (DoS) attacks*: the intruder floods the network to disable the end users and then accesses the network either by performing the attack using physical or logical intervention. Denial of service attack is the most prominent security attack on wireless networks.

7.4.2 Security Challenges in 5G Networks and WiFi

[13] clearly describes the security issues that exist in 5G Networks and WiFi. Since similar network architecture (UMTS, LTE) with different modulation/demodulation, data rates and cell coverage area (sensing/transmission) are used in 5G networks, the issues described in [13] will exist in 5G networks. The "Issue Type" and "Issue Description" for security issues in 5G networks are clearly described in Table.7.1.

The possible security attacks in the packet-switched based radio access network (WiFi) are briefly described in (Table 7.2) below:

Table 7.1 Security issues in 5G Networks [13].

| Security Issue Type | Security Issue Description |
|--------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Based on the type of the security attack | |
| Interception | The intruder intercepts the information through control/data signaling, but does not modify or delete; this kind of attack affects the privacy of the subscriber as well as the network operator |
| Reply attacks | The intruder can insert the unauthentic objects into the system that depends on the target and physical access type (e.g. spurious messages, fake service logic or fake subscriber information) |
| Resource modification | The intruder creates damage to the system by modifying the system resources |
| Interruption | The intruder tries to interrupt the operation by destroying the system resources (e.g. delete signaling messages, subscriber data, or stop delivery, etc.) |
| Based on methodologies used to cause the attack | |
| Attacks based on data | The intruder targets the information stored in the 5G communication system and causes damage by altering or inserting and/or deleting the data stored in the system |
| Attacks based on messages | The intruder targets the 5G system by adding, replacing, replaying and dropping the control/data signaling flowing to and from the 5G network |
| Service logic attacks | The intruder tries to inflict significant damage by simply attacking the service logic running in the various 5G network entities |
| Based on the level of physical access | |
| Class I | The intruder gains access to the radio interface using a physical device and then uses the modified mobile stations (eNodeB's) to broadcast the radio signal at a higher frequency, eavesdrop and execute “man-in-the-middle attacks” |
| Class II | The intruder gains access to the physical cables connecting the 5G network switches and may cause considerable damage by disrupting the normal transmission of control/data signaling messages |
| Class III | The intruder will have access to some of the sensitive components of the 5G network and can cause important impairments by changing the service logic or modifying the subscriber information stored in the 5G network entity |
| Class IV | The intruder has access to communication links connecting the Internet to the 5G network and can create disruption through transmission of control/data signaling flowing between the link and adding some new control/data signaling messages into the link between the two heterogeneous networks |
| Class V | The intruder has access to the Internet servers or cross-network servers providing services to mobile subscribers connected to the 5G network and can cause damage by changing the service logic or modifying the subscriber data (profile, security and services) stored in the cross-network servers |

Table 7.1 (Continued)

| Security Issue Type | Security Issue Description |
|----------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Access of unauthorized sensitive data | |
| Eavesdrop | The intruder intercepts the messages by continuously monitoring the operation of the communication network |
| Masquerading | The intruder frauds an authorized user by pretending that they are the legitimate users to obtain the confidential information from the end user or from the communication network |
| Analysis of the traffic flow | The intruder eavesdrops on the traffic flow through length, rate, time, source and destination of the traffic to trace the user location |
| Browsing | The intruder searches for data storage to trace the sensitive information |
| Data leakage | The intruder obtains sensitive information by exploiting the ways to access the legitimate user data |
| Inference | The intruder checks the reaction from a system by transmitting a query or control/data signal to the system |
| Manipulation of sensitive data | |
| Modification of user information | User information can be modified, inserted, replayed or deleted by the intruder deliberately |
| Unauthorized access to services | |
| Access rights | The intruder will access the services through masquerading as network entities or end user information |
| Physical layer issues | |
| Interference | The intruder intentionally creates man-made interference onto a communication medium, causing the communication system to stop functioning, due to high signal to noise ratio |
| Scrambling | This is a type of interference that is triggered based on short time intervals. A specific frame is targeted to disrupt a service. This kind of security attack is very complex to implement in a communication network |
| Medium Access Control (MAC) issues | |
| Location tracking | The intruder monitors the presence of user equipment in a specific cell coverage or across multiple cell coverage |
| Bandwidth stealing | The intruder creates this kind of attack by inserting the messages during the Discontinuous Reception (DRX) period or through utilizing fake buffer status reports |
| Open architecture security issues | As 5G networks are I-enabled networks with a high density of devices that are highly mobile and dynamic, an open architecture of an IP-based 5G results in an increase in the number of security threats |
| Security issues at higher layers | The departure from proprietary operating systems for handheld devices to open and standardized operating systems and the open nature of the network architecture and protocols results in an increasing number of potential security threats to the LTE wireless network and makes it vulnerable to a wide range of security attacks, including malwares, Trojans and viruses |

Table 7.2 Security issues in packet switched networks [13].

| Security Issue Type | Security Issue Description |
|-------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| De-authentication | <p>Based on attack type</p> <p>The intruder endeavors to defeat the authorization mechanism in Wireless LANs. With this, the intruder steals the legitimate wireless users' identities. In addition, the intruder tries to gain the control of authorized wireless access points' deployment rights and deploy rogue access points without going through the security process and review.</p> <p>The possible security issues that fall into this category are:</p> <ul style="list-style-type: none"> i) <i>MAC spoofing</i>: the intruder bypasses the MAC filtering policies through the modification of the MAC address of a wireless client; ii) <i>IP spoofing</i>: the intruder can evade the IP address-based authentication by pretending to be a legitimate user by altering the source IP address of the end-user; iii) <i>Rogue access points</i>: The intruder can gain open access to the WLAN by deploying an unauthorized wireless access point |
| Eavesdropping and interception in wireless domain | <p>The intruder can eavesdrop/intercept the legitimate wireless traffic flows by compromising the legitimate users' radio access channel. From this, he can gain access to all the information transmitted by the end user.</p> <p>Security attacks that comes under this category are:</p> <ul style="list-style-type: none"> i) <i>Network Traffic eavesdropping</i>: the intruder makes use of the network sniffer to eavesdrop the network traffic in the whole Wireless LAN; ii) <i>Man-in-the-middle attacks</i>: the intruder tries to obtain, intercept, modify and impersonate the end-to-end communication between source and destination, who believe they have a secure channel by monitoring in the middle of the two-way communications; iii) <i>Network injection</i>: the intruder injects fake network traffic into legitimate user traffic and makes an attempt to achieve malicious goals; iv) <i>Session hijacking</i>: the intruder will steal a legitimate authenticated conversation session ID and control the whole session of specific traffic flow between source and destination |
| Traffic jamming | <p>The intruder attacks by heavily consuming the bandwidth of the WLAN to drown the legitimate traffic, by flooding with fake messages or through high radio frequency signals. Attacks that comes under this category are:</p> <ul style="list-style-type: none"> i) <i>Denial of Service (DoS) attacks</i>: the intruder interrupts the legitimate user traffic to reach the receiver, by flooding with high frequency radio signals or fake messages; ii) <i>Spam attacks</i>: the intruder will launch spam attacks by flooding with spam messages over the wireless communication channels |
| Brute force attack to gain the control over Access Point (AP) passwords | <p>The intruder brute forces dictionary attacks to compromise the single shared password (in point-to-point (P2P) communication) of an access point by testing every possible password</p> |
| Attack against security protocols | <p>The intruder targets the information stored in the 5G communication system and causes damage by altering or inserting and/or deleting the data stored in the system</p> |
| Attacks based on messages | <p>The intruder compromises the vulnerabilities of existing Wired Equivalent Privacy (WEP) and WiFi Protected Access (WPA) security protocols</p> |
| Misconfiguration | <p>The intruder exploits the limited security knowledge of the WLAN administrator, human misconfiguration, or improper operations to gain control of the access point</p> |

Since the objective of the 5G network is to interoperate with any of the wireless networks, this chapter describes the abstract overview of the possible security attacks in other wireless technologies, such as WIMAX, Bluetooth, Vehicular Networks (VANETs), wireless mesh networks, Wireless sensor Networks and RFID [14]:

- *Security attacks in WIMAX:* the possible security attacks in WIMAX are:
 - i) parallel session;
 - ii) reflection;
 - iii) interleaving attacks;
 - iv) flaw attacks;
 - v) attack through name omission;
 - vi) cryptographic misuse attacks;
 - vii) identity theft; and
 - viii) Rogue Base Station (BS) attack.
- *Security attacks in Bluetooth:* the possible security attacks in the Bluetooth are namely:
 - i) Blue-jacking;
 - ii) Blue-bugging; and
 - iii) DoS attacks.
- *Security attacks in VANET:* the possible security attacks in the VANET are:
 - i) Sybil attack;
 - ii) node impersonation attack;
 - iii) traffic application attack;
 - iv) bogus information attack;
 - v) non-safety applications;
 - vi) time attacks;
 - vii) social attacks;
 - viii) monitoring; or
 - ix) sniffing attacks.
- *Security attacks in Wireless mesh Networks (WMNs):* the possible security attacks in wireless mesh networks are:
 - i) compromised mesh stations;
 - ii) mesh point disappearance;
 - iii) mesh point re-route; and
 - iv) cloning attacks.
- *Security attacks in the Wireless sensor Networks (WSNs):* the possible security attacks in the wireless sensor networks are:
 - i) jamming;
 - ii) tampering;
 - iii) collision;
 - iv) unfairness;
 - v) exhaustion;
 - vi) Sybil attack;
 - vii) sinkhole attack;
 - viii) wormhole attack;
 - ix) HELLO flooding; and
 - x) de-synchronization attacks.

- *Security attacks in the Radio Frequency Identifier (RFID)*: the possible security attacks in the RFID are:
 - i) illicit tracking;
 - ii) skimming;
 - iii) Mafia fraud;
 - iv) RFID malware; and
 - v) RFID DoS attacks.

7.5 Security Consideration for Architectural Design of WiFi-5G Networks

Section 7.4 briefly describes the state-of-the-art security attacks that are possible in 5G networks (cellular based network architecture) and WiFi networks. The existing research, design and implementation of the security protocols mainly focus on providing the security for specific network technologies and communication mediums [15–18]. However, the increased demand in access to the Internet, and flexibility over the usage of multiple services in a single networked device (e.g. smartphone) results in interoperability of heterogeneous network technologies. Hence, design of security protocols to counter attack intruder activities at the boarder router (Gateways) is crucial to protect user-data. One way of designing the security protocol is to handle the security issues that exist in both wireless networks and cellular networks. With this kind of security protocol design, the existing attacks listed in Tables 7.1 and 7.2 can be handled by a single security protocol at the interoperable gateway. Since the interoperability of the 5G and WiFi can occur at any layer (see Section 7.2.2) of the protocol stack, the design of security protocol needs special considerations for every separate layer of the protocol stack.

7.5.1 User and Device Identity Confidentiality

The term “confidentiality” defines that data privacy is assured between sender and receiver through the Internet. With confidentiality, no one can read the data except for the specific entity. Wireless mobile networks (WMN) only provide confidentiality of data through over-the-air and link encryption. In this sense, the data transmission between the base station (BS or eNodeB) and mobile station (MS) can be encrypted through symmetric or asymmetric cryptographic algorithms, but the data transmission between BSs is plaintext [19]. In addition, even though WMN implement data confidentiality and integrity partially within the network components, they do not provide end-to-end confidentiality and integrity to the user data. In order to provide secure end-to-end user communication, both sender and receiver of the communication must have a common encryption key (EK) for a symmetrical algorithm. Symmetric or asymmetric cryptographic algorithms should be compatible with different network architectures to provide confidentiality at the interoperable gateways.

7.5.2 Integrity

Data Integrity is known to be one of the basic elements of security. The objective of integrity models is to keep the data pure and trustworthy by protecting the system data

from any intentional or accidental changes. Integrity has been used for a long time in computer systems to attain three important goals, namely:

- 1) To protect data by prohibiting unauthorized user access to make any modifications to user data or network data;
- 2) To prohibit authorized users from making unauthorized modifications; and
- 3) To prevent internal and external consistency of data and programs [20].

There are two ways to measure the data integrity, criticality and credibility. Criticality can be described as the system that does not tolerate failures for internal and external events. Credibility can be described as something that is believed to have a degree of trust in place.

7.5.3 Mutual Authentication and Key Management

Password-based authentication is the most widely-used method for remote user authentication. Existing methodologies can be classified into two types, namely: (i) weak password approach; and (ii) strong password approach. The ElGamal cryptosystem is based on the weak password approach, where its advantage depends on the fact that it does not require a user ID-password table to check and verify the user login validity. On the other hand, the weak password approach depends on a heavy computational load on the node that constrains devices' (IoT) lack of capacity, which results in the difficulty in rendering 5G networks. One-way hash function and exclusive-OR (XOR) operations are being proposed as a strong password approach. In 2004, a dynamic ID-based remote user authentication scheme was proposed [21]. Based on this explanation, it requires much less computation and does not require any complex operations. Due to these reasons, this kind of proposed scheme certainly has advantages when implemented in constrained networks.

Key management plays a pivotal role in enforcing access control on the group key and in group communication. In addition, key management supports the establishment and maintenance of key relationships between valid groups based on the security policy being enforced on the corresponding group. The techniques and procedures that can carry out key management protocols are:

- *Member identification and authentication:* Authentication is a key factor to prevent an intruder from impersonating a legitimate group member. Furthermore, it is significant to prevent attackers from impersonating key managers. This clearly states that authentication mechanisms must be used to allow an entity to verify whether another entity is really who it claims to be [22].
- *Access control:* Once a group is identified, its joint operation should be validated. Access control will be performed in order to validate the group members before giving them access to group communication, particularly for the group key.
- *Generation, distribution and installation of keys:* It is a prerequisite to change the key at regular intervals to safeguard the secrecy of the key. In addition, extra care must be taken while choosing a new security key to guarantee key independence. Furthermore, each key must be completely independent from any of the previously used keys and keys going to be used in the future. Otherwise, compromised keys may reveal other keys that will create a security loop-hole.

In centralized techniques, a single entity is employed to control the whole group. Hence, a group key management protocol seeks to minimize storage requirements and computational power in both client and server.

7.6 LiFi Networks

With the increasing demand for wireless data communication, the existing wireless spectrum below 10 GHz (cm-wave communication) has become congested and insufficient. The Federal Communication Commission (FCC) has come up with different alternatives to minimize this challenge. One is to opportunistically re-use the available spread spectrum (below 10 GHz) through Software defined radio (SDR) and Cognitive radio networks. The other alternative is to consider the radio spectrum above 10 GHz (mm-wave communication) and provide wireless network connectivity through visible light communication. However, the higher frequencies (f) result in increased path loss (L) based on the Friis free space equation ($L \propto f^2$). Furthermore, blockage and shadowing in terrestrial communication are more difficult to overcome at higher frequency-based radio communication. Consequently, radio devices need to be designed to enhance the probability of line-of-sight (LoS) through beam forming techniques and by using cells with small radius (~50 m) [23]. The design of small cells is not an issue from a system capacity perspective, because reducing cell sizes has been the major contribution for enhanced system performance in existing cellular communications. On the other hand, usage of higher frequencies for terrestrial communication has become a practical option. However, one drawback is that the challenge to provide supporting infrastructure for smaller cells becomes an important consideration.

Light Fidelity (LiFi) is a bidirectional, high-speed and fully networked optical wireless communication technology that works in a similar way to WiFi. LiFi is a form of visible light communication and a subset of optical wireless communications (OWC). LiFi can be a complement to RF communication (WiFi or cellular networks), or even a replacement in the context of data broadcasting. LiFi takes the concept of visible light communication (VLC) further through light emitting diodes (LEDs) to realize the fully networked wireless systems. Synergies harnessed as luminaries become LiFi at the cells, resulting in enhanced wireless capacity providing the necessary connectivity to realize the Internet-of-Things, and contributing to the key performance indicators for the fifth generation of cellular systems (5G). Visible light communication makes use of “off-the-shelf” white light emitting diodes (LEDs) that are used for solid-state lighting (SSL) as signal transmitters and off-the-shelf p-intrinsic-n (PIN) photodiodes (PDs) or avalanche photo-diodes (APDs) as signal receivers [24]. This shows that VLC communication enables the system that illuminates and concurrently it provides broadband wireless data connectivity.

When the illumination is not desired in the uplink, then the infrared (IR) LEDs or indeed RF would be the viable solutions. In visible light communication, the application information is being carried by the intensity (power) of the light. This clearly shows that the information-carrying signal has to be real valued and strictly positive. State-of-the-art traditional digital modulation schemes for Radio Frequency (RF) communication use complex valued and bipolar signals. Modifications are certainly necessary where there will be a rich body of knowledge on modified multi-carrier modulation techniques such as OFDM for intensity modulation (IM) and direct detection (DD). The data rates of 3.5 Gb/s have been reported from a single LED. It is noteworthy that visible light communication is not subject to fast fading effects, as the wavelength is significantly smaller than the detector area. Although the link-level demonstrations are important steps to prove that VLC is a viable technique to help in mitigating the spectrum

bottlenecks in RF communications, it is important to depict that fully-fledged optical wireless networks can be developed by using state-of-the-art lighting infrastructures. This includes MU access techniques, interference coordination, and others. Visible light communication and mm-wave technologies are considered as energy efficient wireless communication solutions that are going to be deployed in 5G wireless systems. Let us consider an example where the VLC systems consuming the energy in one bulb are much less than that in its RF-based equivalent for transmitting the same high-density data [25].

7.7 Introduction to LiFi-5G Networks Interoperability

The interoperability of LiFi networks with 5G cellular networks is similar to the interoperability of WiFi-cellular networks. This clearly shows that interoperability of LiFi-5G networks can happen at any layer of the TCP/IP protocol stack. Hence, the interoperability of LiFi networks with 5G cellular networks can be incorporated at multiple levels in the TCP/IP network protocol stack of the mobile device's operating system or at LiFi Gateway nodes:

- *MAC Layer*: A-MAC is proposed as part of the AdaptNet protocol suite to support interoperability for wireless and cellular networks. Similar kinds of LiFi-based Adaptive MAC protocols need to be designed to provide interoperability between LiFi and 5G cellular networks through heterogeneous network interface cards (cellular-5G radio interface, LiFi radio interface). At the time of packet switching (frame switching) from one communication technology (LiFi) to another (5G), ensuring the security for the application data is of utmost importance to protect the data from intruders. The security issues in LiFi network interoperability with 5G networks will be different when compared to WiFi networks due to visible light communication.
- *Network Layer*: Ensuring security for the L3 data (IP packets) is prominent for Network layer interoperability. Mobile IP, which is a network layer protocol, is providing everywhere and anywhere Internet connectivity through IP-in-IP Encapsulation and L3-tunneling. State-of-the-art research has come up with extensive security design to provide the security for the L3 data protection through IP security (IPSec). Mobile IP with IPSec is one way to provide security for application data (IP packets) at the network layer. Optimized Mobile IP is designed to select the better packet traversal path between source and destination. When the interoperability is at the network layer, then the encapsulated packets should be secured at the gateway nodes (LiFi).
- *Transport Layer*: Interoperability implementation at the transport layer involves the major challenge of maintaining end-to-end TCP connections across switching between multiple interfaces. A TCP connection consists of 4-tuples (Source IP, Source Port, Destination IP and Destination Port). Since different interfaces have different IP addresses, switching the network interface will break the end-to-end TCP connection. Redirectable Sockets, RedSocks, is one of the solutions to this problem, which is explained in [3]. With pTCP [12], smooth interoperability can be attained at the transport layer through bandwidth aggregation, which is achieved by stripping data across the multiple TCP connections.

- *Application Layer:* Security provision to application level interoperability is important to protect the data of different applications for LiFi-5G heterogeneous network communication. In general, applications are going to provide different kinds of confidential service to the end user, which needs to be protected from the intruders. Standardization in the Internet area is designing new application level security protocols to provide end-to-end application level security compatible with heterogeneous networking technologies.

7.8 5G-LiFi Security Challenges

Security architectural design in LiFi to the high-speed wireless backbone network is also similar to WiFi technology. This chapter extends the discussion of LiFi to 5G security considerations during network architectural design.

7.8.1 LIFI-5G Security Challenges with Respect to a Large Number of Device Connectivity

The two classes of security requirements in LiFi or 5G networks are mobile user's privacy and data integrity protection. The most common types of security challenges with increased device connectivity are [26]:

- impersonation attacks at gateways;
- network security loop holes through authentication flaws;
- End user anonymity access gains;
- attacks against user confidentiality; and
- Denial of Service (DoS) attacks.

The security attacks at LiFi and WiFi network gateways will be common in most of the network operations.

7.8.2 Security Challenges in 5G Networks and LiFi

The security attacks on the LiFi-5G cellular interoperability can be similar to 5G-WiFi. Once the LiFi is deployed and interconnected with 5G-cellular networks, then new kinds of security issues in LiFi networks are investigated [27,28]. The types of security attacks are clearly explained in Table 7.1.

7.9 Security Consideration for Architectural Design of LiFi-5G Networks

This section briefly describes the state-of-the-art security attacks that are possible in 5G networks (cellular-based network architecture) and LiFi networks. The existing research, design and implementation of the security protocols mainly focus on providing security for specific network technologies and communication mediums. However, the increased demand in access to the Internet, and flexibility over the usage of multiple services in the single networked device, result in interoperability of heterogeneous

network technologies [29–31]. Hence, design of security protocols to counterattack the intruder activities at the boarder router (Gateways) is crucial to protect the user-data. One way of designing the security protocol is to handle the security issues that exist in both wireless networks and cellular networks. With this kind of security protocol design, the existing attacks in Tables 7.1 and 7.2 can be handled by single security protocol at the interoperable gateway. Since the interoperability of the 5G and LiFi can happen at any layer (see “Interoperability of LiFi with 5G Networks”) of the protocol stack, the design of security protocol needs to be designed separately for every layer of the protocol stack. The confidentiality of user and device identity, integrity, mutual authentication and key management are the same as in the WiFi-5G networks as mentioned in Section 7.5.

7.10 Conclusion and Future Work

Security considerations are crucial in everywhere, anywhere and anything connectivity. This chapter briefly explains the security consideration for interoperability of WiFi and LiFi networks with 5G networks. In addition, interoperability of WiFi and LiFi networks at different layers of the protocol stack is explained. Furthermore, security challenges and security considerations for the architectural model of WiFi and LiFi networks is also briefly described. In the future, new security challenges and their possible solutions will be briefly discussed once the 5G network solution is implemented and tested.

References

- 1 Marsch, P. et al. (2016) 5G Radio access network architecture: design guidelines and key considerations. *IEEE Communications Magazine*, 54(11), 24–32.
- 2 Luo, F.-L. and Zhang, C. (2016) *5G Standard Development: Technology and Roadmap*, in *Signal Processing for 5G: Algorithms and Implementations*, 1st edition. Wiley-IEEE Press, 616 pp.
- 3 Bangerter, B., Talwar, S., Arefi, R. and Stewart, K. (2014) Networks and devices for the 5G era. *IEEE Communications Magazine*, 52(2), 90–96.
- 4 Han, F., Zhao, S., Zhang, L. and Wu, J. (2016) Survey of strategies for switching off base stations in heterogeneous networks for greener 5G systems. *IEEE Access*, 4, 4959–4973.
- 5 Jahed, K., Fawaz, M. and Sharafeddine, S. (2015) Practical device-centric WiFi/cellular link aggregation mechanism for mobile devices. *Proceedings of the 11th International Conference on Innovations in Information Technology (IIT)*, Dubai, pp. 17–22.
- 6 Zahid, T., Hei, X. and Cheng, W. (2016) Understanding performance bottlenecks of a multi-BSS software defined WiFi network testbed. *Proceedings of the First IEEE International Conference on Computer Communication and the Internet (ICCCI)*, Wuhan, China, pp. 153–156.
- 7 He, Y., Chen, M., Ge, B. and Guizani, M. (2016) On WiFi offloading in heterogeneous networks: various incentives and trade-off strategies. *IEEE Communications Surveys & Tutorials*, 18(4), 2345–2385.
- 8 Jamali, A., Hemami, S.M.S., Berenjkoub, M. and Saidi, H. (2014) An adaptive MAC protocol for wireless LANs. *Journal of Communications and Networks*, 16, 311–321.

- 9 Akyildiz, I., Altunbasak, Y., Fekri, F. and Sivakumar, R. (2004) AdaptNet: an adaptive protocol suite for the next-generation wireless Internet. *IEEE Communications Magazine*, 42(3), 128–136.
- 10 Hsieh, H.-Y. and Sivakumar, R. (2002) pTCP: an end-to-end transport layer protocol for striped connections. *Proceedings of the 10th IEEE International Conference on Network Protocols*, pp. 24–33.
- 11 Wang, J., Jiang, N., Li, H., Niu, X. and Yang, Y. (2007) A simple authentication and key distribution protocol in wireless mobile networks. *Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing*, Shanghai, pp. 2282–2285.
- 12 Duan, X. and Wang, X. (2016) Fast authentication in 5G HetNet through SDN enabled weighted secure-context-information transfer. *Proceedings of the IEEE International Conference on Communications (ICC)*, Kuala Lumpur, pp. 1–6.
- 13 Baraković, S. et al. (2016) Security issues in wireless networks: an overview. *Proceedings of the XI International Symposium on Telecommunications (BIHTEL)*, Sarajevo, Bosnia and Herzegovina, pp. 1–6.
- 14 Singh, N. and Saini, M.S. (2016) A robust 4G/LTE network authentication for realization of flexible and robust security scheme. *Proceedings of the 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, pp. 3211–3216.
- 15 Sheela, S.J., Suresh, K.V. and Tandur, D. (2015) Security of industrial wireless sensor networks: a review. *Proceedings of the International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15)*, Bangalore, India, pp. 1–6.
- 16 Modares, H., Salleh, R. and Moravejosharieh, A. (2011) Overview of security issues in wireless sensor networks. *Proceedings of the Third International Conference on Computational Intelligence, Modelling & Simulation*, Langkawi, pp. 308–311.
- 17 Cao, J., Ma, M., Li, H., Zhang, Y. and Luo, Z. (2014) A survey on security aspects for LTE and LTE-A networks. *IEEE Communications Surveys & Tutorials*, 16(1), 283–302.
- 18 Rekhis, S., Chouchane, A. and Boudriga, N. (2008) Detection and reaction against DDoS attacks in cellular networks. *Proceedings of the 3rd International Conference on Information and Communication Technologies: From Theory to Applications*, Damascus, pp. 1–6.
- 19 Kutscher, D. (2016) It's the network: Towards better security and transport performance in 5G. *Proceedings of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, San Francisco, CA, pp. 656–661.
- 20 Arkko, J., Norrman, K., Näslund, M. and Sahlin, B. (2015) A USIM compatible 5G AKA protocol with perfect forward secrecy. *IEEE Trustcom/BigDataSE/ISPA*, Helsinki, pp. 1205–1209.
- 21 Liu, D. et al. (2016) User association in 5G Networks: a survey and an outlook. *IEEE Communications Surveys & Tutorials*, 18(2), 1018–1044.
- 22 Shao, S. et al. (2014) An indoor hybrid WiFi-VLC internet access system. *Proceedings of the IEEE 11th International Conference on Mobile Ad Hoc and Sensor Systems*, Philadelphia, PA, pp. 569–574.
- 23 Burchardt, H., Serafimovski, N., Tsonev, D., Videv, S. and Haas, H. (2014) VLC: beyond point-to-point communication. *IEEE Communications Magazine*, 52(7), 98–105.

- 24 Haas, H., Yin, L., Wang, Y. and Chen, C. (2016) What is LiFi? *Journal of Lightwave Technology*, 34(6), 1533–1544.
- 25 Aslam, M.U. and Rehman, A-U. (2005) Distributed authentication and authorization mechanism for wireless networks. *Pakistan Section Multitopic Conference*, Karachi, pp. 1–9.
- 26 Lopez, G., Gomez, A.F., Marin, R. and Canovas, O. (2005) A network access control approach based on the AAA architecture and authorization attributes. *Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium*, p. 8
- 27 Ma, W., Song, M., Wang, J., Wu, J., Ren, Z. and Song, J. (2009) A multi-role authorization method used in heterogeneous wireless network. *Proceedings of the 5th International Conference on Wireless Communications, Networking and Mobile Computing*, Beijing, pp. 1–4.
- 28 Um, H. and Delp, E.J. (2006) A new secure group key management scheme for multicast over wireless cellular networks. *Proceedings of the IEEE International Performance Computing and Communications Conference*, Phoenix, AZ, pp. 8–30.
- 29 Um, H. and Delp, E.J. (2006) A secure group key management scheme for wireless cellular networks. *Proceedings of the Third International Conference on Information Technology: New Generations (ITNG'06)*, Las Vegas, NV, pp. 414–419.
- 30 Ammayappan, K., Sastry, V.N. and Negi, A. (2009) Authentication and dynamic key management protocol based on certified tokens for manets. *Global Mobile Congress*, Shanghai, pp. 1–6.
- 31 Faisal, M. and Mathkoor, H. (2015) SDTP: Secure data transmission protocol in Ad Hoc Networks based on link-disjoint multipath routing. *Proceedings of the 2nd World Symposium on Web Applications and Networking (WSWAN)*, Soussepp, pp. 1–5.

8**Safety of 5G Network Physical Infrastructures**

Rui Travanca and João André

Portuguese National Laboratory for Civil Engineering

8.1 Introduction

The failure of crucial communications is one of the most shared characteristics of all disasters. Whether partial or complete, the failure of telecommunication infrastructures leads to damage to critical assets causing, for example, delays in emergency and disaster relief efforts. Despite the increasing reliability and resilience of telecommunications networks to physical damage in general, the risk level associated with communications failures remains severe because of the growing dependence upon these tools [1]. For example, on December 26, 2004, an earthquake of magnitude 9.2 hit the South Asia region causing a tsunami. Both events caused a large number of injuries, accidents, property damage and destruction to the telecommunications system, seriously damaging monopoles, towers and local switching equipment. Similarly, on May 12, 2008, a strong earthquake occurred in Sichuan, China, causing serious damage to all telecommunication systems. Fortunately, satellite communication was available and was used in relief operations. Other similar examples in recent past years include the July 16, 2007, earthquake of 6.7 magnitude in Japan; the August 15, 2007, earthquake of 8.0 magnitude in Peru; the January 12, 2010, earthquake of 7.0 magnitude in Haiti; the February 22, 2011, earthquake of 6.3 magnitude in New Zealand, or the March 11, 2011, 9.1 magnitude in Japan [2].

Communication structures are critical facilities and should be designed to remain operational during and after a major disaster. Consequently, the logical consequence should be to design buildings and facilities with higher standards, for example, by using technologies such as base isolation or passive energy dissipation systems in the case of buildings, and adopting the most recent design practices for masts and towers. Unfortunately, and contrasting with buildings (e.g. data centres), with the current state-of-the-art of design of masts and towers, it will not be possible to achieve the desired safety and performance objectives and, consequently, this chapter will be committed exclusively to these types of structures. The present chapter discusses the expected future of physical infrastructure used for communication networks, for example, monopoles, lattice towers and guyed masts, and also providing a solid background overview.

Figure 8.1 illustrates some examples of these structures, which are specially designed to support the network equipment, such as the antennas.

The exponential growth in the use of cellular phones has meant a new era for communication structures, smaller in height but larger in number. It is highly expected that this scenario will be more empathised with the 5G. The continued survivability of these infrastructures is often taken for granted, as parties involved frequently focus on the services provided, rather than on the structures that support the network equipment [3–6]. Though the design of monopoles used in the communication sector appears as a simple engineering problem and their unit costs are limited, a deeper investigation of these structures reveals a complete different scenario, that is, under wind action they are subjected to dynamic effects that are complex and they are built in such large numbers, which represents an important economic problem [4].

Failures involving these structures are amongst the most common types of accidents in communication engineering, often leading to disproportionate consequences in terms of economic and social impacts [4,6]. This reality calls for a paradigm change regarding the design, construction and operation of communication structures.

There are several stakeholders directly or indirectly concerned with these structures: researchers, designers, manufacturers, clients, consultants, insurers, contractors and sub-contractors. In this context, the design, construction and maintenance of these structures is usually done by specialised sub-contractors, in accordance with a standard design or with a specially developed design depending on the work complexity. In principle, good planning, design, construction and maintenance of communication structures are the keys for the success of every project. In particular, it is vital that synchronised planning and continuous knowledge exchange exists between the structural designer, the contractor, the subcontractors, and other players. Unfortunately, this is not always a reality [3,5–7]. Often, the design, maintenance and/or operation of communication structures are not usually treated as carefully as in the case of building or bridge structures. Also, clients' knowledge and interest about the structural requirements of these structures is usually very limited, possibly due to their cost being small when compared with that of the network equipment it supports [3–6]. These structures also do not receive the same level of research attention and research funding as occurs in buildings or bridges [3].

Until recently, national and international design codes/standards and/or guidance documents concerning communication structures were based on simple design procedures, whose use has been found to be inappropriate for some structural solutions, namely monopoles [4,6]. Nevertheless, even the provisions specified in the most recent design codes do not sufficiently address the dynamic interaction of the wind action between the structure and the equipment installed [8,9]. Quality management requirements are also usually not provided, as well as the principles to assess the extension of their design lifetime. As a result, the structural reliability levels implied in the design codes may not be achieved or maintained during the design working life of these structures. The previous statement is significant since the communication sector is becoming increasingly important in our modern societies over the last few decades, in particular with the globalisation of wireless communications.

Moreover, the increased significance of these structures comes at the same time when the life-span of existing structures is reaching its maturity, and the market has changed from being managed by a single public company to being liberalised into an open global



Figure 8.1 Examples of communication structures: (a) Monopole 50 metres high; (b) Lattice tower 50 metres high; and (c) Guyed mast 100 metres high.

market, while there is not enough information available and adequate communication within the sector concerning vulnerabilities. In addition, the physical infrastructures have recently started to be used as an additional resource of income by leasing their exploitation to third parties. As a result, the ageing communication structures suffer from lack of overall regulation and accountability and there is a lack of a clear understanding of system dependencies.

The above limitations often lead to vulnerabilities of these structures with respect to man-made (e.g. human error, explosions, impacts of objects) and natural disasters (e.g. hurricanes, tsunamis, floods or earthquakes). In a context of increasing threats due to terrorism and hazards due to climate change, the risks to which the safety of mobile communication networks structures are exposed are rising.

Climate change, natural catastrophes and failure of critical infrastructures are ranked at the top of the 2015 Global Risks database prepared by the World Economic Forum [10] and for which less progress has been made. In the period between 1980 and 2014, more than 20 000 natural disasters occurred worldwide, mainly meteorological and hydrological events [11–13]. Examples of recent natural disasters include, for example, the hurricane Katrina in the USA, which caused severe losses to the mobile communications infrastructure in Louisiana and Mississippi on August 29, 2005, and the great earthquake on March 11, 2011, in Japan, which significantly affected the Japanese mobile communication network.

To properly plan and manage the mobile communication networks structures, it is necessary to gather a correct understanding of the structural behaviour of such structures. Therefore, there is an urgent need to apply advanced research methods, for example, structural health monitoring, wind tunnel testing and numerical simulation and calibration, and develop a comprehensive risk framework to elaborate new, and review existing design guidelines, which will surely contribute to overcome the existing weaknesses.

This chapter will present and explore the most recent advances in research and innovation in the field of structural engineering applied to structures used in mobile communication networks.

8.2 Historical Development

8.2.1 Typology

Communication structures typologies vary widely across countries according to their uses, location and more commonly used materials. For example, in Portugal and Spain, where masts and towers are often installed in locations of high altitude (e.g. 500 m), the height of the structures will rarely exceed 60 metres. In contrast, in Denmark and in the Netherlands, where the elevation of the land is practically nonexistent, masts and towers with heights of over than 250 metres are common [4,5]. The choice of type of structure is frequently dictated purely by technical requirements. For example, certain antennas demand larger face widths for mounting and a high degree of resistance to angular structural deflection under extreme wind loading, so that signal quality is maintained. In the latter cases, heavier construction self-supporting towers are usually more suitable [3,6]. On the other hand, it is more efficient to attach radio and television antennas at the maximum height possible. This led to the advent of guyed masts.

However, there has been a sustained increase in the importance of environment aspects. The problem is effectively to achieve a balance between minimising intrusion to the environment and providing the services society has come to expect and demand. For example, monopoles have a higher material cost but lower land cost and are more aesthetically appealing when compared with lattice structures.

In terms of materials, traditionally hot-rolled high strength steel is the preferred choice. The weight of a self-supported tower, and thus its cost, will vary approximately with the square of the height, H . For a guyed mast, the cost will vary proportionally with the expression $H^{1.5}$ [4,5]. In recent years, alternative materials started to be used, such as glass-fibre reinforced concrete (GRC) [14], in new designs or high-modulus carbon fibre reinforced polymer (CFRP) [15] in retrofits of existing structures. However, at the present time, these materials still do not constitute a viable alternative solution to steel structures.

In general, it is possible to define the following five main typologies [3]:

- 1) self-supported lattice tower with a square base, generally with angle sections and bolted connections;
- 2) self-supported lattice tower with a triangular base, generally with tubular sections and with bolted connections;
- 3) monopole with octagonal, dodecagonal or hexadecagonal cross-sections, with several modules generally linked by forced fit;
- 4) monopole of tubular cross-section, with several modules linked by bolted connections;
- 5) lattice guyed mast with triangular or square sections.

Taking the case of Portugal as an example, Figure 8.2 presents the distribution of 385 communication structures according to different typologies. It can be observed that there is a preponderance of monopoles (typologies 3 and 4), representing 56.1% of all structures studied. It is also possible to verify that self-supported lattice towers with square or triangular bases (typologies 1 and 2) represent 35.8% of all structures [3].

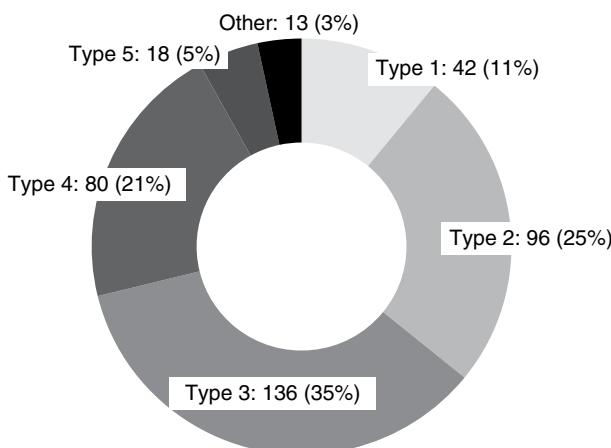


Figure 8.2 Distribution of a sample of tower structures by typologies [3].

8.2.2 Codes

The design of engineering structures can essentially be defined as a continuous process of making difficult engineering decisions based on the available knowledge and under the severe constraints imposed by society and nature. In the traditional approach, engineers resort to structural design codes to make decisions. These documents are developed specifically to address areas where significant past experience exists and where critical societal risks are not involved. Thereby, design codes are established for the purpose of providing a general, simple, safe and economically efficient basis for the design of ordinary structures under normal loading, operational and environmental conditions [16].

The absence of specific design codes or design guidance applicable to towers and masts, was recognized in the early 1970s and developments followed to address this situation (see [6] for details). Of significant importance were the studies performed by the International Association of Shell and Spatial Structures (IASS), from which the rules given in modern design codes for steel lattice towers and masts are still based.

In Europe, examples of early design codes are the German Code, DIN 4131, and the BS 8100 series (Part 1 to 4) [17–20] in the UK published in 1986. In the last decade, a joint European effort resulted in the publication of a suit of euronorms for structural design: the Eurocodes. Among the set of ten standards, EN 1993-3-1 [21] is directly concerned with steel towers and masts. It should not be forgotten that the Eurocodes are only valid if used together with the corresponding National Annexes published by every European Union member state, which contain the national choices for the Nationally Determined Parameters (NDPs). In the present book, the UK National Annexes will be used as an example. In the USA, the most important design code is the TIA-222 standard. The first version of this document was published in 1959. The current version is Revision G, published in 2016 [22] by the Telecommunications Industry Association (TIA). Other relevant documents are ASCE/SEI 48-11 [23] for monopoles and ASCE/SEI 10-15 [24] for lattice towers.

In this chapter, Eurocodes will be used as the reference design codes for communication structures. The versions of the documents used are those current at the time of writing. Code comparison exercises are presented in [6].

8.2.3 Outlook

There are sound reasons for believing that communication engineering will continue to innovate and advance technologically in the future; 5G is just the next step. Consequently, physical infrastructures of the existing and future mobile communication networks will continue to be subjected to difficult challenges as they need to adapt to service requirements in an increasingly competitive market, in particular, the growing demand for and importance of mobile communications services, namely 5G.

Concurrently, irrespective of the success of our mitigation efforts, the impact of climate change will increase in the coming decades. While efforts must continue towards mitigating its effects, there is no other choice but to take adaptation measures. Extreme weather and climate changes leave physical infrastructure systems exposed to different and more extreme and recurrent conditions. Since the available amount of resources is finite, it is highly likely that design thresholds, which are built into physical infrastructure project designs, may be breached more frequently in a future changing climate. This may result in threshold failures once considered

exceptional but acceptable, becoming unexceptional (i.e. normal) and unacceptable [25], such as accelerated degradation and interruption of vital services.

Therefore, there is a need to limit the consequences of failures and accelerate service resumption capabilities, both through engineering solutions and by managing consumer expectations [26]. Yet, research shows that the great majority of organisations managing critical infrastructure networks do not include climate change mitigation options, let alone adaptation strategies and resilience assessments in their strategic plans [27]. Insurers begin to consider this reality as a serious vulnerability and insurance premiums will surely rise for those who fail to demonstrate that they developed appropriate infrastructure resilience strategies, including climate change adaptation measures [28].

It is essential that stakeholders can turn the page to inefficient past practices and commit themselves to comprehensive and continuous planning and management policies of critical infrastructure assets, with the goal of reducing uncertainties, risks and magnitude of adverse consequences, and increasing sector and society safety, resilience and sustainability.

In recent years, many steps have been made towards understanding climate change and its effects, and developing sectoral plans to target resilience. However, all these plans lack the contribution from structural engineering, which is critical to understand the performance of physical infrastructures.

Of particular importance is the use of structural health monitoring techniques and equipment, such as fibre-optic sensors based on Fibre Bragg Gratings (FBG) [29]. In the last decade, there has been a growing interest in the field of structural health monitoring, resulting in the development of new techniques and equipment such as the fibre-optic sensors based on FBG. The recent improvement of sensors, based on all optical technology to study the dynamic behaviour of structures, presents itself as a valuable tool for the assessment of structural integrity and dynamic response of communication structures. Though conventional electronic accelerometers can be used, the high level of electromagnetic radiation near the antennas can easily mislead the interpretation of results and can also interfere with radio operation [29]. Another approach is the use of all-optical instrumentation like FBG accelerometers (Figure 8.3). Therefore, newly developed SHM techniques could be used to obtain valuable data about the structural behaviour, which would be used to validate, calibrate and/or verify numerical model simulations.

8.3 Structural Design Philosophy

8.3.1 Basis

Almost all modern structural design codes for the design and analysis of civil engineering infrastructures are based on the Limit State Design (LSD) principles, which in the USA is termed Load and Resistance Factor Design (LRFD).

The LSD principles are semi-probabilistic. In this methodology, the format for structural design verification is expressed by a simple comparison between factored resistances and factored actions (or action effects) without explicitly assessing the reliability or the risks. Due to the fact that resistances and actions are subject to uncertainties, probabilistic analyses were performed to derive statistically representative values (characteristic values) taking into account the design working life of the structure and the uncertainty of different physical properties and conditions.



Figure 8.3 Structural health monitoring of a 45 metre high monopole: (a) General view; (b) Detail of the sensor on top; and (c) Interrogation monitor unit.

To ensure that the basis for design provides an appropriate level of structural reliability (or probability of failure), partial factors are introduced to take into account the effects of aleatory and epistemic uncertainties in the methods used to assess the characteristic values but also in the specified analysis and verification procedures. Therefore, design values for resistances are determined by dividing the characteristic values by a partial factor, γ_R , (larger or equal to 1.0) and design values for load effects are obtained by multiplying the characteristic values by a partial factor, γ_S , (typically larger than 1.0).

LSD specifies the verification of the structural reliability for several limit states, that is states beyond which the structure no longer fulfils the relevant design criteria: Ultimate Limit States (ULS) in which all possible failure modes must be evaluated, Serviceability Limit States (SLS) in which it is verified that specified service requirements are met (e.g. maximum number of hours per year where the signal level is reduced below an acceptable limit), and other limit states such as fatigue resistance.

One should note that under certain hazard events, communication structures may be allowed to lose their functionality while not collapsing. However, it is often when such hazard events occur that demand for mobile communication services increase as civil protection and emergency services intervene, people try to contact and connect with their families, and remote access requests to businesses expand. In these situations, systems already severely strained can be pushed beyond their operational limits.

Finally, several load combinations must be checked to guarantee that all reasonable possible sets of physical conditions that can occur during a certain time interval, also known as design situations, are taken into account. In general, three different design situations are defined, each one representing a certain time interval with associated hazards, conditions and relevant structural limit states: persistent, transient and accidental situations, which refer to normal, temporary and exceptional situations. Each load combination is formed by the permanent loads, a leading variable action and the relevant accompanying variable actions, which are multiplied by combination factors (smaller than 1.0) in order to obtain concomitant actions values [30].

LSD has been established in the European Structural Eurocodes, namely BS EN 1990 [31,32], BS EN 1993 series, such BS EN 1993-3-1 [21,33], and in the USA, standard TIA-222-G [22] for example. By using BS EN 1990 and BS EN 1993-3-1, it is possible for the stakeholders to manage the target reliability level of the structure by specifying three reliability classes (Table 8.1). Regarding the design working life, BS EN 1993-3-1 recommends using a value of 30 years.

Table 8.1 Reliability classification in accordance with Eurocodes for ULS.

| Reliability classes | Reliability index, β_1 (1 year reference period) | Examples of mobile communication structures |
|---------------------|-----------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| RC1 | 4.2 | Towers and masts whose failure would likely result only in small economical, social and environmental consequences |
| RC2 | 4.7 | All towers and masts that cannot be defined in RC1 or RC3 |
| RC3 | 5.2 | Towers and masts whose failure would result in severe loss of human life, and/or in major economical, social and environmental consequences |

8.3.2 Actions

Actions are classified by their variation in time as [31,32,34]:

- 1) *Permanent action (G)*: an action that is likely to act continuously during a given reference period and for which the variation in space and in magnitude with time is insignificant;
- 2) *Variable action (Q)*: an action that is likely to act during a given reference period but for which the variation in space or in magnitude with time is not insignificant;
- 3) *Accidental action (A)*: an action, usually of short duration but of abnormal magnitude, that is unlikely to occur on a given structure during a given reference period.

In general, the most important actions applied to communication structures are wind action and the combined effect of icing and wind actions. Human errors during design, assembly, maintenance and operation activities are also a common occurrence, enabling cause of collapses. In this chapter, attention will be given only to wind action.

The wind velocity, V , can be decomposed into a mean wind velocity, V_m , with direction, u , and into fluctuating terms, V' , along three orthogonal directions u , v , and w [35]. The design wind velocity can be determined from BS EN 1991-1-4 [36,37], with additional rules specified in Annex B of BS EN 1993-3-1. In BS EN 1991-1-4, the reference wind velocities correspond to characteristic values of the wind velocity equal to the 50-year return period values. The reference wind velocity is considered equal to the maximum mean wind velocity averaged over 10 min, referenced to a height of 10 m over flat open country terrain.

Based on some prior information, namely location, altitude and orography of the site, the basic wind velocity is determined by the reference wind velocity. In BS EN 1991-1-4 + UK NA, this translates to:

$$V_b = c_{\text{dir}} \cdot c_{\text{season}} \cdot c_{\text{alt}} \cdot V_{b,0} \quad (8.1)$$

where [36,37] $V_{b,0}$ is the (reference) fundamental basic wind velocity, at 10 m above the ground of terrain roughness corresponding to “area with low vegetation such as grass and isolated obstacles (trees, buildings) with separations of at least 20 obstacle heights, that is terrain category II in the BS EN 1991-1-4 terminology: c_{dir} is a directional factor, with a recommended value equal to one; c_{season} is a seasonal factor, given in UK NA [37]; and c_{alt} is an altitude factor, given in UK NA [37], since the $V_{b,0}$ map reports to the sea level.

The UK NA specifies values of c_{dir} in increments of 30° (interpolation is allowed). The seasonal factor can only take values lower than one for structures during the construction phase. However, due to the uncertain nature of general construction activities, the codes recommend that a value of c_{season} equal to 1.0 be used. The distribution of the mean wind velocity, V_m , with height, z , needs also to be accounted for. This depends on the structure height, but also on the site orography and terrain roughness:

$$V_m(z) = c_r(z) \cdot c_o(z) \cdot V_b \quad (8.2)$$

where [36,37] c_r is a roughness factor, given in UK NA [37]; and c_o is an orography factor, given in UK NA [37].

The UK NA adopts the Deaves and Harris model to express the change of mean wind velocity with height and gives charts for the roughness factor that depend on three terrain types, instead of five categories suggested in BS EN 1991-1-4: Sea, Country terrain and Town terrain. When there is the choice between two or more terrain types in a given area, then the lowest value of terrain roughness length should be used. The UK NA reduces the complexities of orography assessment and effects on wind velocity by implementing the altitude factor, c_{alt} . Orography effects must only be evaluated according to the rules specified in Annex A.3 of BS EN 1991-1-4 for sites that fulfil the criteria specified in Figure NA.2 of UK NA [37]. The effects of interference between neighbouring structures are also addressed in BS EN 1991-1-4. For example, when a structure is located close to another structure that is at least twice its height, it could be exposed to higher wind velocities for certain wind directions. Annex A.4 of BS EN 1991-1-4 gives a conservative method to estimate this effect. Additionally, when groups of structures of similar height are packed closely together, they provide a shielding effect against wind action in a zone extending from the ground to about the average height of the top level of each structure: termed displacement height, h_{dis} . Therefore, the effective height to be considered for wind action assessment is $z - h_{dis}$. Rules to obtain the value of h_{dis} are given in BS EN 1991-1-4. The use of the latter concept implies that the neighbouring structures can only be removed after the end of the work period of the structure under consideration, and that the shielding effect is uniform in plan.

In the UK NA, the fluctuating component of the wind velocity is simulated by a turbulence intensity parameter, I_v , whose value for flat terrains can be determined from one of the several charts provided. Correction factors are also given to calculate the value for other types of terrain roughness. For sites where orography is important (see Figure NA.2 of the UK NA), the values of I_v must be divided by the orography factor, c_o , given in Section 4.3.3 of BS EN 1991-1-4. The two components of the wind, mean and fluctuating, are then combined to obtain the maximum peak gust velocity, \dot{V} :

$$\dot{V}(z) = [1 + g \cdot I_v(z)] \cdot V_m(z) \quad (8.3)$$

In the UK NA, a value of peak factor, g , equal to 3.0 is specified. Having determined the wind velocity, the wind effect on a body (element/structure) can be calculated. The simplest analysis is to analyse the along wind structural response, with more complex analyses needed for the across wind and torsional responses [38]. All modern design codes require the effects of wind on a structure to be accounted for, either by wind pressures or by wind forces. For mobile communication structures, the latter method is preferred.

The complex processes and relationships of wind engineering have been translated to code rules by assuming simplifications. One of the most important assumptions made in most of modern design codes dealing with wind effects is the quasi-static hypothesis, which assumes [35]:

- 1) the wind turbulence intensity is low;
- 2) the fluctuations of the effects of the wind action follow the variations of the direction of the mean wind velocity; and
- 3) there is a perfect correlation between fluctuating sectional forces.

Thus, according to the quasi-static hypothesis, the peak forces resulting from the mean and turbulent wind action components, along the mean wind velocity direction, can be determined by [36,37]:

$$F_w = c_s c_d \cdot c_f \cdot q_p(z_e) \cdot A_{ref}, \quad F_w = c_s c_d \cdot \sum_{\text{element } j=1}^{j=n} c_{f,j} \cdot q_p(z_{e,j}) \cdot A_{ref,j} \quad (8.4)$$

where $q_p(z_e)$ is the peak dynamic pressure at height, z_e , given by:

$$q_p(z_e) = \frac{1}{2} \cdot \rho \cdot \left\{ 1 + \left[g \cdot I_v(z_e) \right]^2 + 2 \cdot g \cdot I_v(z_e) \right\} \cdot \left[V_m(z_e) \right]^2 \quad (8.5)$$

z_e are reference heights defined in Section 7 of BS EN 1991-1-4, and F_w is the wind force acting on the whole structure or structural component. It may be given by the sum of the forces on all elements multiplied by the structural factor. $c_s c_d$ is the structural factor, which takes into account the fact that the peak velocity does not act simultaneously on a surface, c_s , and the dynamic response of the structure due to wind turbulence, c_d . The value of $c_s c_d$ can be calculated using the formulas presented in Section 6 and Annex B of the code (Annex C should not be used); c_f is the force coefficient given in Section 7 of the code and in the UK NA for various types of structure and structural elements; A_{ref} is a reference area, frequently the projected area of the body in the direction of the mean wind velocity.

Note that BS EN 1993-3-1 gives improved equations to calculate the peak forces due to the along wind action for lattice towers and guyed towers. To ease the calculation of the peak forces, it is customary to find in modern codes design charts with values of the exposure factor, c_e , and then determine the peak dynamic pressure by:

$$q_p(z_e) = c_e(z_e) \cdot q_b = c_e(z_e) \cdot \frac{1}{2} \cdot \rho \cdot V_b \quad (8.6)$$

Thus, the exposure factor combines wind gust, terrain roughness, height profile and orography effects into a single factor, and enables the 10 min mean wind velocity to be easily converted into a peak gust wind velocity, and wind pressure. The UK NA [37] recommends that the size factor, c_s , and the dynamic factor, c_d , be calculated separately, providing a table with values of the former factor and figures for the latter factor.

Estimates of the wind loading on mobile communication structures can be determined by summing the forces on individual members (including its ancillaries such as ladders, platforms, antennas, feeders and cables), using the force coefficients provided in Sections 7.6 to 7.10 of BS EN 1991-1-4 or in Annex B of BS EN 1993-3-1. For rectangular sections, the force coefficient values range between 2.4 for wide sections and 0.9 for long sections (see Section 7.6 of BS EN 1991-1-4), but usually a value equal to 2.0 is used. For circular sections (including cables, feeders, etc.), the value most often used is 1.2, although more accurate values can be obtained from figures that relate the force coefficient with the Reynolds number, Re . Finally, for angle sections, the most often used is 2.0 [35].

The above-mentioned procedure can be time-consuming and will give overly conservative values, except in cases where the solidity ratio is low. For specific cases of lattice structures, namely those having three or four columns (i.e. three or four faces), Section 7.11 of BS EN 1991-1-4 and Annex B of BS EN 1993-3-1 provide values of

overall (drag) force coefficients, which account for group and shielding effects. They should not be applied to the area of a single element, but directly to the area equal to the sum of the area of the elements of the most unfavourable exposed face (reference face), projected normal to the face.

The force coefficients provided in Sections 7.6 to 7.11 of BS EN 1991-1-4 correspond to infinitely long bodies (represented with symbol $c_{f,0}$). To account for the reduced wind effects caused by wind flows around the ends of a finite body, the code multiplies the force coefficients by an end-effect factor, $\psi\lambda$, which depends on the body slenderness, λ , and solidity ratio, φ . Care should be paid when determining the value of the end-effect factor: for the force coefficients provided in Sections 7.6 to 7.10, the slenderness and relative position should be determined for the single element, whereas for Section 7.11, the corresponding values should be determined for the reference face [35].

When there are elements with different shapes in a single face, the overall force coefficient could be determined by a weighted average:

$$c_f = c_{f,c} \cdot \frac{A_c}{A_{\text{face}}} + c_{f,c,\text{sup}} \cdot \frac{A_{c,\text{sup}}}{A_{\text{face}}} + c_{f,a} \cdot \frac{A_a}{A_{\text{face}}} \quad (8.7)$$

where A_c and $c_{f,c}$ represent the exposed area and force coefficient of the elements with circular shape in sub-critical flow regimes, respectively; $A_{c,\text{sup}}$ and $c_{f,c,\text{sup}}$ represent the exposed area and force coefficient of the elements with circular shape in supercritical flow regimes, respectively; A_a and $c_{f,a}$ represent the exposed area and force coefficient of the elements with angular shape, respectively.

Figure 8.4 illustrates the force coefficients provided in BS EN 1991-1-4 [36,37] and BS EN 1993-3-1 [21,33] for three or four faces lattice structures with angle elements. The data for solidity ratios smaller than 0.2 or greater 0.6 are merely indicative and should be used with prudence. When analysing Figure 8.4, it can be concluded that there is a reasonable agreement between the values given in different codes for solidity ratio values in the range of 0.2 to 0.6.

For lattice structures with circular elements, as the behaviour of the wind flow around a circular body depends on the Reynolds number, the definition of the value of the force coefficient to be used is more complex. For these structures, it should be evaluated if it

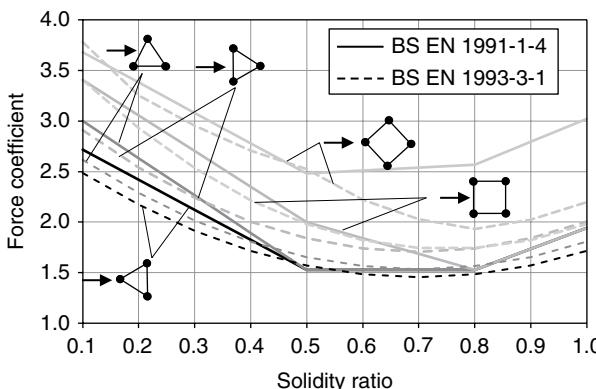


Figure 8.4 Force coefficients for three or four faces lattice structures with angle elements.

is possible that the highest values of wind effects occur for wind velocities smaller than the design wind velocity.

Some structures have attached several ancillaries. In these cases, the ancillaries' arrangements provide a certain degree of internal shielding to the structural elements. This positive effect is accounted for in BS EN 1993-3-1 + UK NA by reduction factors, K_A , used to determine the force coefficients of the structural elements (see Table 8.2).

Disturbed flows occur when wind flows past a bluff body causing the flow to separate from the surface of the structure rather than follow the body contour. At relatively low wind velocities, disturbed flows result in spiral vortices that are created periodically and symmetrically on either side of the body. However, for velocities higher than a limiting value, the vortices are shed alternatively, that is on one side first then the other. As a result, alternating low pressure zones are formed on the downstream side of the body and a fluctuating transverse force is created. This dynamic phenomenon is called vortex shedding [39,40] and may be significant for very slender and tall monopoles with circular cross-sections. Galloping is a self-induced dynamic phenomenon of flexible bodies and occurs when there are large amplitude lateral or torsional oscillations due to aerodynamic forces that are in-phase with the motion of the body [39,40]. Guyed masts are more susceptible to this phenomenon. Flutter is also a self-induced dynamic phenomenon that occurs when the natural frequencies of the torsional and lateral modes are very similar [39,40]. A famous example of a structural collapse due to flutter is the Tacoma Narrows Bridge failure of 1940.

BS EN 1991-1-4 [36,37] offers in Annex E expressions for determining the equivalent static wind load distribution due to the resonant response in the crosswind direction. These loads should also be accounted for in the structural analysis. It is conservative to simply combine the maximum values of the loads due to the along wind (see Equation (8.4)) and due to the crosswind. A well-known rule of thumb is to determine the fundamental natural frequency of the structure, and if this value is below 1 Hz, the resonant response may be significant [39,41,42]. Through most monopoles having a fundamental natural frequency lower than 1 Hz, dynamic analysis should therefore be performed.

In the case of lattice towers, the simplified quasi-static design procedures are in general conservative. For guyed masts, however, such a procedure may not be appropriate, as the vibration modes are not well separated and several modes can all contribute significantly to the response of the structure under a fluctuating wind action. Since a full

Table 8.2 Reduction factor, K_A , for ancillary items.

| Position of ancillaries | Ancillaries conforming to BS EN 1993-3-1, B.2.3(2) | | Ancillaries not conforming to BS EN 1993-3-1:2006, B.2.3(2), and circular sections in supercritical flow |
|-------------------------|----------------------------------------------------|----------------------|----------------------------------------------------------------------------------------------------------|
| | Square or rectangular plan form | Triangular plan form | |
| Internal to the section | 0.6 | 0.5 | 1.0 |
| External to the section | 0.7 | 0.6 | 1.0 |

dynamic analysis is complex and time-consuming, a simplified analysis method was developed by using an equivalent quasi-static analysis through “patch” loading techniques (see Annex B of EN 1993-3-1).

8.3.3 Structural Analysis

The general format for load combinations specified in the various codes for ULS verification (applicable both to persistent and transient design situations; not applicable to accidental and seismic design situations) can be expressed by [31,32]:

$$\frac{\gamma_{G,\max}}{\gamma_{G,\min}} \left| \times \sum G_{k,j} + \gamma_{Q,1} \times Q_{k,1} + \sum_0^{1,0} \right| \times \gamma_{Q,i} \times \psi_{0,i} \times Q_{k,i} \quad (8.8)$$

where $\gamma_{G,\max}$ and $\gamma_{G,\min}$ represent the maximum and minimum partial factors to be applied to the characteristic value of permanent (dead) loads, $G_{k,j}$, respectively; $\gamma_{Q,1}$ represents the partial factor to be applied to the characteristic value of the leading variable (live) action, $Q_{k,1}$; $\gamma_{Q,i}$ represents the partial factors to be applied to the characteristic value of accompanying variable loads, $Q_{k,i}$ and $\psi_{0,i}$ represents the combination factors of each $Q_{k,i}$.

For ULS, in terms of actions partial factors, these should be taken as equal to $\gamma_G = 1.10$ and $\gamma_Q = 1.45$ for RC2, according to UK NA of BS EN 1933-3-1, while the values of $\psi_{0,i}$ are given in BS EN 1990 + UK NA [31,32]. Regarding resistance partial factors, values for structural members of towers are equal to 1.0, while for connection elements they are equal to 1.25. For SLS, all partial factors are set to 1.0.

The designer of communication structures can use many types of structural analysis: from the simple first-order elastic analysis, which can only approximately simulate the “real” behaviour of the structure, to the complex but more accurate geometrical and material nonlinear imperfect analysis (GMNIA).

The calculation model and basic assumptions for the calculations should reflect the structural behaviour at the relevant limit state with appropriate accuracy and reflect the anticipated type of behaviour of the cross-sections, members and joints. The analysis shall be based upon calculation models of the structure which are appropriate for the limit state under consideration. The method used for the analysis shall be consistent with the design assumptions. For example, there are limits for the validity of application of linear analyses: they return accurate results only if the design loads are sufficiently smaller than the critical buckling load of the structures. In all other cases, second-order analyses should be performed [43].

Additionally, the analyses can be static or dynamic. In the former case, and when the resonant part of the response is not significant, equivalent static loads may be used to take into account conservatively the dynamic effects. The latter is usually performed by multiplying the static loads by notional dynamic factors.

Finite element analyses (FEA) of structures can use beam, shell or solid elements, each with their own advantages and disadvantages. If beam elements are used, local buckling effects could be simply simulated by considering an effective cross-section, or by other equivalent methods. Appropriate allowances should be incorporated into the structural analysis to cover the effects of initial imperfections, including residual stresses and geometrical imperfections such as lack of verticality, lack of straightness, lack of flatness, lack of roundness, dimples and minor eccentricities present in joints due to fabrication and/or erection operations. The joint effect of the most common types of

deviations from the idealised perfect state of the structure is often considered in a simplified manner using equivalent initial geometric imperfections whose values were calibrated to obtain conservative results [43].

The following equivalent initial geometric imperfections are typically introduced in the analysis:

- global imperfections expressed as initial sway imperfections of the system; and
- local imperfections, such as element out of straightness, joint looseness and load eccentricities.

8.3.4 Steel Design Verifications

8.3.4.1 Ultimate Limit States

Regarding the verifications of the ultimate limit states (ULS), code rules usually divide them into five different parts [16]:

- 1) resistance of the steel cross-section against tensile, compressive, shear and torsion forces, individually or combined;
- 2) resistance of the steel element: flexural buckling, torsional buckling, flexural-torsional buckling, coupled local and global buckling, shear buckling, flange-induced buckling and web resistance to high local transverse forces (web buckling, crippling and crushing);
- 3) resistance of the connections against tensile, compressive, shear and torsion forces, individually or combined;
- 4) resistance to fatigue and brittle failure; and
- 5) loss of equilibrium of the structure or any part of it, considered as a rigid body.

The design procedures to verify the safety of steel elements to these ULS are given in the various parts of Eurocode 3, namely BS EN 1993-1-1 [44,45], BS EN 1993-1-6 [46], BS EN 1993-1-8 [47,48], BS EN 1993-1-9 [49,50] and BS EN 1993-3-1 [21,33].

Of particular relevance to the safety of monopoles is the resistance of the bolted joints connecting adjoining tubular sections and also connecting the base of the tower to the foundation element. However, guidance is not provided in the Eurocodes. A suitable method is proposed in the Steel Design & Construction Bulletins 65, 66, 67 and 78 [51] published by the New Zealand Heavy Engineering Research Association (HERA); reformulated in [52] to suit the requirements of the Eurocodes.

For the design of anchoring elements to concrete, reference is made to the rules given in design documents such as [53–57] or to the design guidelines included in the technical manuals developed by the manufacturers of proprietary fastening solutions, as long as they comply with the principles and requirements for structural design set in the applicable design codes. Safety against relevant anchor failure modes should be evaluated, namely [16]:

- tension resistance of the steel anchor;
- shear resistance of the steel anchor;
- tension resistance of the concrete;
- shear resistance of the concrete;
- bond resistance in the interface between the anchor and the surrounding concrete; and
- concrete splitting resistance.

Concerning design of geotechnical structures, such as foundation elements, reference is made in Europe to BS EN 1997-1 [58,59]. In particular, Annex D provides a simple analytical method for bearing resistance calculation, which can be used for the basis of calculation.

Finally, with respect to design against seismic events, if simplified analysis methods that consider the linear elastic response of the structural system, design codes may allow for a reduction of seismic forces (dividing them by behaviour factors equal to or larger than one) in order to account for the nonlinear response of the structure. The values of behaviour factors strongly depend on the structural system configuration, elastoplastic behaviour and robustness.

8.3.4.2 Serviceability Limit States

The serviceability limit states (SLS) verifications are usually specified in the form of deflections and rotations limits, as well as vibrations and stress limits. Regarding load combinations, the same guidance as for ULS is used, although the loads involved can differ in order to take into account specific operating conditions.

Under the most adverse load combination, it is often considered the SLS to be expressed by a maximum lateral displacement at the top of the structure in the along wind direction equal to $h/50$ (where h represents the height of the structure). In the crosswind direction, for the RC2 class, the maximum lateral displacement at the top of the structure should not exceed 10% of the diameter that encloses the top section of the tower.

8.4 Survey of Problems

8.4.1 General

There are still large uncertainty levels and many complexities involved in the development of strategic plans for the future of communication structures. Service disruptions caused by physical destruction tend to be more severe and last longer than those caused by disconnection or congestion, because of the time and funding needed to repair and/or replace the structure. Historically, communication networks have been highly vulnerable to physical destruction. Seen as the most sophisticated and fragile urban infrastructure, communications networks are damaged in nearly every major urban disaster.

In the recent years, performance-based design philosophy has been gradually introduced in design standards, but little has been made towards resilience-based design. For example, for certain design scenarios, key physical infrastructures are required to withstand the event but do not need to remain functional. However, it is when such events occur that communication services increase as civil protection and emergency services intervene, people try to contact and connect with their families and remote access requests to businesses expand. In these situations, systems already severely strained are pushed beyond their operational limits. For example, in the 1995 earthquake that struck Kobe, Japan, communications failures prevented outsiders from receiving timely information about the severity of damage. These communications breakdowns delayed relief efforts for days, stranding tens of thousands of homeless victims outdoors in freezing winter weather. In a near future, such failures could result in unexpected events without precedents.

Therefore, a comprehensive public policy for critical infrastructure protection must begin with an understanding that “protection” should not be the main goal. Unlike other civil structures, such as buildings and bridges, even if one accepts the word “protection”, it is important to understand that what is being protected is not the infrastructure itself but the services provided. Unfortunately, our societies rarely reward investments that reduce vulnerability, in particular with respect to events so exceptional that there is no statistical basis for quantitative risk assessment. Management often focuses on keeping expenditure to a minimum and increase short-term revenues. Investments are therefore short-term intended, focusing on replacing and renewing as needed rather than modernising key physical infrastructure, and expenditure takes place ultimately in response to a crisis rather than proactively planning and managing key physical infrastructures. This is often ill-justified in economic terms since the economic risk from natural or man-made causes perceived by network operators often represents a small percentage of the capital at risk. Furthermore, the focus is on operating at near maximum operational capacity of the physical infrastructure, which is viewed as being an optimal and efficient management decision. However, such an option causes the systems to be less resilient against anticipated or unknown climatic and socio-demographic changes during the infrastructures lifespan. Investment in research and innovation towards physical infrastructure resilience and climate change adaptation is given low priority and only a small portion of the available amount is redirected to it. In the long term, network operators often expect that climatic changes can be addressed when old assets are replaced, using improved or, more likely, adapted technology.

8.4.2 Design Failures

The number of failures observed in communications structures is high when compared with other structures of equal economic and social importance. A great number of the failures observed are the result of poor design methods and practices, conducting to unsafe structures that are prone to collapse, partial or full [3,5–7]. The analysis and design of masts and towers require specific knowledge and expertise. In fact, a structurally sound definition of the basis of design of communication structures is a key issue. To address this special need, proper communication between the designer and client is necessary to create an economical structure with adequate reliability and performance. Unfortunately, the combination of an inexperienced client and a designer who is not familiar with the special problems associated with the design of communication structures is common. Over the years, this combination has created many structures that do not fulfil their intended purpose [3,5,6]. From the analysis of the problems recently observed in structures used in the communication sector, it is reasonable to conclude that design errors are the most frequent cause of collapse or early replacement events [3]. In addition, market pressure demanding lighter and more economical solutions has generated structures with structural limitations in terms of fatigue safety [3].

The predominant loading for the analysis and design of such structures are actions of random nature, namely the wind action or the combined action of wind and ice. Wind is a dynamic action and slender structures are sensitive to the turbulent component of the wind. The dynamic response becomes important when these structures exhibit a first natural frequency below 1 Hz. Therefore, dynamic analysis is necessary to determine whether the resonance response could be significant compared with the background response [3].

In particular, in monopoles with tubular cross-section or with a radome mounted on the top, the vortex shedding phenomena causes a random pressure in the plane perpendicular to the direction of the wind flow. If the frequency caused by vortex shedding coincides with the fundamental frequency of the structure, significant resonant vibrations will occur [3,6]. Because this condition occurs for a critical wind speed that is often observed, the assessment of the fatigue resistance may also become important. Although this phenomenon is widely known, its complexity makes the currently used empirical-based procedures for the assessment of the response too simple. Diverse failures caused by these phenomena were observed and are well documented [3,60]. Also, the presence of ice on the structure increases the mass, changes its dynamic characteristics and could severely alter the behaviour under wind action on the structure due to the increase of the exposed area coupled with potential asymmetries of the structural geometry [3,6]. These design shortcomings could be aggravated by fabrication faults, for example in the welding of the stiffeners where undersize welds are often used.

In structural design, many aspects are frequently ignored when using numerical model simulation, for example, soil-structure interaction, a more refined mass distribution and/or stiffness loss. The dynamic analysis of tall slender structures is commonly performed in the frequency domain, based on the frequency dependent character of both wind loads and mechanical properties of the structure. These factors depend on several parameters, including the first natural frequency of the structure, its damping and the characteristics of the wind [4,5]. Consequently, the numerical models currently available and commonly used by the industry do not reflect the actual field conditions and do not provide adequate information of the correct structural behaviour [3]. Also, several studies highlight the great difficulty in the adequate consideration for structural analysis of non-structural elements, which are protected and/or allow protection of the structure when exposed to wind action [8,9]. Shield effect is a well-known fact. However, there is no well-defined method available for the quantification of this important phenomenon. Thus, there may be a significant reduction of the current design values of the wind pressure at areas where there are head frames and/or multiple antennas, which are often determined by a simple sum of the wind action on each element [9].

In this context, there is also an urgent need for the development of the knowledge in this technical and scientific field. It is vital to perform an extensive review of the state-of-the-art and develop new approaches for the analysis and design of this type of structures.

8.4.3 Maintenance Failures

Before the exponential growth in the use of cellular phones, maintenance failures were mainly due to inappropriate maintenance operations, for example, the collapse of a very high guyed mast (over 300 m in height) during maintenance. Nowadays, with the huge number of small existing communication structures, the major failures occur more likely due to the lack of maintenance. In the 1990s, communication networks significantly expanded. During this period, governments and telecommunication companies intended to cover the national territories with the best possible network. As a result, there was a large-scale deployment of communication structures to respond to their needs. After this period, it was recognized as essential to maintain these structures in good structural conditions using periodic inspections, maintenance and occasional testing. Like any infrastructure, maintenance is essential to ensure the extension of the

life of the asset and the safety of its users. For the degradation of communication structures, several factors are involved, for example, the characteristics of the material used, of the construction process, of the environment and of the use of the structure. Still, the most common pathologies in steel communication structures are: paint degradation, steel corrosion and cleaver cracks [3] (Figure 8.5).

In some cases, it is possible to observe the fracture of bolts in connections between adjoining monopole sections. Such situations can occur due to unexpected flexibility

(a)



(b)



Figure 8.5 Inspection of a 30 metre high guyed mast: (a) Deterioration of the foundation; and (b) Corrosion of the cleavers.

of the structure. There are also situations where no sealing is applied, leading to water traps. Some defects in welded connections can also be detected by visual inspection, namely planar imperfection, slag inclusions, pores, undercut and/or profile imperfections [3] (Figures 8.6 and 8.7). The major effect of weld imperfections on the service performance of steel structures is to increase the risk of failure by fatigue or by brittle fracture.

(a)



(b)



Figure 8.6 Inspection of a 50 metre high monopole: (a) Planar imperfection; and (b) Defects in welded connections.

(a)



(b)



Figure 8.7 Inspection of a 40 metre high monopole: (a) Defects in welded connections, and (b) Defects in welded connections.

8.4.4 Vandalism or Terrorism Failures

Since the early 1970s, there has been an increasing number of cases of damage to communication structures from vandalism [6]. Communication structures are vulnerable to vandalism due to their own nature, that is, many are located in remote locations making access for protection difficult. Though communication structures can be a target for those who fully disagree with the service they are providing, the scale of such vandalism varies from carelessness to full terrorism. Carelessness should not strictly be considered as vandalism, but more correctly covers any type of accidents that should be prevented,

for example, adjacent civil engineering works having an adverse effect on the stability of the structure in question [6]. But of most concern are the higher-risk activities such as true vandalism or terrorism. In both cases there is a conscious intention to cause damage. For example, in May 2016, two communications guyed masts, key parts of Sweden's infrastructure, collapsed, on suspicion of sabotage.

Although the main responsibility for setting goals for the protection of critical infrastructure rests primarily with the governments, the application of the necessary measures to reduce the vulnerability of privately owned assets depends mainly on the private-sector action. Though the private sector surely understands their operations and the hazards they entail, it is clear that there is not currently an adequate incentive to fund vulnerability reduction, that is, the cost of reducing vulnerabilities outweighs the benefit of reduced risk from terrorist attacks as well as from natural or any other disasters [61]. In addition, risks to critical infrastructure industries are becoming more interdependent as the economic, technological and social processes of globalization continue to intensify.

With regard to terrorist threats, the policy goal should be to build capabilities for prevention of attacks that interrupt such services and for effective response and rapid recovery – if and when – such attacks do occur. As the scale and reach of these large technological systems have increased and certainly will continue to increase, the potential economic and social damage of failures has increased also. and will continue increasing. The sources of such major disruptions lie in technical and managerial failures as well as natural disasters or terrorist attacks, as demonstrated in the Northeast Blackout of August 2003 [61].

Also, economic and social activities are becoming more interdependent, so that the actions taken by one sector will affect others. As terrorist attacks have emerged as potential threats to critical infrastructure, private-sector executives and policymakers must face far greater uncertainties than ever before. In contrast, actions can be taken to reduce damage from future natural disasters with the knowledge that the probability associated with the hazard will not be affected by the adoption of these protective measures. The likelihood of an earthquake will not change if the structural design is focused on more quake-resistant structures. But the likelihood and consequences of a terrorist attack are determined by a mingling of strategies and counter-strategies developed by a range of stakeholders and always changing over time. This dynamic uncertainty makes the likelihood of future terrorist events extremely difficult to estimate and increases the difficulty of measuring the economic efficiency of both public policies and private strategies.

In that context, even if private-sector actors reduce system vulnerability by reducing dependence on vulnerable external services, for example decentralizing critical assets, decentralizing core operational functions, and adopting organization practices to improve resiliency, these organizations may face sanctions from the markets for taking such actions if they reduce efficiency, raise costs and/or reduce profits [61].

Codes of Practice for communication structures should be extended to include a risk assessment considering factors relating to vandalism, sabotage or dangerous trespass. The owner needs to consider carefully the likelihood of hazards occurring at the chosen site. The designer needs to minimize elements of the structure that are worth stealing and/or are accessible. The severity of any attack needs to be considered on both economic and risk to life terms. As our dependency on communication structures increases, more attention will need to be given to the risk of failures due to vandalism or terrorism.

8.5 Opportunities and Recommendations

The communication sector is becoming increasingly important in our modern society as communication within social, economic and industrial systems is becoming increasingly digital, wireless and interdependent, consumer market is globally expanding and there is an escalating offer/demand input. Standing on the edge of the 5G technological revolution, because of its scale, scope and complexity, its impact on society involves remarkable opportunities and benefits, but also large threats. It is not possible to know yet just how it will unfold, but one thing seems clear: the process must be an integrated and comprehensive initiative, involving all stakeholders, both from the public and private sectors, to academia and civil society.

Communication structures are critical infrastructures. One of today's most pressing challenges is to allocate limited resources to reduce as low as reasonable practicable the risks posed by natural and/or man-made hazards to these key physical infrastructures. Failure to do so will certainly have enormous consequences, as the value at risk is massive. Not by the economic value of the physical infrastructures themselves but because the present and future society welfare and sustainable development strongly depend on the capability of this system to continue to progress and to provide reliable public services and to adapt to our civilisation evolution.

Management of key physical infrastructures of communication networks can be achieved by adapting the existing or developing new standards and design methodologies.

Existing codes were calibrated so as to assure performance expectations focusing on the single asset level and based on historical practice, but with significant limitations in certain fields. Current code design philosophies continue to be mainly based on prescriptive rules, many of them calibrated only at the individual element level, from which the final design solution is deemed to satisfy a variety of different goals. When the engineer is confronted with an omission on the code about a given problem, related for instance to modelling or human errors, generally there is no other option than to resort to heuristic methods such as what is considered to represent good practice. The more subtle aspects of this approach are based on intuition, and are often referred to as "engineering judgement". However, examples abound of new issues and new problems, where the experience of previous work does not provide an adequate guide due to structural or economical aspects.

Furthermore, uncertainties can appear in the process of extrapolating past experience to existing problems due to differences in the current and past design methodologies. An evident example where such problems will arise is the planning, analysis, design, construction and management of key physical infrastructure of the communication networks. Additionally, there is a lack of recommendations and guidance towards resilience-based design, including adequate consideration of climate change effects on key physical infrastructures. Therefore, the present basis for design does not assure optimal design in terms of resources allocation and risk acceptance. The traditional standards-based approach is becoming increasingly inadequate to handle the allocation of limited resources for key physical infrastructures plan, design, operation or management, in a climate of growing public scrutiny.

In order to develop a methodological guide with recommendations to achieve more effective standards and design methodologies, it is indispensable to determine the

most important variables that control the structural fragility and system resilience. The need to protect critical infrastructure and save human lives in the case of a major event must be balanced with the amount of available resources, for example technical and financial. Usually, minimum levels of public services of communication are defined based on a reliability criterion, where a maximum number of service disruption events per year are fixed, or based on an availability criterion, where maximum service downtime hours per year are fixed. Even though the latter criterion is the current state-of-the-art, the approach used to define and demonstrate minimum levels of communication services does not find support in a comprehensive and holistic approach, lacks the contribution of trans-disciplinary disciplines, in particular of structural engineering, and does not account for systems resilience, climate change effects or other unexpected events.

As the frequency and severity of climate-related natural catastrophes is expected to increase and the current risks of disasters of geological origin pose a serious threat to physical assets located in vulnerable locations, including critical infrastructures along their life cycle, it is extremely important to develop a risk management framework that could be used to assess and help to manage risk to which physical infrastructures used for communication systems are exposed to.

In order to be able to prepare solutions, first it is necessary to consolidate the context of the problem. Therefore, the relevant technical, social, economic and regulatory specificities, past, present, and in the predictable future, of communication systems needs a proper discussion. For example, the characterisation of the key physical infrastructures, topology of the existing networks, and the precise identification of the main challenges to be solved to attain the intended objectives are part of this evolving process.

Next, risk assessment methods should be applied. This task encompasses risk identification, risk analysis and risk evaluation. The former consists in the formal, systematic and comprehensive compilation, review and use of the available information concerning relevant hazard scenarios; with appropriate consideration of the uncertainties involved. Subsequently, links should be established between hazards, consequences and causes, and their sensitivity to each individual contribution evaluated. The information obtained from various sources relative to natural and man-made hazard events relevant to communication systems should be combined, for example, earthquakes, storms, floods and terrorist attacks. In this regard, research is needed concerning hazards of which existing data is incomplete, insufficient or even non-existent, and taking into account the present and foreseeable exposure level of the key physical infrastructures to each hazard.

Risk analysis involves a suitable combination of numerical, experimental and monitoring methods. Nowadays, and due to the current state-of-the-art, all the participants in the construction and maintenance activities of communication structures use simplified procedures for the analysis and design of such structures. To meet the urgent and necessary evolution in this field, it is important to combine different areas of knowledge toward a common and primordial goal, that is the correct modelling and calibration of numerical models based on the results of experimental tests such as wind tunnel tests and the structural monitoring of existing structures. Currently, this interaction is very limited or non-existent. Consequently, the numerical models currently available and commonly used by the industry do not reflect the actual field conditions and do not

provide adequate information about the correct structural behaviour. Advanced and complex finite element models should be developed to study the behaviour and resistance of these structures, subjected to the major hazards that will be exposed to, in particular due to wind action.

In the last decade, there has been a growing interest in the field of structural health monitoring resulting in the development of new techniques and equipment such as the fibre-optic sensors based on Fibre Bragg Gratings. The recent improvement of sensors based on all optical technology to study the dynamic behaviour of structures presents itself as a valuable tool for the assessment of structural integrity and dynamic response of communication structures. Therefore, newly developed SHM techniques could be used to obtain valuable data about the structural behaviour, which would be used to validate and verify the numerical simulation models. After which advanced stochastic simulations, including sensitivity analysis and uncertainty quantification, could be performed. Results of the models should be then analysed to determine structural robustness and structural fragility based on innovative procedures [62].

Consequence models could be then developed to estimate the vulnerabilities of the system to each specific hazard, including direct and indirect consequences. Structural resilience models could also be elaborated on, based on the structural fragility and consequence models, but also on specific models that simulate relief and recovery activities after a disruptive event takes place. These models should be analysed using a consistent systems of systems framework. To this end, considerable attention should be given to the formulation of a modelling framework that captures relevant features of a complex system, including the systems intra- and inter- dependencies, and to the development of a simulation tool, integrating the new and innovative capabilities and improved knowledge obtained in the abovementioned tasks.

Risk evaluation is the process of examining and judging the significance of risk. First, the risk acceptance criteria should be established (e.g. using equity, utility, technology and sustainability principles) and the acceptable and the unacceptable risk (or resilience) levels are defined, for example following the ALARP (As Low As Reasonably Practicable) principle. Furthermore, a list with a range of alternative measures (active or passive, preventive or protective) for managing the risks which are higher than the acceptable risk level could be developed, encompassing planning to operation phases and including social, technical and economic considerations.

The last step of risk management is risk control. It incorporates the identification of the measures most suitable to manage risks, the definition of the performance objectives and requirements of the implementation methods, as well as the definition of the monitoring, evaluation criteria and review methods of the selected measures. For each one of the selected risk treatment measures, residual risks should be estimated and resource allocation optimised.

8.6 Acknowledgement

This Chapter is based upon work from COST Action CA15127 (Resilient communication services protecting end-user applications from disaster-based failures – RECODIS) supported by COST (European Cooperation in Science and Technology).



RECODIS

Resilient communication services
protecting end-user applications
from disaster-based failures



References

- 1 Townsend, A. and Moss, M. (2005) *Telecommunications Infrastructure in Disasters: Preparing Cities for Crisis Communications*. The Center for Catastrophe Preparedness & Response, New York, 45 pp.
- 2 Bendimerad, F. (2013) *Earthquake Risk Considerations of Mobile Communication Systems*. Earthquakes and Megacities Initiative Location, Diliman, Philippines.
- 3 Travanca, R., Varum, H. and Vila Real, P. (2013) The past 20 years of telecommunication structures in Portugal. *Engineering Structures*, 48, 472–485.
- 4 Hawkins, D. (2010) Discussion of current issues related to steel telecommunications monopole structures. *Proceedings of the Structure Congress*, pp. 2417–2438.
- 5 Stottrup-Andersen, U. (2014) Masts and towers. *Journal of the International Associations of Shell Spatial Structures*, 55(2), 69–78.
- 6 Smith, B. (2007) *Communication Structures*. Thomas Telford, London.
- 7 Marques de Souza, J. (2005) Stability analysis of VIVO RC poles (in Portuguese). *Proceedings of the Brazilian Congress on Bridge Structures*. Rio de Janeiro, Brazil, 10 pp.
- 8 Carril, C., Isyumov, N. and Brasil, R. (2003) Experimental study of the wind forces on rectangular latticed communication towers with antennas. *Journal of Wind Eng. Ind. Aerodyn.*, 91(8), 1007–1022.
- 9 Filipe, J., Travanca, R., Pipa, M. and Baptista, A. (2014) Monopoles for telecommunications – the influence of the equipment for wind action definition (in Portuguese). *Proceedings of the 5th Portuguese. Proceedings of the Conference of Structural Engineers*, Lisbon, Portugal, 16 pp.
- 10 WEF (2015) *Global Risks*, 10th edition. World Economic Forum, Geneva, 69 pp.
- 11 Munich Re (2015) *NatCatSERVICE. Loss Events Worldwide 1980–2014*. Munich Reinsurance Company, Munich, 9 pp.
- 12 UNISDR (2015) Making development sustainable: the future of disaster risk management. *Global Assessment Report on Disaster Risk Reduction*, United Nations Office for Disaster Risk Reduction, Geneva, 316 pp.
- 13 UNISDR (2012) *Impacts of Disasters since the 1992 Rio de Janeiro Earth Summit*. United Nations Office for Disaster Risk Reduction, Geneva, 1 pp.
- 14 Ferreira, J. and Branco, F. (2007) Structural application of GRC in telecommunication towers. *Construction Building Materials*, 21(1), 19–28.
- 15 Lanier, B., Schnerch, D. and Rizkalla, S. (2009) Behavior of steel monopoles strengthened with high-modulus CFRP materials. *Thin-Walled Structures*, 47(10), 1037–1047.
- 16 Beale, R. and André, J. (2017) Design codes and general design guidance. *Design Solutions and Innovations in Temporary Structures*. IGI Global, Pennsylvania.
- 17 BSI (1986) *BS 8100-1:1986 – Lattice Towers and Masts – Part 1: Code of practice for loading*, British Standards Institution (BSI), London.

- 18 BSI (1986) *BS 8100-2:1986 – Lattice Towers and Masts – Part 2: Guide to the background and use of Part 1: Code of practice for loading*, British Standards Institution (BSI), London.
- 19 BSI (1999) *BS 8100-3:1999 – Lattice Towers and Masts – Part 3: Code of practice for strength assessment of members of lattice towers and masts*, British Standards Institution (BSI), London.
- 20 BSI (1995) *BS 8100-4:1995 – Lattice Towers and Masts – Part 4: Code of practice for loading of guyed masts*, British Standards Institution (BSI), London.
- 21 BSI (2006) *BS EN 1993–3–1:2006 – Design of Steel Structures – Part 3–1: Towers, masts and chimneys – Towers and masts*, British Standards Institution (BSI), London.
- 22 TIA (2016) *TIA-222-G – Structural Standard for Antenna Supporting Structures and Antennas*. Telecommunications Industry Association (TIA), Arlington, USA.
- 23 ASCE (2011) *ASCE/SEI 48–11 – Design of Steel Transmission Pole Structures*, American Society of Civil Engineers (ASCE), USA.
- 24 ASCE (2015) *ASCE/SEI 10–15 – Design of Latticed Steel Transmission Structures*, American Society of Civil Engineers (ASCE), USA.
- 25 European Commission (2013) *An EU Strategy on Adaptation to Climate Change. Adapting Infrastructure to Climate Change*, 37 pp.
- 26 The Royal Academy of Engineering (2011) *Infrastructure, Engineering and Climate Change Adaptation – Ensuring Services in an Uncertain Future*, 108 pp.
- 27 Altvater, S., McCallum, S., Prutsch, A. et al. (2011) Recommendations on priority measures for EU policy mainstreaming on adaptation – *Task 2 report*, 49 pp.
- 28 CEA (2009) Tackling climate change. *The Vital Contribution of Insurers*, 64 pp.
- 29 Antunes, P., Travanca, R., Varum, H. and André, P. (2012) Dynamic monitoring and numerical modelling of communication towers with FBG based accelerometers. *Journal of Constr. Steel Res.*, 74, 58–62.
- 30 Turkstra, C., and Madsen, H. (1980) Load combinations in codified structural design. *Journal of Struct. Div.*, 106(12), 2527–2543.
- 31 BSI (2005) *BS EN 1990:2002+A1:2005 – Basis of Structural Design*, British Standards Institution (BSI), London.
- 32 BSI (2009) *UK National Annex to BS EN 1990:2002+A1:2005*, British Standards Institution (BSI), London.
- 33 BSI (2010) *UK National Annex to BS EN 1993–3–1:2006*, British Standards Institution (BSI), London.
- 34 ISO (2015) *ISO 2394: 2015 – General Principles on Reliability for Structures*, International Organization for Standardization (ISO), Geneva.
- 35 Beale, R., and André, J. (2017) Actions. In: *Design Solutions and Innovations in Temporary Structures*. IGI Global, Pennsylvania.
- 36 BSI (2010) *BS EN 1991–1–4:2005+A1:2010 – Actions on Structures – General Actions – Part 1–4: Wind actions*. British Standards Institution (BSI), London, UK.
- 37 BSI (2011) *UK National Annex to BS EN 1991–1–4:2005+A1:2010*, British Standards Institution (BSI), London.
- 38 Stathopoulos, T. and Baniotopoulos, C. (2007) *Wind Effects on Buildings and Design of Wind-sensitive Structures*, Springer, Vienna.
- 39 Holmes, J. (2015) *Wind Loading of Structures*. CRC Press, Florida.
- 40 Simiu, E. and Scanlan, R. (1996) *Wind Effects on Structures: Fundamentals and Applications to Design*, John Wiley & Sons, New York.

- 41 ASCE (2010) *ASCE 7–10 – Minimum Design Loads for Buildings and Other Structures*, American Society of Civil Engineers (ASCE), USA.
- 42 SAA (2011) *AS 1170.2:2011 – Structural Design Actions – Part 2: Wind actions*, Standards Association of Australia (SAA), Australia.
- 43 Beale, R. and André, J. (2017) Structural analysis. In: *Design Solutions and Innovations in Temporary Structures*, IGI Global, Pennsylvania.
- 44 BSI (2005) *BS EN 1993–1–1:2005 – Design of Steel Structures – Part 1–1: General rules and rules for buildings*, British Standards Institution (BSI), London.
- 45 BSI (2008) *UK National Annex to BS EN 1993–1–1:2005*, British Standards Institution (BSI), London.
- 46 BSI (2007) *BS EN 1993–1–6:2007 – Design of Steel Structures – Part 1–6: Strength and stability of shell structures*, British Standards Institution (BSI), London.
- 47 BSI (2008) *UK National Annex to BS EN 1993–1–8:2005*, British Standards Institution (BSI), London.
- 48 BSI (2005) *BS EN 1993–1–8:2005 – Design of Steel Structures – Part 1–8: Design of joints*, British Standards Institution (BSI), London.
- 49 BSI (2005) *BS EN 1993–1–9:2005 – Design of Steel Structures – Part 1–9: Fatigue*, British Standards Institution (BSI), London.
- 50 BSI (2008) *UK National Annex to BS EN 1993–1–9:2005*, British Standards Institution (BSI), London.
- 51 HERA (2001) *DCB 65, 66, 67, 78*. Heavy Engineering Research Association, Auckland, New Zealand.
- 52 André, J. and Pipa, M. (2011) *Technical Specifications and Provisions for the Design of Telecommunication Towers* (in Portuguese). Laboratório Nacional de Engenharia Civil (LNEC), Lisbon, Portugal, 75 pp.
- 53 ACI (2011) *ACI 355.3R–11 – Guide for Design of Anchorage to Concrete: Examples using ACI 318 Appendix D*, American Concrete Institute (ACI), Washington, DC.
- 54 ACI (2014) *ACI 318–14 – Building Code Requirements for Structural Concrete and Commentary*, American Concrete Institute (ACI), Washington DC.
- 55 CEN (2009) *CEN/TS 1992–4:2009 – Design of Fastenings for Use in Concrete – Parts 1–5*, European Committee for Standardization (CEN), Brussels.
- 56 EOTA (2013) *ETAG 001: Metal Anchors for Use in Concrete – Parts 1–6*, European Organisation for Technical Assessment (EOTA), Brussels.
- 57 EOTA (2013) *ETAG 029: Metal injection Anchors for Use in Masonry*, European Organisation for Technical Assessment (EOTA), Brussels
- 58 BSI (2013) *BS EN 1997–1:2004 + A1:2013 – Geotechnical Design – Part 1: General rules*, British Standards Institution (BSI), London.
- 59 BSI (2014) *UK National Annex to BS EN 1997–1:2004 + A1:2013*, British Standards Institution (BSI), London.
- 60 Repetto, M., and Solari, G. (2010) Wind-induced Fatigue Collapse of Real Slender Structures. *Engineering Structures*, 32(12), 3888–3898.
- 61 Auerswald, P., Branscomb, L., La Porte, T. and Michel-Kerjan, E. (2005) The challenge of protecting critical infrastructure. *Issues Sci. Technol.*, 22(1).
- 62 André, J., Beale, R. and Baptista, A. (2015) New indices of structural robustness and structural fragility. *Struct. Eng. Mech.*, 56(6), 1063–1093.

9

Customer Edge Switching: A Security Framework for 5G

Hammad Kabir, Raimo Kantola, and Jesus Llorente Santos

Department of Communication and Networking, Aalto University, Finland

9.1 Introduction

5G is the next phase in the evolution of mobile networks. The recent ITU-T statistics [1] reveal that active mobile broadband subscriptions are on the rise, and have overtaken both fixed broadband subscriptions and individuals connected to the Internet. This segment is rapidly growing and expectedly the number of Internet-connected wireless devices will rise. The initial estimates for 5G are to handle 100 devices per inhabitant of the world, and that data traffic will grow 1000-fold from the years 2010 to 2020. However, besides faster mobile broadband and increased capacity, 5G will have to support new trust models, service delivery models, use cases and withstand challenges from an evolved threat landscape [2].

The divergent requirements of different use cases will require the ability to tailor the network to meet the particular use-case requirements, for example with respect to reliability, services availability and level of security. For this, the concept of *network slicing* is proposed. A network slice is a set of resources and communication nodes handling a significant user or device segment with intended coverage area. Compared to provisioning such communication through regular mobile broadband access, in the slice, the network control and data planes can be enhanced with additional (virtualized) network functions and the resources can be dimensioned in the way that meets the requirements of the use case. To take an example of the potential of slicing, laws on privacy protection apply to communication when at least one human is involved. Contrary to humans, machines do not have human rights, so if needed, the network can extensively monitor machine-to-machine communication for security reasons. It seems that allowing such monitoring in the same network (or network slice) where humans also communicate, would be a legal nightmare; however, when suitable isolation is achieved by network slicing, arranging and controlling such monitoring would become much easier.

Besides others, the threat landscape would evolve due to new valuable services being carried over the network, new service delivery models and a multitude of end-devices

connecting to the Internet. For example, the advent of Industrial Internet (II) and Internet of Things (IoT) would connect many previously unconnected devices as well as completely new ones to the Internet, increasing the attack surface. The ubiquitous reliance of users on smart mobile devices and in particular 5G, makes end devices more lucrative for hackers. Very often these end systems run poorly developed software and have unpatched vulnerabilities that put the system and its network at risk [3]. As a result, networks suffer from persistent hacking attempts to the end systems, service compromises, fraud, theft of information and DoS. Mobile networks already today rate *availability* as their top concern [4]. The evolved threat landscape of 5G will further stress this concern, and challenge the thriving potential of new services that would rely on 5G for ubiquitous connectivity. For example, 5G aims to support massive machine-to-machine communications and ultra-high reliability, such as in life-critical automotive applications. There is a concern that 5G will put more value (e.g. human life) at stake, and that hackers can compromise more value than before, if efforts to better security are not made.

Besides the evolution of a threat landscape, the Internet has yet to address challenges from some of the classical Internet weaknesses, such as address spoofing, unwanted traffic, network scans, traffic floods and DoS. We recognize that much of the traditional Internet weaknesses are the result of any-to-any communication paradigm in the Internet, which allows any host in the Internet to send flows to another Internet host. Hackers often abuse this paradigm to target their victims, and leverage poor identification of hosts, possibility of spoofing and the lack of authentication mechanisms in the Internet to their advantage, and as a result invalidate the network auditing. Consequently, the volume of malicious activities, such as network scans, DoS and unwanted traffic is on the rise [3], which often result in network outages, computing downtime and waste of resources. Often the hackers use these classical Internet weaknesses as a launch pad for more advanced attacks.

From the security perspective, it is pertinent that 5G addresses the challenges from the classical Internet weaknesses, as well as deploys mechanisms for better tackling the evolving threat landscape. In particular, 5G must ensure better than state-of-the-art security to achieve its goals of provisioning the ubiquitous access and ultra-reliable services. 5G security can be broadly categorized into three categories:

- 1) Access security (subscriber-operator relation);
- 2) SDN-style virtualized-core security (function-to-function relation); and
- 3) End-system security.

This chapter in particular is concerned with the latter, using various network-based methods. The chapter introduces a security framework called Customer Edge Switching (CES) (Figure 9.1), to address the challenges from the classical Internet weaknesses and to contribute towards better handling of the emerging threat landscape. It promotes:

- a) policy-based communication, whereby the interests of the receiver are met with the interests of the sender. This is unlike the traditional best-effort Internet, where any



Figure 9.1 Customer Edge Switching.

host can send packets to another Internet host, and often the interest of the receiver differs from the interest of the sender by an amount called *unwanted* traffic. In addition, CES overcomes the classical weaknesses of the Internet, namely source address spoofing and DoS prior to admitting a flow.

- b) CES (can also act) as a cooperative firewall, which in addition to attack detection can also attribute the misbehaviour to the sender network and filter the malicious (non-cooperative) hosts and their activities due to the cooperation of networks. In this context, CES is essentially an extension of the classical firewall functionality into a cooperative firewall, such that in addition to the typical accept or drop decision on a packet, CES can issue additional queries to the packet sender to eliminate spoofing and assert identities prior to making a final decision. In particular, the elimination of spoofing enables attributing the misbehaviour evidence back to the identity of the sender, and lays the foundation for
- c) sharing and aggregating evidences across the Internet, for example via an Internet-wide trust management system, that is, to limit the scope of attack strategies and to better tackle the evolving Internet threats. However, we generally consider the latter beyond the scope for this chapter.

The particular advantages of our approach are that it can be deployed one network at a time; the costs of deployment are well aligned with the benefits and the system suits the needs of mobile and wireless devices. The deployment of CES happens at network edges, where it replaces NAT. The adoption of CES does not require any changes in the end-hosts, since CES limits all the changes to the edge network, thus minimizing the deployment challenge. To fulfil the need for incremental deployment, CES supports the Realm Gateway (RGW) function, which allows communication between the legacy Internet and hosts in the private network, behind CES. In addition, it overcomes the drawbacks of the classical NAT traversal solutions that do not scale well to the battery-powered mobile devices.

The rest of this chapter is structured as follows: Section 9.2 describes the state-of-the-art in mobile network security; Section 9.3 briefly describes the CES framework and implemented security mechanisms; Section 9.4 presents an evaluation of the security mechanisms; Section 9.5 considers the deployment aspects and presents different use cases for CES deployment; and finally Section 9.6 concludes.

9.2 State-of-the-art in Mobile Networks Security

Mobile networks are continuously evolving and are becoming platforms for various services. The upgrade to 5G mobile networks and expected support for new technological evolutions, such as IoT and Industrial Internet, requires careful consideration of mobile network security. This section discusses the state-of-the-art in mobile network security, and how it may evolve to support 5G and its requirements.

Figure 9.2 presents the state-of-the-art in current mobile networks, which connect via the Gi/SGi interface to external packet-data networks, such as the public Internet or other corporate networks. Clearly, the Gi/SGi interface of the mobile networks is susceptible to attacks from the Internet and external packet data networks, such as customer networks connected to the PGW (for LTE/4G networks) or GGSN (for 3G networks). This section presents some of the most common threats to the Gi/SGi

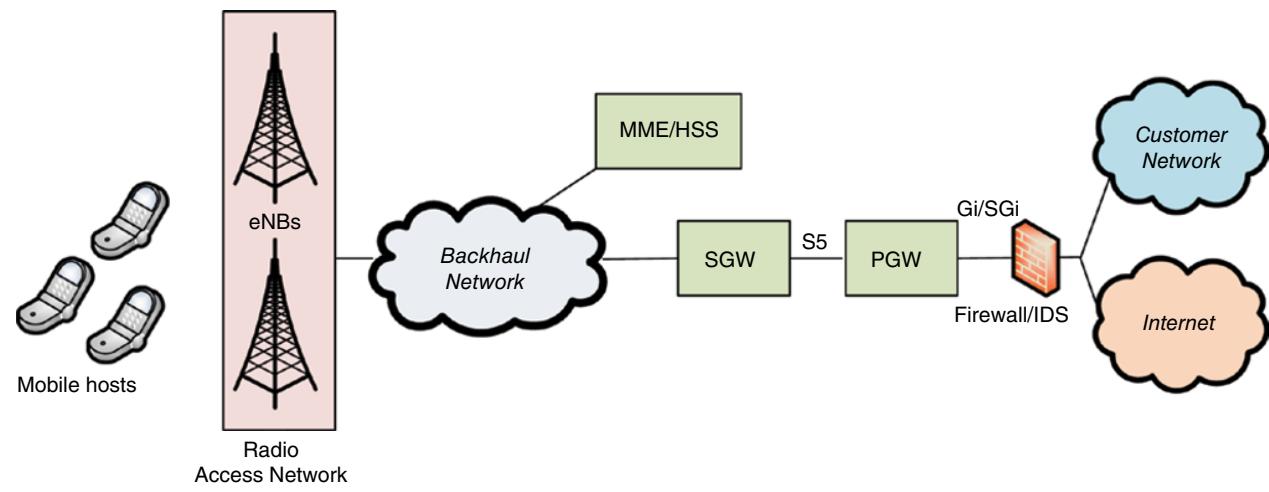


Figure 9.2 Mobile network architecture.

interface, for example: address spoofing, traffic floods, unwanted traffic, viruses, network scans, botnets and DoS, and their mitigations by the mobile network operators. If these ills of today's Internet are not properly addressed, network operators may incur huge losses in terms of customer loyalty and trust, due to frequent end-system and network compromises and service disruptions. Network operators typically employ Carrier Grade NAT (CGNAT), stateful firewalls, IPSec, intrusion detection system (IDS) and Intrusion Prevention Systems (IPS) to mitigate these attacks [4].

For example, hackers may send TCP SYNs from spoofed source addresses as well as from the botnets to one or more publicly reachable addresses of the network (or served devices). Similarly, by sending spoofed DNS queries, hackers can use the responses from DNS servers (or mobile infrastructure) to launch reflection and amplification attacks. Besides hogging the capacity of Gi/SGi interface, a response from one of the public addresses allows hacker to learn of an open port, and thus subject the corresponding service or connection state to attack. Networks typically mitigate these attacks by employing an integrated stateful inspection firewall and IPS engine, which either respond by rate-limiting or resetting such connections. For all the *attack* traffic that bypasses the network filters or firewalls, it consumes the network capacity, clogs up the precious radio spectrum, and affects the battery consumption of the wireless device by disturbing its sleep cycle. By keeping the radio spectrum in continuous use, the hacker denies the use of network resources for legitimate users and needs.

In addition to bandwidth saturation attacks, hackers also subject the security infrastructure, for example Gi/SGi firewall and IPS to attacks. Firewall and IPS are stateful inline devices and thus are inherently vulnerable to DDoS, which can overwhelm the state capacity of these systems. Networks typically handle the bandwidth attacks by enforcing bandwidth policing, which limits each class of traffic in the admitted bandwidth. This ensures that each traffic type gets required resources to ensure a minimal level of quality of service (QoS) at all times. Moreover, some traffic types, such as IPSec, can be prioritized over others, to ensure the promised QoS to customer networks. Similar rate limits on the inbound traffic prevent DoS on the security systems, such as Firewalls and IDS/IPS.

Besides exposure to Internet malpractices, mobile networks also face challenges from connected mobile devices. In particular, hackers can compromise a weak device, a vulnerable service on the end device, or may trick a careless user into becoming part of a botnet. These infected devices are then used to launch attacks on other devices or the mobile network itself. Mobile operators employ multi-stage detection methods to filter botnet traffic. This typically involves combining and triangulating the information from searching attack signatures, the application context and limiting the connection rates. In general, the mobile networks employ Carrier Grade NAT (CGNAT), Deep packet Inspection (DPI), proxy agents, web-content filtering, anti-spam filters, and application-specific filters to secure the network and their hosts against Internet attacks and compromises.

Network Address Translation (NAT) and NAT traversal are of particular interest in the state-of-the art. The communication from host in the private network to a public host in Internet is trivially based on the NAT, which follows the pattern of communication and establishes a NAT binding to admit the responses. In the opposite direction, the traversal method recommended by IETF is based on self-address fixing, that is, the server behind a NAT (e.g. an App that can receive calls) learns the NAT outbound

address assigned by the NAT by exchanging messages with a server in the Internet and keeps that binding alive by keep-alive signalling. Once the address is learned, it is passed to the remote end on the application layer. The initial learning requires several tens of messages, easily consuming tens of seconds and the learning algorithm needs to be repeated from time to time. The approach makes NAT traversal an Application level issue, which, in addition to consuming the device battery, deteriorates the architecture.

To summarize, the current state-of-the-art in mobile networks and end-system security mostly follows the reactive approach. The prevalent culture in security is such that Internet entities (i.e. networks or hosts) are responsible for their own security. However, in protecting the air interface against bandwidth saturation, mobile networks to a certain extent also filter the attacks directed to the end systems. To that end, mobile networks are more advanced than fixed broadband networks, which generally leave the end-user host exposed to the Internet. For end systems, some security vendors are already today leveraging the tools of sharing vulnerability and exposure information. Using the shared information, each vendor can update the counter-measures in its security solution; leaving it up to the users of the software to install the security updates. According to a security specialist [5], the majority of the security breaches in end systems are due to known vulnerabilities in the software that is not updated in time.

9.2.1 Mobile Network Challenges and Principles of Security Framework

Besides the need to better handle the classical Internet threats, the future mobile networks also have to tackle the challenges from introduction of new technologies, such as Software Defined Networking (SDN) and Network Function Virtualization (NFV), which are critical to the realization of 5G. For example, the adoption of SDN enables pursuing a split of control and data plane in the mobile networks, such that the network intelligence is gathered into a set of control nodes that actively monitor traffic, generate flow rules, and due to global visibility of the network can diagnose threats and mitigate the security challenges. The data plane consists of basic forwarding nodes that enforce the rules generated by the controller. While SDN brings the desired level of flexibility, it also presents the controller as a single point of failure for mobile networks if the Internet-borne attacks are not mitigated. The paper in [6] identifies some of the challenges raised by the separation of planes and aggregation of the control functionality into centralized nodes. We argue that in a broad context, 5G security can be categorized into: i) access-network security; ii) virtualized-core security; and iii) end-system security. The CES framework is concerned with the end-system security and mitigation of the Internet-borne threats on (Gi/SGi interface of the) mobile networks.

Clearly, the existing and emerging Internet threats assert that mobile networks do better than the state-of-the-art to tackle the security challenges. In this context, we present some of the learnings from Internet security and ambitions for 5G in terms of security principles, and argue that future mobile networks shall:

- limit the flow acceptance to verifiable sources, to tackle the problem of source address spoofing and traffic floods, and hence prevent the resource exhaustion;

- eliminate source address spoofing, to establish grounds for attributing misbehavior evidence to the identity of the sender or the network serving it;
- make it possible to aggregate misbehavior evidences under a stable source identity:
 - Such aggregation can contribute towards building and employing the reputation of the Internet sources/networks, and demote the networks that do not take corrective actions and hence keep forwarding the malicious traffic;
 - The aggregation of misbehavior evidences can also present an overview of the developing threats, and thus contribute to better preparedness against attacks, by generating indicators for trusted network monitoring.
- under network stress, grant resources based on the source reputation;
- meet the interests of the sender with the interests of the receiver, to tackle the problem of unwanted traffic;
- allow defining dynamic (reachability) policies for hosts, applications and services. The management and control of the policies could be in the cloud while enforcement takes place in standard data-plane nodes, as well as control nodes on trust boundaries. This is in contrast to the current mobile networks where policies are tightly coupled to the physical resources and are not scalable to services/applications;
- deploy a security solution that does not require changes in the end-hosts or protocols, and limits all the changes to edge nodes, to minimize the adoption cost.

In addition, the mobile networks can leverage the current state-of-the-art and future adoptions of technology to strengthen their security. For example, the future mobile networks shall:

- leverage the mechanisms in state-of-the-art security to harden defence against known attacks, in particular, to protect the controller from the Internet-borne attacks;
- present a multi-tier approach to end-system security, where at least the mass-attacks are handled by a user's agent in the network, such as a cooperative firewall; while the application-level security can either stay in the end devices or can be also delegated to a set of cloud-based security entities, such as Web Application Firewalls or virus detection environments;
- analyse and manage the policy configuration of the data-plane elements, to deploy a robust and uniform security policy across the network. The logically centralized controller can provide a global view of different network device configurations and hence mitigate the conflicts and inconsistencies in the network security procedures.

In summary, these principles address the classical Internet weaknesses, such as spoofing, unwanted traffic, malicious flows and DoS. The ability to attribute misbehaviour in particular allows identifying the malicious sources and contributes to filtering the malicious traffic close to the sender, due to the cooperation of networks. Such cooperation of networks is possible due to aggregation of the misbehaviour attributed to a source, under a trusted entity such as a trust management system, and reflecting the corresponding sender reputation in the network admission decisions. We argue that the goal shall be to locate the malicious sources and early filtering of attacks, thus making defection or hacking a less favourable/practiced strategy in the Internet [7]. This is unlike the traditional security, which only drops the malicious traffic after it has reached the destination or its network.

9.2.2 Trust Domains and Trust Processing

The challenges of using trust and reputation concepts in mobile networks for improving security include:

- 1) host generated evidence of misbehaviour cannot be trusted;
- 2) due to laws on communication privacy, ISPs cannot monitor the end-user traffic unless the whole network is under attack; and
- 3) corporations could be reluctant to share misbehaviour evidence, as it could damage their business reputation and also reveal their weaknesses to even more attackers.

We have studied the ways to tackle the first challenge, for example in [8] and challenges 2 and 3 in [9].

Under the concepts of trust and reputation, to achieve a significantly higher level of security in 5G (and in software-defined sliced-network connected systems) compared to the state-of-the-art in the end systems, we propose a new security principle “all-for-one and one-for-all”. This principle suggests that the good guys join forces and cooperate against the brotherhood of hackers. This can be implemented by deploying elaborate methods of cooperation between networks and among end systems, for processing of security attacks and incident information, mapping and automatically containing the resources used for the attack as soon as detected.

This principle can be implemented by first introducing the concept of a *trust domain*. A trust domain is a set of network entities and administrations that agree to share the security incident information; follows some common rules in the area of security and contains attacks with joint efforts. The security incident information is collected ubiquitously by all the end systems and network functions, aggregated in a secure manner and validated by network-based monitoring in the admitted traffic [8]. The results of the aggregation are distributed to customer network gateways that act as cooperative firewalls in the form of black lists, penalty lists, grey lists and white lists of the remote (network) entities.

A trust domain can also be an operator’s network together with the end systems and network entities serving a critical sector of economics. To improve defence against Internet attacks, the larger is a trust domain, the better. Therefore, several administrations may decide to join and form multi-administration trust domains based on a trust alliance agreement. However, in state-of-the-art, corporations do not prefer sharing of the security incident information due to attached privacy concerns and fear of a bad reputation, which may damage the public image of the firm. Nevertheless, it is possible to set up a trust domain by combining operations of most networks under one administration, or one ISP, or by the use of regulation.

To encourage incident reporting within a trust domain, we propose the use of homomorphic security [9] to encrypt the incident information from the customer end and aggregating the evidence in an encrypted form at the serving eye-ball ISP. The ISP will anonymize the reports and send them to the cloud-based global trust operator (GTO) for final processing. The GTO will be able to decrypt the reports, because the end systems use the GTO’s public key for encryption, but the GTO will see only the encrypted identity of the reporting system. This allows keeping the customer-to-ISP relation a business secret between the two parties. When GTO detects sufficient evidence of

malicious activity, it will authorize the serving ISP to monitor the offending (remote) system. Once the ISP, using network monitoring, observes conclusive evidence of malicious activity, it will report back to the GTO that will blacklist the infected system. This kind of multi-stage detection of malicious activity is justified by several factors:

- 1) it does not solely rely on the end systems, since they cannot be trusted to report accurately; and
- 2) ISPs cannot simply monitor human-to-human traffic at will, because of communication privacy laws, unless the whole network or its customers are under attack.

Under the premise of a global GTO, the blacklists are distributed within the trust domain to all cooperative firewalls and IPS. The cooperative firewall hosting the infected system will put the device into a sandbox where it can do no harm to other systems, and gets cleaned of viruses, Trojans and installs software upgrades to patch the vulnerabilities.

Trust processing must be accurate, efficient and robust [8], such that the probability of false positives is (nearly) zero. Efficiency means that malicious activity is detected quickly, while robustness comes from the fact that it is resistant to system attacks that try to submit false evidences. The way to achieve robustness is that trust processing at both the ISP and GTO maintains the reporting credibility of the reporting entity. This value gives a likelihood that the entity is reporting honestly and in a timely manner. We studied the question of robustness, for example in [8].

The starting point for all the evidence collection is that a suspected entity must be identified reliably. Evidence that does not point to a reliable identity is not actionable and therefore not very useful. Since IP addresses can be spoofed and most end systems either use a dynamic IP address or are NATted, an alternate and more reliable identity is needed. In our proposed system, we make use of different types of end-system identifiers, such as Fully Qualified Domain Names (FQDN) for end-system identification. It is a requirement that cooperative firewalls maintain reliable and traceable identification of all the communicating entities.

9.3 CES Security Framework

CES is a proposed replacement of Network Address Translators (NAT) at the network edges. NAT effectively hides the private network from the Internet, and connects the hosts located in its network to the Internet via a translation of public and private IP addresses. By default, NAT allows outbound connections to the Internet and creates a flow state to admit the respective inbound flows in the private network. However, the Internet-originated inbound connections towards private hosts are typically dropped, due to absence of a prior state in NAT for flow admission. As a consequence, inbound connections to private hosts are accepted following the officially recommended NAT-Traversal methods [10], which are cumbersome and do not scale well to the battery-powered mobile hosts.

In contrast, the deployment of CES separates the customer network from the public Internet, such that CES:

- 1) acts as NAT for outbound connection to legacy Internet hosts; and

- 2) supports Realm Gateway (RGW) to handle inbound connections towards private hosts from the legacy Internet. RGW offers a better-than classical NAT traversal solution that scales well to mobile devices.
- 3) can act as a cooperative firewall that negotiates the communication policies of its host with the policies of the destination located in another CES network.

Both the NAT and RGW functions aim to provide the backwards compatibility and interconnection of CES with the legacy Internet, whereas CES-to-CES connectivity allows secure interconnection between two customer networks. We briefly present the comparison of CES functionality with NATs in Figure 9.3.

In the context of security, CES can be seen as an extension of the traditional stateful firewall functionality into a cooperative firewall. Compared to the classical stateful firewalls, which either accept or drop an inbound packet, CES can issue additional queries to the source (network) of the packet, for example, for the purpose of eliminating spoofing, authenticating the sender and establishing the base-level policy compliance. The queries can also be issued to other Internet entities, such as Certificate Authorities (CA). This establishes the basic trust at the network level, and provides grounds for attributing the evidence of misbehaviour to the source address (or to the source network). Thus, if a traditional firewall mostly executes the destination network rules, a CES firewall also provides efficient methods related to the source addresses. A CES node can also use the black, grey and white lists maintained by its trust domain. The reaction to a particular flow can be controlled by a policy at CES and host levels. The policies can be dynamic and will react to the type and level of malicious activity.

The establishment of trust at the network-level allows hosts in the respective networks to communicate following a negotiation of policies. A CES node acts as a connection broker for the hosts/applications that it serves and exchanges their policy offers and requirements with the remote hosts using Customer Edge Traversal Protocol (CETP). For two hosts connected to their respective CES node, the negotiation of CETP policies must succeed prior to the relaying of user data between the hosts. The scope of policy negotiation is limited to CES only and is a must to establish the data connection between two hosts communicating across their respective CES nodes. Subsequently,

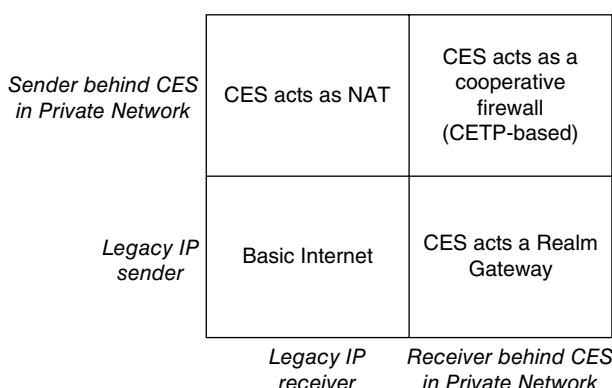


Figure 9.3 CES Functionality comparison with NAT.

the user-generated data related to a host-to-host session is tunneled across networks using session tags agreed in the policy negotiation.

Under the premise of trust between networks, when a misbehaviour (or attack) is observed in a host session, the misbehaving host could be reported to the remote network and cooperative firewalling will lead to its filtering close to the source network. Naturally, the effectiveness of cooperative firewalling to filter the attacks will increase as more and more networks adopt the CES firewall. If the remote CES node does not cooperate, the CES node of the victim can send evidence to trust processing at GTO, which can blacklist the remote CES everywhere for a time.

A possible result of deploying a CES node as gateway to the rest of the network blurs the boundary between closed private networking and open style Internet networking. In addition, for the case of closed and secure networking, all the logic resides at network edge (additional logic may be executed at the end host), while the network core does not need to be burdened with numerous virtual network routing tables as in the case of BGP supported VPNs.

9.3.1 DNS to Initiate Communication

In an Internet where default gateways will be replaced largely by CES nodes, most hosts will have just a private address, and the legacy interworking methods such as NATs and RGW will grant connectivity to and from legacy hosts. The use of private addressing, which could be assigned dynamically by DHCP servers, requires that a host also has a rather stable globally unique identity. It is natural to resort to the use of Fully Qualified Domain Names (FQDN) for identifying hosts, as well as the services running on the hosts.

As a result, the CES architecture is tightly coupled with the use of Domain Name System (DNS), such that hosts initiating the communication are required to issue DNS queries for FQDN of the destination hosts. The use of DNS makes it possible for CES to decouple the endpoint FQDN identifier from the advertised routing locator for destination. Consequently, the CES system maintains a delegated DNS zone of authority with the records of the served hosts, that is, to respond to the DNS queries for the served domains.

The DNS query for a host FQDN also triggers a phase of CETP service discovery, which precedes the CES-to-CES policy negotiation. The service discovery process is triggered by DNS query from an outbound CES (oCES) node, and determines whether the destination is behind another CES and the proper course of action to reach it. The procedure relies on the standard Naming Authority Pointer (NAPTR) [11] DNS queries to ascertain the CETP service on the remote edge. An example of a valid NAPTR response offered by an inbound CES (iCES) node is as follows:

```
b.nwb.ces. 40 IN NAPTR 20 6 "U" "CETP+cesid"
"!^(.*)$!cesid:1=nwb.ces.?ip=182.3.2.55?alias=GRX!"
```

The value *CETP + cesid* indicates the availability of the CETP service. The NAPTR response carries the publicly reachable address called Routing Locator (RLOC) for connecting to the remote CES, in this example indicated by “ip = 182.3.2.55”. The response can also carry other RLOC types, such as IPv6, and it can be used in conjunction with fields, such as port numbers for identifying the exact location of the CETP service in the remote network. The CETP service can be available in future on top of other protocols, such as TLS/TCP, HIP, etc. The *alias* field allows the use of private transit links instead of the public links. A NAPTR response may also contain a CES identifier in “cesid:1 = cesb. ces” for routability checks.

The NAPTR records could be hosted by third-party DNS servers or cached by intermediate DNS proxies. However, if the CES node has the authority over the DNS leaf, that is, the network that it serves, this would improve the CES functionality and administration. For hosts, the decoupling of endpoint identifiers from routing locators means that an IP address can no longer be used to identify a host. Instead, CES leverages the concept of proxy-address to represent the remote hosts in its local network. These proxy-addresses are allocated from the IPv4 private address space [12] or the IPv6 Unique Local Address (ULA) [13], depending on the addressing of the host.

9.3.2 CETP Policy-based Communication

The CETP protocol has been developed to convey signalling and data across CES nodes. The scope of signalling is exclusively CES related and is used to exchange:

- a) network-level CES policies for establishing trust between CES networks; and
- b) individual host or application policies, for establishing data connection between two hosts across their CES nodes.

The information exchanged in the signalling depends on these policies. The use of CES in mobile networks allows the policies that can scale up to individual hosts, services or applications and are not tightly coupled to the physical resources.

The data plane only carries the user-generated data, which is tunnelled using CETP session identifiers. The session identifiers are negotiated during the CETP signalling phase and are used to distinguish between different user connections on the data plane. Moreover, the negotiation of data RLOCs during signalling makes it possible to carry user data on different links than the signalling, to achieve a higher level of assurance and improved heuristics of the algorithms.

A CETP packet starts with a mandatory 4-byte header that identifies the protocol version and CETP header length. This is followed by source and destination session tags (or identifiers), and optional signalling or payload information. CETP signalling carries the host or network policies encoded in a flexible Type-Length-Value (TLVs) format. These TLV-encoded policy elements are classified according to the Type field, which can be further decomposed into *operation*, *group* and *code* fields. A list of currently supported TLV elements is classified as per the group and code field and is presented in Table 9.1.

Table 9.1 CETP policy elements organized into groups.

| Group | Code | Description |
|---------|-----------------|------------------------------------------|
| CES | Pow | The proof-of-work computation |
| | cesid | FQDN-based ID of the CES node |
| | headersignature | Signature of the CETP packet |
| | caces | The CA address for CES validation |
| Control | Dstep | FQDN-based destination endpoint ID |
| | caep | The CA address for endpoint validation |
| | terminate | Contains session terminating information |
| | warning | Contains the warning information |
| | ack | The acknowledgement number |
| | ttl | The time to live for session |
| ID | fqdn | FQDN-based ID of the sender |
| | maid | The Mobile Assured ID |
| | moc | The Mobile Operator Certificate |
| | msisdn | The MSISDN number of the host |
| RLOC | ipv4 | An IPv4 address (RLOC) of the CES |
| | ipv6 | An IPv6 address (RLOC) of the CES |
| | eth | An MAC address (RLOC) of the CES |
| Payload | ipv4 | IPv4 encapsulation of the user payload |
| | ipv6 | IPv6 encapsulation of the user payload |
| | eth | Eth encapsulation of the user payload |

A policy is defined by three distinct sets: *offer*, *requirement* and *available*. Each of these sets carries the policy elements, which collectively define a policy. The operation field in the TLV encoding can take following values:

- a) *info* to indicate the value of an offered policy element;
- b) *query* to indicate the request for a policy element; and
- c) *response* is an answer to a previous query from CES and could be empty if the policy element is not available.

Outbound Policy for Host-A (a.cea):

```

Offer      = {id.fqdn, rloc.ipv4, rloc.ipv6, payload.ipv4}
Requirement = {id.fqdn, rloc.ipv4, rloc.ipv6, payload.ipv4}
Available   = {id.fqdn, rloc.ipv4, rloc.ipv6, payload.ipv4, id.hash}

```

Inbound Policy for Host-B (b.ceb):

```

Requirement = {id.fqdn, rloc.ipv4, payload.ipv4}
Available   = {id.fqdn, rloc.ipv4, payload.ipv4}
Offer       = An inbound policy does not make any offer

```

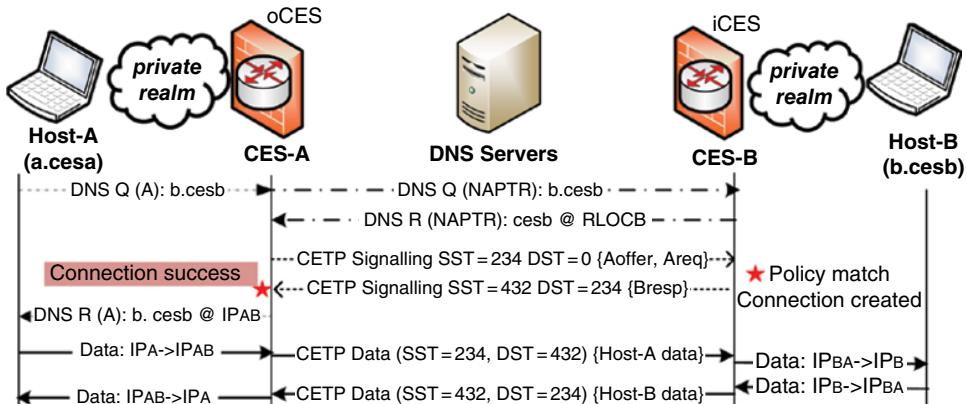


Figure 9.4 CETP signalling between two hosts in different CES networks.

Policy Negotiation

The process of establishing communication between two hosts across their CES networks is illustrated in Figure 9.4 and can be decomposed into four phases:

- 1) Host-A resolves the FQDN `b.cesb` via CES-A to communicate with Host-B. CES-A initiates the CETP discovery process with an NAPTR query for destination FQDN. The valid NAPTR response confirms the availability of CETP service at a remote network served by CES-B. The DNS response also provides the signalling RLOCs of the CES node serving the destination.
- 2) *CETP Signalling*: CES-A initiates CETP signalling (i.e. policy negotiation) towards signalling RLOCs of CES-B, by presenting the policy offers and requirements of Host-A (A_{offer}, A_{req}). The signalling to CES-B bears a source session tag (SST), while the destination session tag (DST) is empty. CES-B checks for local policy match, as to whether both hosts satisfy the policy requirements of each other. In the case of a local policy match, CES-B responds to the policy requirements A_{req} with B_{resp} and completes the negotiation by assigning the destination session tag (DST) in (SST = 432, DST = 234). CES-A validates the answer and also completes the negotiation. Both CES nodes allocate the proxy addresses to represent the remote hosts in their local network.
- 3) CES-A responds to the earlier DNS query of Host-A with a DNS response containing the allocated proxy address, IP_{AB} .
- 4) *User-Data*: Host-A and Host-B can then communicate through their respective proxy IP addresses. The CES nodes perform tunnelling of the user-data packets using the session tags, across the networks, on the data RLOCs negotiated during signalling.

In step-2, the inbound CES (iCES) node can either accept the policy offer as it is, or it can request the sender for additional policy requirements. The latter effectively postpones the connection establishment and allows the sender to make a better offer that satisfies the destination requirements and thus achieve a policy match. This also allows iCES to specify policies that assert identities, eliminate spoofing and ensure compliance. Depending on the evaluation of policies, the CETP negotiation can either result in success or failure, and this is typically achieved in 1–2 round trips. For attack resistance, the iCES node remains stateless until the session establishes in the last round trip of the CETP signalling.

In case a prior trust does not exist between the CES nodes, a CES-to-CES level negotiation of policies precedes the host-to-host policy negotiation in step-2. The aim of CES policies is to negotiate network-layer capabilities, and to achieve the trust between networks. This includes elimination of the classical internet weaknesses, such as source address spoofing, and hence establishing grounds for putting blame on the remote network or its hosts, in case a misbehaviour is detected or reported. We discuss these CES policies and corresponding security mechanisms briefly in Section 9.3.4.

9.3.3 Policy Architecture

All aspects of communication in CES are governed by policy. For example, CES-level policy defined by a network administrator is concerned with establishing trust between CES networks, besides guiding other aspects of CES-to-CES signalling, such as cooperative firewalling and reputation handling. Similarly, all the host, service or application-level communications are governed by the corresponding policies defined by the end users or its subscriptions within constraints agreed with the corporate network admin, the eyeball ISP or the mobile operator.

The policy creation can be based on business contracts or software licenses. Examples of such business contracts include subscription to ISP or mobile operator; purchase of mobile network packages and services; or outsourcing contracts between firms or between the firms and the consumers. In order to produce executable policies for a CES node, there is a need to leverage information in different data repositories, such as mobile application stores, trusted electronic contracting services, security vendor databases and company level security policy databases, at both the CES and host level. In addition, the CES nodes need to leverage the black, grey and white lists dynamically maintained by the trust domain.

To use policy architecture in the context of mobile networks, CES would, for example, use a Diameter client that can query HSS for mobile issued identities. The storage and management of policies for hosts and applications can be delegated to the current 3GPP architecture, where the controller can also configure and update PCRF with a set of its security policies. The policies can be retrieved by the CES node, cached and reused when needed. The principle of allowing a subscriber to modify its policy within a given contract or license will be that it can always make policies of its users, hosts or services more restrictive. For example, the subscriber could limit the access to a “Payroll” server only to a set of devices that can provide an identity defined in the policy.

The idea is that by executing the flow level policies in the control plane of a CES node, residing in the cloud, we can provide per-host or per-service firewalling. In the context of cellular networks, the security policy management can be seen as an extension of the 3GPP policy management architecture, which has so far focused on the Quality of Service policies.

9.3.4 CES Security Mechanisms

CES provides several policy-controlled mechanisms for establishing flow legitimacy, such as negotiation of a variety of ID types, return-routability checks, proof-of-work mechanisms and the use of secure locators. The mechanisms enable CES to identify the sender and its network on a required level of assurance. For this, CES eliminates

the classical Internet weaknesses of address spoofing, unwanted traffic and DoS prior to admitting a flow from the sender. CES provides the above mechanisms as policy-controlled features for network administrators, and gives different level of assurances based on the configured mechanisms. We briefly describe these mechanisms that minimize the risks to CES from Internet abuse:

- *Proof-of-Work*: is an effective method in many anti-spamming and anti-flooding proposals, since it makes it difficult for the sender to flood the victim with a large number of flows. The use of this mechanism in network-level policy negotiation pushes the burden of communication to the sender, since the *burden-of-computation to the sender >> burden-of-verification on the receiver*. In addition, the elimination of spoofing due to this mechanism contributes to sender identification together with the CES-Authentication mechanisms.
- *CES Authentication*: CES nodes can either use the CES-ID, which is FQDN of the served network, or their registered RLOCs for identification purposes. By principle, the CES node needs to authenticate the remote edge or its identity, prior to admitting a flow from it. After spoofing has been eliminated, which is a pre-requisite of authentication, a centralized *trusted* server can verify the authenticity of the CES-ID or its RLOCs. In more practical cases, CES nodes can leverage the trust on X.509 certificates issued by well-known Certificate Authority (CA) to authenticate CES-ID of the remote network. A remote edge can also be authenticated if its certificate is endorsed by one of the CES-IDs that the CES node trusts, similar to a Web-of-Trust (WoT) model.
- *Header Signature*: the CETP signalling is generally agnostic to the underlying technologies, that is, IPv4, IPv6, HIP, TLS/TCP and others. Some of these underlying protocols already have some level of certificate enforcement, for example in TLS. However, for all the other protocols, it is possible to provision the signed CETP header within the CETP signalling such that the receiver can verify it to authenticate CES node using the CA certificate for CES-ID.
- *Secure signalling*: although the CETP signalling is agnostic to the underlying protocols, the use of protocols such as TLS/TCP, HIP, and IPSec. can contribute to the security and reliability of the signalling channel between CES networks, due to underlying technology. Similarly, the CES nodes can choose to exchange signalling over private-links, which could be more secure and hence preferred compared to the public links between networks.
- *Policy-based Communication*: compared to the best-effort principle of the Internet, which favors the senders, the policy-based communication in CES allows meeting the interests of the receiver with interests of the sender, and hence filters the unwanted traffic. In addition, it asserts the identities, authenticates the hosts, and executes host-level firewalling to filter the unwanted traffic.

9.3.5 Realm Gateway

The CES framework constitutes of Realm Gateway (RGW) function for incremental adoption of the technology and inter-operability with the legacy IP networks. RGW can be employed as an independent standalone solution or as a part of CES. For outbound connections, its behaviour resembles a traditional NAT that allows the hosts in its

private network to connect to public networks sharing a single public IP address. However, unlike NAT, RGW allows unilateral initiation of inbound connections from public networks towards private hosts via Circular Pool of Public Addresses (CPPA) [14]. The CPPA algorithm is activated on an inbound DNS query for FQDN of the host/service in the private domain, and offers a scalable NAT traversal solution [14] that has advantages over the classical NAT traversal. RGW allows an inbound connection in three different ways:

- 1) a general purpose connection is served by CPPA upon an incoming DNS query for (a) FQDN of the host or (b) service FQDN of the service running on the host;
- 2) incoming HTTP(S) traffic, e.g. towards www.host-a.rgw; is handled by a reverse HTTP proxy; and
- 3) via inbound mapping on public IP addresses similar to traditional port forwarding in NATs.

Figure 9.5 illustrates the CPPA algorithm and shows how RGW accepts connection initiated from Internet hosts towards hosts or services in the private realm:

- 1) *DNS*: Upon receiving a DNS query for FQDN of a served domain, CPPA temporarily allocates (i.e. ~2sec), an available address from its pool of public IP addresses and replies with a DNS response carrying the allocated address, and a TTL=0 to avoid caching. Correspondingly, it creates a temporary half-connection state in RGW. The half-connection state applies endpoint-independent filtering [15] relative to the client. The state ($H:iP_H$, $R_X:oP_H$, P_{protocol} , T_{Tout}) is unique and includes the IP address and port of the private host ($H:iP_H$), the IP address and port on the public side of the RGW ($R_X:oP_H$), the protocol (P_{protocol}), and the lifetime of the entry (T_{Tout}). The TTL=0 of DNS response allows control over the address assignment, e.g. taking out addresses that are under attack, and can offer advantages over static port forwarding in NATs.
- 2) *Data*: Upon receiving a new flow matching the half-connection state, RGW upgrades the half-state to a full connection state and returns the corresponding public address to the circular pool for future use. The inbound packet is forwarded to the private host and subsequent data packets are admitted as a part of the ongoing flow.

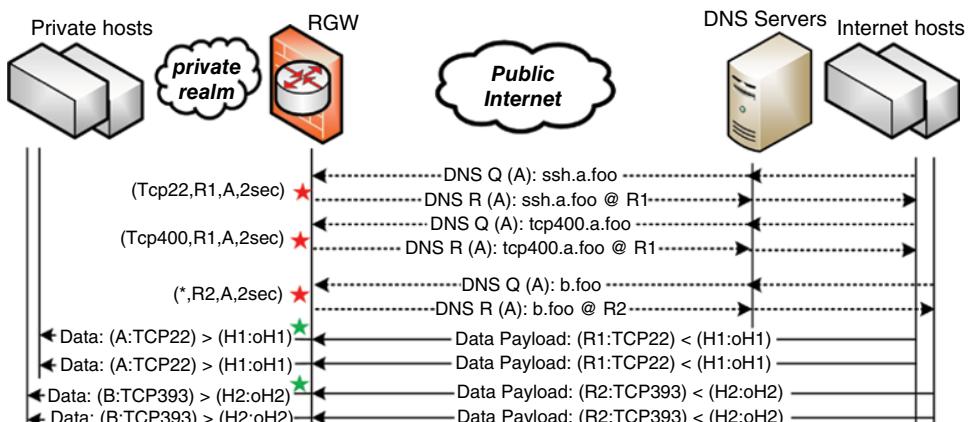


Figure 9.5 RGW serving the connections from the legacy Internet.

Since RGW relies on standard inbound DNS resolutions, it does not require any changes in the current hosts, protocols or applications to support its adoption. The adoption of RGW is transparent to the end hosts, and limits all the changes to edge nodes, allowing the adoption of the technology in the edge networks with minimal disturbance.

9.3.6 RGW Security Mechanisms

RGW could be susceptible to the inherent Internet weaknesses, such as:

- 1) the DNS abuse and source address spoofing. For example, a DNS flood to FQDNs of the private hosts (via open DNS Resolvers, i.e. GoogleDNS) can force the CPPA into DoS, by reserving all the addresses in the CPPA. As a result, the RGW would be unable to accept new inbound DNS requests and corresponding connections.
- 2) Hackers can also send the malicious traffic from spoofed addresses or botnets to claim a connection state reserved by a legitimate user. This would lead to DoS to the legitimate user of the service and would also leak unwanted traffic into the private network. It is pertinent to mention that the already established connections are not vulnerable to these abuses, instead only new inbound connections are affected.

RGW adheres to a set of principles to effectively tackle the impact of these inherent Internet weaknesses:

- a) UDP flow initiations are admitted after the connection was signalled through a secure channel, e.g. SIP(S) [16];
- b) flow acceptance is limited to verifiable sources only;
- c) under network stress, resource access is granted based on the source reputation;
- d) the security mechanisms shall limit all the changes to the edge network.

Although it is possible to take a clean-slate approach and design an architecture free of security weaknesses, we take the deployment constraints as a cornerstone of our design and thus lay special emphasis on the last rule. Adhering to these principles, we define the following mechanisms to tackle the threats to RGW communications.

9.3.6.1 Name Server Classification and Allocation Model

RGW classifies the public name servers into white, grey and black lists. Servers in each category are treated differently by RGW. A DNS server can be white-listed based on the service-level agreement (SLA) with the remote network; or if it executes ingress filtering and indicates the source of DNS query, for example in DNS extensions or Additional records, or if it meets some other pre-conditions. By default, the rest of the name servers have grey-listed and -non-priority access to RGW resources. The name servers are constantly monitored and dynamically redefined into white, grey or black lists based on the influx of the attack traffic. For example, a server that repeatedly exceeds its SLA is penalized with a degraded service and serves the time penalty.

In addition, RGW employs a CPPA address allocation model that rate-limits the number of simultaneous queries from a DNS server or towards a host. The model limits greylist requests to a portion of the circular pool, and under network stress prefers address allocation to white-list servers, and thus assures that whitelist servers always have resources to their disposal.

9.3.6.2 Preventing DNS Abuse

Remote DNS servers can also connect to the RGW via a TCP connection, which provides a spoofing-free channel. The mechanism can be employed as a part of whitelisting a DNS server. Together with DNS extensions or additional records that identify the source of DNS query, it is possible to trace an aggressive host and rate-limit it.

The deployment of RGW in hierachal fashion with the serving-ISP also provides an opportunity to protect the DNS leaf node against attacks. This is possible due to a DNS Relay function that acts as a front end and prevents RGW from direct exposure to the inbound DNS queries. The relay function allows a separation of concerns and makes possible to delegate the security against DNS abuses to a dedicated entity that can independently leverage best practices in tackling the DNS abuse, and thus only forwards legitimate traffic to RGW.

9.3.6.3 Bot-Detection Algorithm

The arrival of the first packet of a flow, that is, TCP SYN that does not correspond to connection state is monitored for hacking activity. Once such mismatching packets exceed a threshold in an observed duration, Bot-detection can process the subsequent mismatching packets to determine if the sender is not spoofed. A non-spoofing check together with the history of misbehavior leads to blacklisting of the sender in RGW, where it undergoes a time penalty during which its connection initiations, that is, TCP SYNs, are dropped.

9.3.6.4 TCP-Splice

To tackle the problem of spoofing in Internet originated connections, RGW may challenge the sender of TCP SYN with a cookie embedded in the Initial Sequence Number (ISN) [17] of the SYN/ACK segment. If the TCP handshake completes, we ascertain authenticity of the sender as non-spoofed and accept the connection. The subsequent data packets are forwarded in compliance with TCP-splicing technique. However, a packet from a blacklisted source is dropped to prevent the connection hijacking. A source can be blacklisted following the Bot-detection algorithm. The Bot-detection method together with the TCP-Splice aims to filter unwanted traffic from both the spoofed and non-spoofed sources.

CES provides these mechanisms as policy-controlled features for network administrators. The mechanisms contribute to establish safer connections from the legacy Internet towards the private network. They protect RGW from hazardous effects of the most common Internet weaknesses and facilitate the deployment of RGW in Internet networks.

9.4 Evaluation of CES Security

This section introduces the prototype network developed as CES proof-of-concept. We define the key performance indicators (KPIs) for evaluating the CES prototype, and present the performance evaluation as well as the results from security testing. Figure 9.6 presents our CES testbed, which is implemented in the Linux environment and simulates the network nodes using Linux containers, connected via standard Linux networking capabilities. The architecture employs Ryu [18] as the SDN Controller, realizing the control plane, whereas the data plane is implemented with OpenvSwitch, which enforces the rules generated by the controller. The communication protocol between the planes is OpenFlow v1.3.

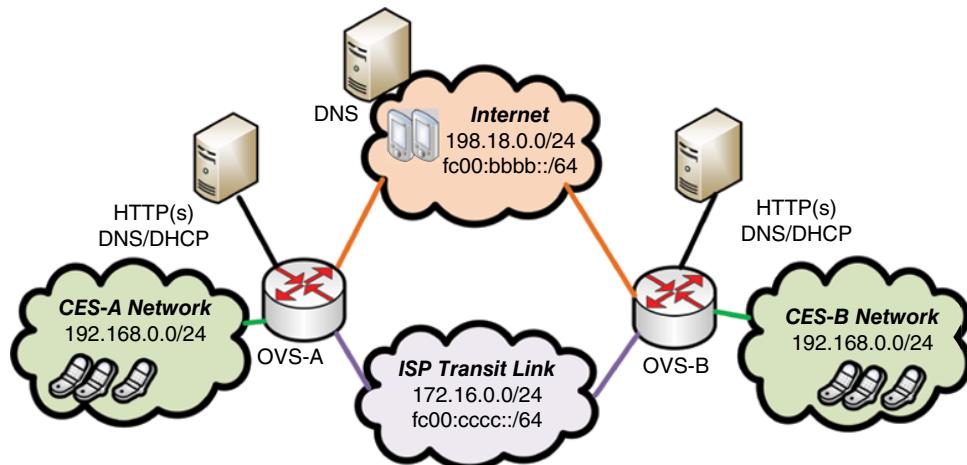


Figure 9.6 Implementation of CES testbed.

The setup comprises of two networks served by their respective CES nodes. Each network has an open flow switch at its edge, which processes the traffic passing it as per the controller generated rules. The edge of the network has two links/interfaces to send and receive the traffic from: (a) legacy Internet, simulated with the public IPv4 addresses; and (b) networks connected via private transit links. The setup allows testing both modes of CES technology: (a) *CETP-based policy communication*, for establishing communication between hosts in different CES networks; and (b) *RGW* for interoperability of CES with legacy IP networks.

We define the following KPIs for CES security testing:

- *Session setup delay*: is the average CETP session setup time when using the CES prototype, including the delay due to control/data plane split architecture;
- *Signalling round trips*: for creating new outbound and inbound CETP sessions;
- *Cost of Security*: is the processing delay introduced to the session setup time;
- *False Negatives*: is percentage of hacker flows wrongly admitted into the network;
- *False Positives*: is percentage of ordinary users classified as attacker;
- *Contribution of security*: compared to CES without the security mechanisms.

9.4.1 Evaluating the CETP Policy-based Communication

As discussed above, the CETP signalling can be split into:

- 1) DNS NAPTR query and response for identifying the locator of the remote CES node;
- 2) CETP policy negotiation, which upon success, allows tunnelling of the subsequent data packets between hosts on the negotiated data RLOCs, across the networks.

We tested and analyzed the connection setup delay for over a hundred new CETP flows between CES nodes. The delay perceived by the originating host is measured as the time to resolve a DNS query for a destination domain, as well as the time to forward

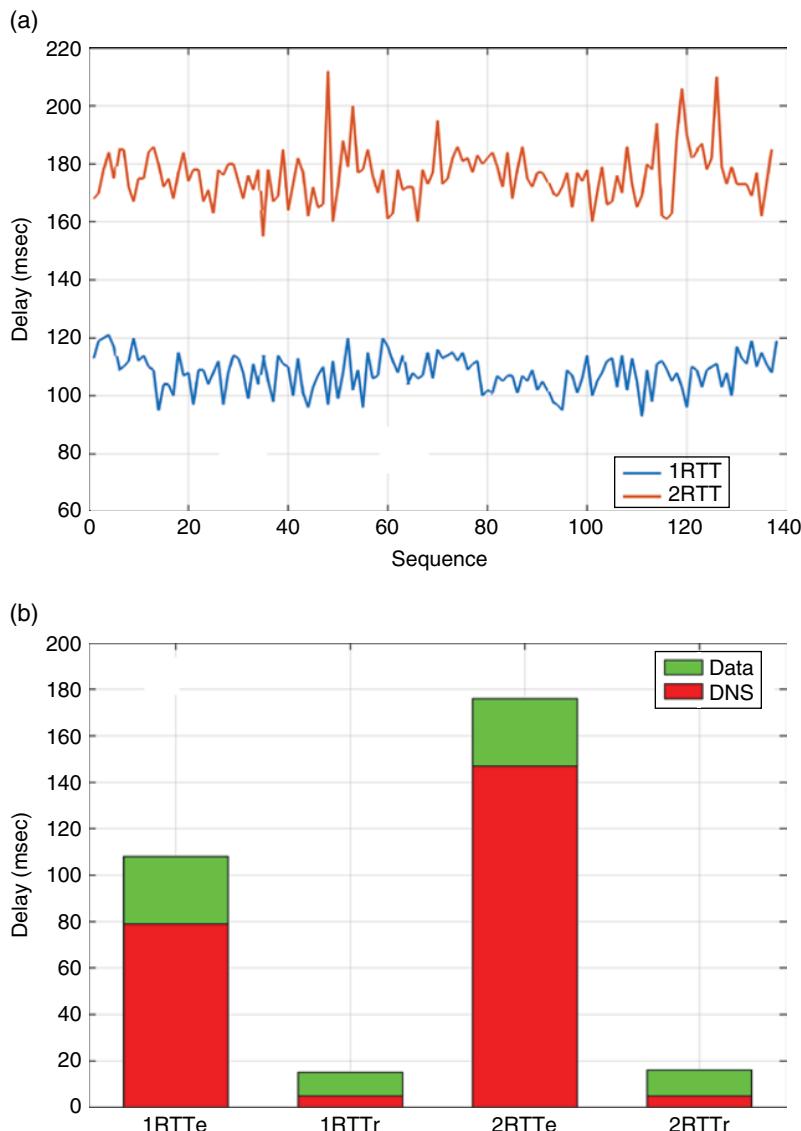


Figure 9.7 CETP Connection setup delay.

the first packet of the data connection (i.e. TCP SYN segment). Figure 9.7 presents the impact of the delay caused by the CETP policy negotiation on setting up of the data connection. The evaluation includes the delay introduced by the separation of control and data plane realized in the local test network, where the CETP signalling can either complete in 1 or 2 RTT, depending on the policies in use. However, due to zero-latency of the simulation links, we will have to add the RTT delays of a real network to obtain the actual connection setup delay.

Most of the delay for establishing a new connection is attributed to the signalling phase. So a new connection setup takes time, but a host that re-utilizes an existing connection undergoes a lesser delay, since the connection state for establishing a data connection already exists in the control and data planes. The only significant delay in this case is for forwarding of the data packet. This is shown in Figure 9.7(b), which splits the session setup into new connection establishment (e) and connection reuse (r). To account for network uncertainties, the CES state machine can absorb any host retransmissions while the CETP signalling is still converging.

9.4.1.1 Security Testing

We conducted a set of tests to evaluate CES security mechanisms. Table 9.2 presents the result of security testing, which reveals effectiveness of the security mechanisms as well as their penalty on CES performance in terms of processing delay.

In our testing, the spoofed addresses failed to place or claim a connection state, as well as from leaking traffic into the CES served network, due to use of the signalling over spoofing-free underlying protocols such as TCP and TLS/TCP. CES uses the proof-of-work mechanism to push the burden of communication towards the sender, such that the sender spends more computing cycles than receiver. This de-incentivises CETP-level floods from remote sources.

The CES authentication mechanism effectively identifies the remote CES node at the cost of a minimal processing delay, using X.509 certificate and an optional signed CETP header for validation. CES authentication is triggered on the first CETP flow from a new source, and is carried together with other network level policies for achieving the trust. Depending on the complexity of policies, there can be one or more round trips in addition to the normal host-to-host policy negotiation, for the first CETP flow. As a result, the first host-to-host level CETP flow establishes in approximately 300 msec for 2-additional RTTs of the CES-policy negotiation. Naturally, end-to-end link latency of real network must be added to obtain the real first session setup delay. In case the connection setup takes longer than the host's retransmission time, the oCES state machine can absorb any host retransmissions while the CETP signalling is still converging.

The subsequent interactions re-utilize this validation result, and thus host-to-host signalling completes in 1 or 2 RTTs depending on the policies, in the duration shown in Table 9.2. The authentication mechanism does not exhibit any false positives or false negatives and hence safeguards CES against CETP-level attacks from unauthorized sources.

Table 9.2 CES Security testing.

| KPIs and Processing delays | Testing results | |
|-------------------------------------------------------|-----------------|-----------------------|
| CETP signalling roundtrips (RTT) | 1-RTT | 2-RTT |
| CETP Signalling delay (msec) | 80 ms | 145 ms |
| False negatives/positives (percentage) | None | |
| Burden of proof-of-work mechanism | 3 ms to Sender | 0.001 ms for receiver |
| CES Authentication burden (on 1 st packet) | 2 ms to sender | 1.8 ms for receiver |

9.4.1.2 Outcomes of the Security Testing

The security testing shows that the CES security mechanisms can:

- eliminate risks from spoofed sources, due to possibility of exchanging CETP signalling on the underlying spoofing-free protocols, i.e. TCP, TLS/TCP, IPSec;
- use of X.509 certificate allows authenticating a signalling source at network-level using CES-level policy negotiation for trust establishment. This can happen both at the level of underlying protocols or exclusively in CETP signalling;
- proof-of-work mechanism in CES pushes the burden of establishing trust to the sender, and prevents CETP level floods towards CES at the signalling level;
- the elimination of spoofing and source authentication allows attributing the misbehavior at signalling or data-connection level to the ID of the remote network and its host;
- the performance penalty (i.e. delay), introduced by security mechanisms to the connection setup is minimal, and the mechanisms do not exhibit any false positives or negatives.

Our implementation is in Python, and is open to further optimizations at the architecture level, which could further reduce the session-setup delay at processing level.

9.4.2 Evaluation of RGW Security

This section presents an evaluation of the RGW security mechanisms in tackling the inherent Internet weaknesses. For this, we implemented the RGW security mechanisms in our prototype, and subjected it to a set of Internet abuses, such as source address spoofing, network/port scans and DNS floods from legacy Internet, to determine the bounds of the RGW security.

We utilize Scapy [19] to craft malicious packets and to launch attacks on RGW. For our testing, this enables legacy nodes in the testbed to initiate: i) spoofed traffic; or ii) network floods from multiple sources (i.e. botnets). The legacy nodes use virtual network interfaces to provide an illustration of many hosts participating in the attack. The attack load is measured in flows per second, whereas the network delay between the nodes is artificially induced. The outcome of the testing reveals the effectiveness and cost of the RGW security, in terms of the ratio of hijacked connections and processing delay introduced in the prototype, respectively.

Figure 9.5 shows how legacy hosts can initiate connections towards RGW. As described in Section 9.3.6, RGW can be susceptible to the abuse of DNS, address spoofing and connection hijacks. The goal of RGW security is to neutralize these abuses that can otherwise force CPPA into a blocking state or launch DoS by hijacking connections of legitimate clients.

First, we analysed the impact of DNS floods and the arrival of different DNS queries on RGW. For this, we pre-configure the whitelist servers in RGW, and submit it to a DNS flood from multiple greylist sources. Figure 9.8 shows that without security, the DNS flood could reserve all the CPPA resources and force RGW into blocking state. In contrast, the address allocation model notes that the DNS source is greylisted and limits the resource allocations for greylist servers to a portion of the circular pool. In this manner, the allocation model prevents the exhaustion of CPPA under DNS floods, and ensures that whitelist servers can access RGW, even under load/attack conditions. Moreover, the rate limitation enforced by the address allocation model makes it difficult to launch DoS by flooding through a single server.

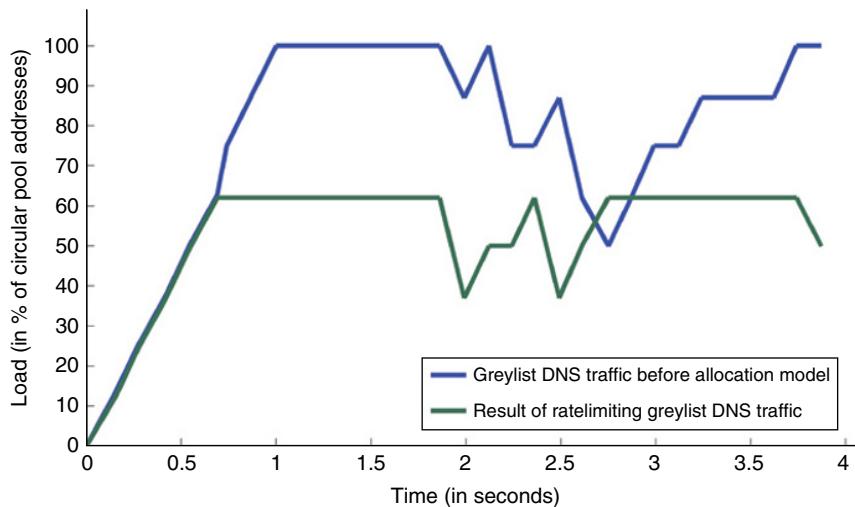


Figure 9.8 Tackling DNS flood towards RGW.

Besides admitting connections based on FQDN queries of the served hosts, RGW also allows connections towards a service running on the end host, using a service-FQDN (SFQDN). The use of SFQDN allows CPPA to make a more specific address allocation that binds the port number and protocol of the service to the allocated IP address. This makes it possible to reuse a public IP address from the pool for any combination of the port (2^{16}) and transport protocol (i.e. UDP and TCP). As a result, the scalability of CPPA improves and it becomes more difficult to enforce the blocking state on CPPA. In this case, the CPPA intermediate state is port- and protocol-dependent but endpoint agnostic. In practice, the upper bound of CPPA scalability is determined by the most popular service that leads to the allocation of a new circular pool address for the already reserved port and protocol combination. Our work in [20] further discusses the scalability due to use of SFQDN.

Our testing shows that a resource depletion attack using SFQDN is more challenging. This is because it now requires a high-rate DNS flood to constantly force CPPA into a blocking state, where the CPPA would be reserved for all the combination of its public IP addresses, ports and protocols. Such a high rate of flood makes it easier to detect the attack and to grey- or blacklist the server that is constantly serving the DNS flood. Moreover, the rate limits on simultaneous domain queries from a DNS server also hinders the attacker's ability to launch floods from a few named servers or open resolvers, for long durations.

The more specific address allocation due to SFQDN also hardens RGW against malicious flows from spoofed and non-spoofed addresses, which aim to hijack the valid user connections. This is because besides the public address, a hacker also has to target the correct port and protocol to compromise the state allocated to a user. This is in contrast to general-purpose FQDN query where IP address allocation is not restricted to a port or transport protocol, and thus applies end-point independent filtering relative to the client.

For the purpose of testing, we forward different concentrations of DNS queries that include both FQDN and SFQDN towards RGW. Traffic patterns: *Test-1* carries 100% FQDN traffic towards RGW; *Test-2* carries 50% FQDN and 50% SFQDN queries; *Test-3*,

carries 25% FQDN and 75% SFQDN queries; and in *Test-4*, 100% of the inbound queries are SFQDN. We stress RGW with these traffic patterns from legitimate clients at a constant rate of 4 requests per second. The connection load is distributed between the private domains located in RGW.

In parallel to the client DNS queries, we trigger a network scan of 40 SYNs/sec from the legacy hosts to CPPA addresses. The figure reveals that for *Test1*, FQDN initiations only, nearly all the connections are hijacked. This is because the hacker is constantly scanning the CPPA at a high rate and would overtake the allocation of a legitimate host when its packet meets the end-point independent state. However, as the share of SFQDN grows and reaches 100% in total inbound DNS queries, the ratio of hijacked connections starts declining to zero (for an all SFQDN traffic in *Test-4*). This is because, besides scanning for the allotted public IP addresses, the attacker also has to scan through the whole port number space (2^{16}) to compromise an allocation. As a result, the hijacking of host allocations drops and legitimate users do not face disruptions or DoS. We present our evaluation of the DNS query types and contribution of SFQDN on a circular pool of 4 and 6 addresses, denoted by C4 and C6 in Figure 9.9. Note that this result applies to the case where the hacker has no prior knowledge nor can the hacker guess which port and protocol is being used. This may apply, for example, when one admin can control both the legitimate client and the server.

Next, we evaluate the RGW security against the unwanted flows and network scans that might originate from spoofed or non-spoofed sources. RGW employs TCP-Splice to filter the risks from spoofed flows. Our testing revealed that spoofed flows (i.e. TCP SYNs) failed to compromise an allocation of a valid client. This was due to the SYN cookie mechanism, which admits a flow only after the spoofing is eliminated, in the next inbound TCP ACK segment. In terms of performance, this limits the reusability of the public IP address (and the port combination) by the same duration for the next

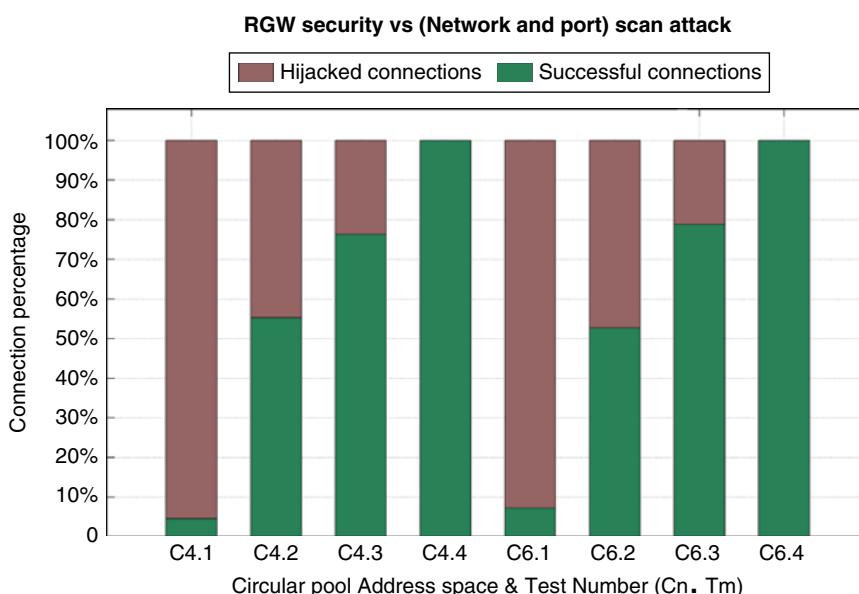


Figure 9.9 Impact of inbound DNS query type (SFQDN) against versus network/ports scans.

inbound connection. Figure 9.10 presents this delay in assigning a TCP-half connection state due to use of TCP-Splice. In a real network, the end-to-end latency for TCP messages would be added to compute the total delay in assigning the half-state. It is possible to reduce the average delay penalty caused by the TCP Splice by using it selectively, that is, on privileged ports, or under network attacks only.

To test the effectiveness of the Bot-detection method, we subjected RGW to malicious flows (i.e. network scans) from the non-spoofed sources, emulating a botnet. Figure 9.11

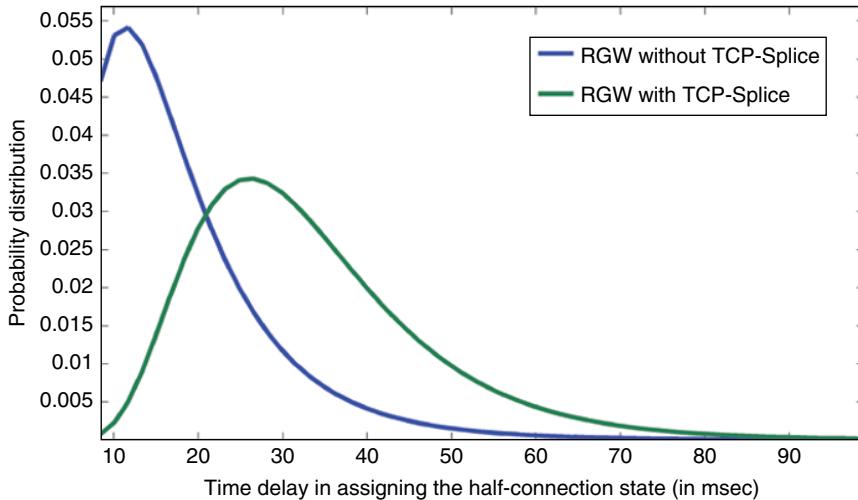


Figure 9.10 Delay in assigning TCP half-connection state due to TCP-Splice.

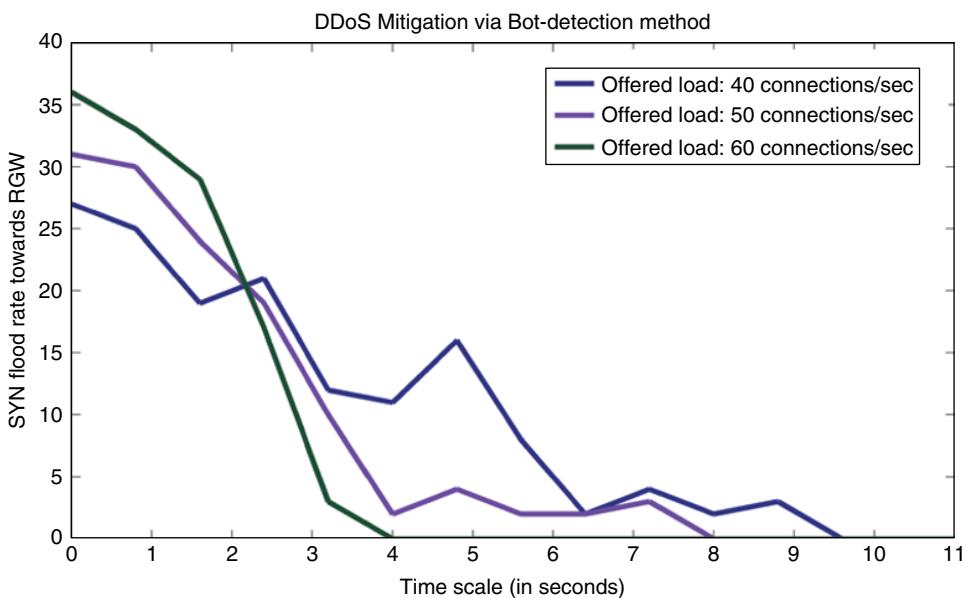


Figure 9.11 Bot-Detection method to mitigate SYN flood from botnets (non-spoofed sources).

presents the effectiveness of Bot-detection method. The figure shows that Bot-detection is more reactive to high-rate SYN floods and filters them earlier, as they quickly meet the detection threshold. Bot-detection constantly tracks the packets that are dropped for not meeting any connection state, and once a sender exceeds a threshold, it is blacklisted following a non-spoofing test. As a result, subsequent packets from the hacker fail to claim any state allocations in RGW and ration of legitimate connection rises. A more detailed analysis of the Bot-detection method and different influencing parameters for RGW security is presented in [21].

To gain a more realistic view of security, hackers can be divided into: i) probing or scanning hackers; and ii) advanced hackers. A probing hacker scans the entire CPPA address space and port range to discover the available services, IP addresses, ports or NAT mappings. It is likely that such an attacker, due to its limited knowledge of the victim and thus random network scanning, will fail to pose risk at RGW for SFQDN based traffic, as shown in Test-4 of Figure 9.9. In comparison, an advanced hacker may already know services/ports to target, via knowledge sharing among hackers or using botnets that perform the service discovery. As a result, the hacker can target the SYN floods to the specific ports.

Clearly, the results show that RGW attains best-case security against regular network scans for SFQDN admitted traffic, where the hacker is not advanced and simply scans the network for available services or IPs. Under the premise that the hacker targets the served ports, it is possible that SFQDN naming is changed to new service ports. This will force the attacker to restart its service/port discovery cycle and allow RGW to regain its best case security. Such use of SFQDN is possible in cases where a single administration owns or manages both the remote hosts and the RGW. For example, in Internet of Things (IoT), both the communicating nodes and gateway can fall under a single administration. In the absence of such a scheme, Figure 9.12 shows the security of SFQDN admitted traffic against an advanced hacker.

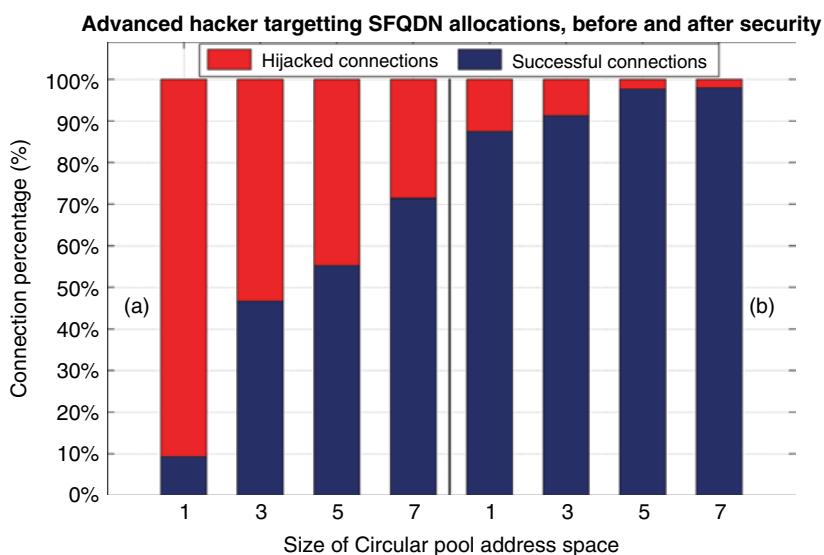


Figure 9.12 Security of SFQDN allocations against advanced hackers: (a) without; and (b) with RGW security.

Table 9.3 Security mechanisms and their performance.

| Security threats | Mechanisms | Cost of Security | Outcome |
|-------------------------|-------------------------------------------------------------------|-------------------------------------------------|--------------------------------------------------|
| Source address spoofing | TCP-Splice | Extends duration of assigning the half-state | Spoofing eliminated |
| Bot-controlled flows | Bot-detection | Processing delay | Possible False Negatives |
| Malformed ACK segments | SYN cookie verification | Verification cost (<0.01 ms) | Malformed ACKs filtered |
| Spoofed DNS requests | DNS/TCP, DNS Relay and Ingress filtering | Slow first DNS resolution due to TCP connection | Spoofing eliminated |
| DNS-floods | Address allocation model Rate limitations Server reputation | Server tracking | Less trusted servers face congestion, under load |

Figure 9.12 shows security of SFQDN allocations against an advanced hacker. It is pertinent to mention that in our testing, no state allocation is compromised by spoofed flows. However, before a non-spoofed flood is mitigated, some of its packets can beat a legitimate host in claiming the allocated state, causing DoS to the actual client. Thus the security of RGW can exhibit false negatives during attack. However, these false negatives reduce as the attack progresses, since the active bots will be filtered upon exceeding the Bot-detection threshold. The ratio of false negatives can further reduce by:

- 1) proper dimensioning of the network that presents more opportunities for a hacker to meet the detection threshold; or
- 2) by building and employing reputation of the source addresses.

Though our testing identified few false negatives, RGW did not exhibit any false positives, that is, classifying a valid client as attacker. We argue that in the RGW networks, a false negative is not as severe as a false positive; since a client that suffers hijacks can always re-attempt to access the service hosted in the private realm.

The use of TCP-Splice together with the Bot-detection method aims to protect RGW against abuses from spoofed and non-spoofed sources, respectively, and hence ensure that only a valid client is admitted to the private network. A new version of our prototype is to consider replacing TCP-Splicing with the SYN proxy for improved performance of the prototype. Table 9.3 presents an overview of the implemented mechanisms for securing RGW against typical Internet abuses, their contribution to security and impact on the RGW performance.

9.5 Deployment in 5G Networks

The principle is that each network administrator, mobile operator or ISP makes independent CES deployment decisions. The deployment of CES can take place one stub network at a time, since it supports the RGW functions for interoperability with the legacy networks.

For example, in the case of 5G mobile networks, the CES function can be deployed as part of the gateway application that provides connectivity to the Internet and other public data networks. This effectively complements the functions currently performed by the Packet Data Gateway (PGW) in 3GPP networks. Since 5G core network will be expectedly defined and controlled by a set of virtualized network applications, we developed a demonstrator of the 5G-network control plane using a set of SDN applications [22], where CES serves as the access provisioning application. The demonstrator carried the traffic between LTE user devices and the Internet via an eNodeB supplied by Nokia.

The deployment of the proposed CES framework as a part of 5G, or other 3GPP mobile networks, can provide enhanced traffic management by eliminating a part of unwanted traffic. This can be initially done in a specific slice of the 5G EPC carrying ultra-reliable services. In the long run, there is good motivation to integrate CES with all PGWs in mobile networks.

For instance, by using existing control nodes and seeking policy requests, CES can collect sufficient information from a source to establish its legitimacy. CES allows users, hosts and applications to define their reachability policies and hence control the traffic they deem interesting. Therefore, the contribution of CES to the future networks can be in reducing or eliminating malicious traffic from being a cause of failure to the legitimate services, and thus contributing to ultra-high reliability of services in 5G. CES adheres to the SDN-principles and can be leveraged in other networks also, for example for providing access to objects in the Industrial Internet or Internet of Things (IoT). For deployment into corporate network gateways, CES functionality needs to be integrated into firewall products. Compared to various tunnelling proposals that focus on solving the core scalability issues in the Future Internet, CES is focused on trust and in providing added value to end customers as a result of the improved security.

Figure 9.13 shows an overview of CES deployment in mobile networks, where it can be co-located with Packet Gateway (PGW) for fine-grained security. In the following sub-sections, we discuss a set of use cases for CES deployment, and further elaborate the use case in terms of the CES operations, security benefits, scalability and reliability of services.

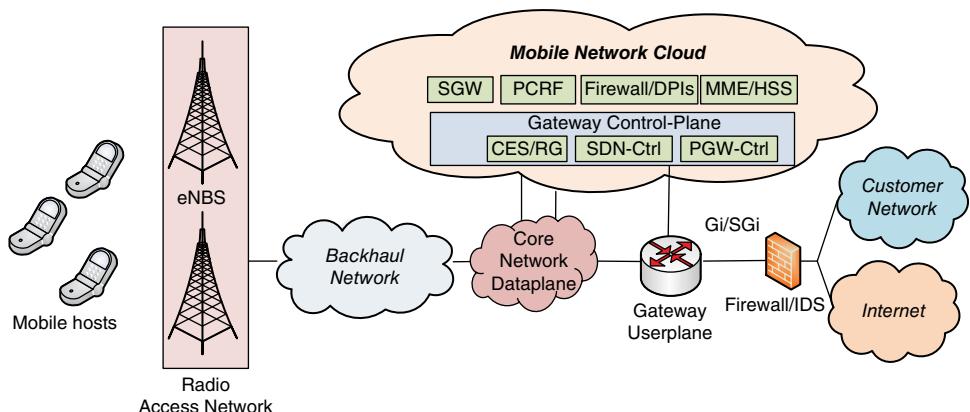


Figure 9.13 CES deployment in 5G mobile networks.

9.5.1 Use Case 1: Mobile Broadband

9.5.1.1 Deployment and Operations

A mobile operator decides on the adoption of Customer Edge Switching independent of the others. Within the operator network, the deployment progresses one network segment at a time. CES complements the Packet Data Gateway (PGW) in the mobile core network.

The mobile operator has millions to hundreds of millions of subscribers, as well as many supporting systems such as edge computing platforms and connections to content providers. Therefore, the mobile operator may also deploy network-based security monitoring, security incident aggregation and trust processing for the end system security. Using these systems, a mobile operator defends all its subscribers from the rest of the world based on a uniform, dynamic policy that makes use of all evidence collected by its own and the subscriber-owned devices. In addition, the mobile operator may have contracts with security software vendors that will deploy regular security updates, and may limit the policies applicable to the mobile devices. This gives the option of patching vulnerable applications on a network level before a proper application software update is available and has been deployed by the end users.

Initially, the policies for CES, RGW and host could be simple. Improved security is achieved gradually, as more fine-grained policies are deployed using the different repositories for sourcing the policies and policy constraints. For the purpose of improved heuristics, CES and RGW can treat differently the DNS servers that make queries to the mobile operator DNS service. Servers that reveal the source of DNS query, that is, via DNS extensions, and apply strict ingress filtering, can be preferred compared to DNS servers that serve any query and any sender. For corporate customers, the mobile operator can use the policies defined by the corporate admin in a suitable policy database, fully administered by the corporation.

The mobile operator may allow its subscribers to define a policy for each of its users within some constraints set into the mobile-operator-administered policy database. The updates from the security software companies can also feed this database, with the security perspective. A subscriber fundamentally selects its desired services (i.e. subscription) from a set of operator offered packages, which can be translated into user policies. The principle of allowing the user to modify policies is that a user can always make its policy more restrictive as well as deploy new Apps that imply a certain policy – template for such a policy should be given in the App license agreement. For example, a user can employ a service “me and my gadgets”. It would allow limiting access to the subscriber-owned sensors or servers to a set of devices that can provide an identity defined in the policy.

9.5.1.2 Security Benefits

Ultimately, each mobile device can have its own security policy that admits traffic only to applications that are actually deployed on the device. If some device is running an outdated software or application, and its network traffic can be identified by the deployed CES release, all traffic to and from the application on the device can be restricted. Such restriction can be based on the type of subscription the user has. This possibility may be of particular interest to corporate customers who may want to handle the policy management by themselves, possibly in cooperation with one or

more device security software vendors. The mobile operator could also itself cooperate with security software vendors in the area of virus detection, and sell to its customers the improved security services based on application level security processing in its edge cloud.

9.5.1.3 Scalability

A CES in its basic architecture is composed of a separate data-plane node and a data-centre hosted control plane. The data-plane node is stackable; the operator buys more boxes or blades as a function of the number of subscribers and traffic volume per subscriber. The scalability of policy services can be achieved by a suitable grouping of subscribers into classes or packages, where each class has the same policy for all subscribers. Such grouping will also make it easier to sell mobile services to consumers who are unwilling to be bothered with too much detail.

Using the same addressing principle as current mobile networks, each mobile device resides in its own private address space. Device-to-device communication takes place through CES nodes. The number of CES RLOCs depends on the type of outbound interfaces supported in one or more CES data-plane nodes, for example on 1 Gbit/s levels to 100 Gbit/s levels.

9.5.1.4 Reliability

Each served mobile device will be able to reach another mobile or Internet host through the mobile backhaul and core network. In case a CES DP node fails, the virtualized SDN orchestration will move the users served by the failed DP node to other DP nodes. The failure disables the use of failed DP routing locators for any following use. Correspondingly, CES will start advertising new priorities for its routing locators (RLOCs) to remote systems. To what extent it makes sense to develop CES-to-CES recovery operations as compared to end-to-end recovery operations is for further research.

9.5.2 Use Case 2: Corporate Gateway

9.5.2.1 Deployment and Operations

A corporation decides to pursue CES deployment. Arguably, the first to deploy CES as a corporate network gateway are the companies with multiple sites and a need for secure extranet connectivity. Since CES uses centralized policy management, the company may also deploy a framework to manage policies in the centralized policy database. To maintain a high level of security, the company will make contracts with security software vendors who will support the company in policy creation and policy constraints. Desktop management will link this with the policy management. The company may optionally decide to share security incident information with a security service provider, who may supply black-, grey- and whitelists of remote networks in return. Optionally, the company can have its own trust management. The adoption of CES technology by many companies that cooperate with a security services provider can also gradually lead to formation of a proper trust domain. An Internet service provider would always be in a position to do network-based monitoring in an efficient manner. The company can either perform such monitoring itself or buy this function as a service from its ISP.

9.5.2.2 Security Benefits

Besides the fact that policy management will be centralized, it is possible to have a uniform policy across the whole company. Intranet traffic protection for a company's site-to-site connections is by default available using encrypted CES-to-CES tunnels. Using CES, security will be an integrated feature for end- and network-based systems. A company can enter into extranet contracts with remote partners running CES networks, such that some security rules are agreed and implemented by CES policy management and the corresponding operational processes. For example, if a company is in manufacturing (i.e. a paper mill) and has lots¹ of sensors and actuators served through 5G radio, these devices would have been supplied by a number of vendors that also provide technical maintenance for the manufacturing company. Each outsourcing contract would imply a policy that gives vendor access to a set of wireless devices in the manufacturing company's network. Both parties of the outsourcing contract in this case would use CES nodes to enforce the policies (agreed in the contract) for secure communication.

9.5.2.3 Scalability

The scalability challenge in this case is mostly a subset of the use-case of mobile broadband. It makes sense to connect each Ethernet level VLAN and the corresponding IP subnet to CES for applying a consistent security policy in the intranet, for communication between the subnets. Alternatively, legacy corporate-network routers can be used to connect the subnets in the corporate network, and then from the CES perspective, all the hosts of a corporate network are not in different address spaces and thus CES cannot offer a means for controlling the intranet traffic across subnets. Naturally, outside the scope of CES security, a company can have servers with globally unique IP addresses providing services to the public.

9.5.2.4 Reliability

CES will support multi-homing to several ISPs. It can choose to advertise different priorities of its RLOCs to different remote CES systems.

9.5.3 Use Case 3: National CERT Centric Trust Domain

9.5.3.1 Deployment and Operations

One or several ISPs or mobile operators with packet services deploy CES nodes for providing security services to consumers, both mobile and wire-line. The ISP can also offer Carrier-Grade Realm Gateway services, which allows its subscribers to run servers on their hosts in a controlled manner. The ISP or mobile operator can host CES services for companies offering the option of outsourced security services to other firms. In this case, the ISP will cooperate with security software vendors for policy constraints, etc. The individual corporations can opt to deploy their own CES services, but this will cause reluctance to share security incident information because of the reason discussed earlier. At this stage, the benefits of CES adoption come from allowing servers in private address space, coherent and uniform policy management, and the possibility of patching

¹ A paper mill would, e.g. have some 20 000 to 30 000 sensors for monitoring different aspects of the production process.

vulnerabilities in the end systems using the operator or company CES. For example, when it is known that a consumer has deployed a vulnerable gadget such as an Internet TV or a game-box in its network, the operator can use its policy management and CES to block all suspicious traffic to and from such gadgets. For this to be practical, ISPs must have a way of earning revenue from better security services to the end systems. One possibility could be that the national regulator will set a price for such a service. The recent cases, where hackers built a botnet of nearly 10 million such gadgets and targeted DDoS floods of a Terabit per second, motivates this kind of vulnerability patching, as well as pricing mode. The problem stems from the fact that it is fully accepted to sell the gadgets that require Internet connectivity to a consumer, without testing for their security compliance or even their software update capability.

After this initial stage, the national CERT can decide to deploy dynamic trust management for the national networks, and either runs the Global Trust Operator (GTO) function itself or delegates that function to a firm. Homomorphic security is used in security incident reporting and report aggregation. This will increase the willingness of companies to share their security incidents. The earlier deployed CES nodes at ISPs, mobile operators and companies then execute the black-, grey- and whitelisting policies as instructed by GTO. The adoption of GTO functions will be supported by national regulation that already today sets the rules for ISP-based network monitoring, so that the privacy of communication is maintained. The regulator also mandates the security incident sharing obligations, and as a result national infrastructure will benefit from GTO functions to improve their robustness against the cyber warfare as well as hacking by criminals.

9.5.3.2 Security Benefits

When national GTO has been deployed and most customer networks are protected by cooperative firewalls that automatically react to GTO authorized blacklists, DDoSing of legitimate services will be much harder than before. The patching of vulnerable consumer gadgets would lead to reduced Bot penetration at the national level, and contribute to improved security of national infrastructures. Another possibility could be to use blacklists and build a national firewall against all cross-border traffic, but this is rather costly and perhaps will not be needed on all borders.

9.5.3.3 Scalability

The use of homomorphic encryption contributes to trust management at the national level. We implemented this in [9]. The experiments show that the approach is feasible. Stackable data plane nodes and cloud-based control plane ensure CES level scalability.

9.5.3.4 Reliability

Even if the GTO is temporarily unavailable, the pre-existing black-, grey- and whitelists can still contribute to protect attacks from known blacklists.

9.5.4 Use Case 4: Industrial Internet for Road Traffic and Transport

9.5.4.1 Deployment and Operations

For the data traffic that comes to and from vehicles, which will either have human drivers or be operated remotely by fleet control pilots, a 5G network slice is created for road safety related data. Such a slice can span several countries. The resources for packet

transport in the slice could be obtained by contracts with a number of incumbent operators. The company operating the slice will aim to provide a highly secure and robust network. The slice will have its own core-network control plane and uses CES control/data plane nodes for interconnecting different segments. In addition, the fleet control centres and other road transport related servers will be connected to the slice. A fleet control centre may have a gateway to the Internet on application layer. The slice, for example, can carry the telemetry from vehicles to fleet control: a vehicle will have many cameras, and the video feed from any camera can be streamed through the slice to the corresponding fleet control centre at any time. The slice may have edge computing servers for the purpose of collecting locally relevant information from the vehicles or delivering the locally relevant information. Edge computing infrastructure can, for example, be rented from the incumbent mobile operators. By allowing only security certified devices to connect to the slice, the security can be further hardened.

9.5.4.2 Security Benefits

By re-using 5G technology and choosing CES to control -access to the wireless segments, as well as to the data centres, we create an isolated secure network where all traffic may be encrypted and carried edge-to-edge after mutual authentication.

9.5.4.3 Scalability

Clearly, all the values here reside in the end devices and the network edge. Naturally, the services are available in the coverage area of the underlying mobile networks. The allocated transport resources do not need to be static. The orchestration functions in the slice and the underlying mobile network controller may have an on-line interface for requesting additional transport capacity to and from particular base stations, where more vehicles are arriving as well as releasing capacity that is no longer needed.

9.5.4.4 Reliability

By sufficiently generous dimensioning of the resources and an online link to the controller of the underlying mobile networks, the slice can deliver high reliability of services. Each fleet data centre runs on a high availability platform and can have multiple CES nodes for the interconnection to the slice.

9.6 Conclusion

This chapter identifies the security challenges faced by state-of-the-art in mobile networks and introduces CES as a framework that can address the classical weaknesses of the Internet, and provides means to protect Internet networks, such as mobile networks, against a constantly evolving threat landscape. Compared to best-effort nature of the Internet that solely attends to the interests of the sender, and is often abused by hackers, the policy-based communication in CES allows negotiating the interests of the sender with the receiver interests and thus filters the unwanted traffic. The adoption of CES allows individual users, hosts or application to express their interests in terms of a *policy* and hence control the traffic they deem interesting. A policy can specify the set of requirements for establishing a communication. A CES node is deployed at the network edge, where it replaces NATs, and acts as a

connection broker that exchanges and negotiates the policies/interests of the hosts or applications, which it serves, with the remote hosts.

The adoption of CES offers multiple advantages over the state-of-the-art in mobile networks. For example, CES allows treating each user, host and application differently, since the nature of communication can differ between host-to-host and application-to-application. CES also supports several mechanisms to establish flow legitimacy, authenticate the sender, protect the network and its infrastructure against Internet attacks, and provides these mechanisms as policy-controlled features. This decision of applying a particular security scheme is left to the network administrator, for example according to the security conditions. In particular, the elimination of spoofing allows attributing the misbehavior evidence back to the identity of the sender or its network, and lays the foundation for establishing the reputation of Internet entities, for example networks, particularly the ones that do not take corrective actions and hence keep forwarding the malicious traffic. The aggregation of these evidences under a GTO can lead to generating white-, grey- and blacklists of sources and the CES firewall can accordingly admit, rate limit or deny the traffic. The cooperative firewalling of CES can lead to filtering of malicious traffic close to the sender, upon receiving host misbehavior evidence from remote CES node or information from GTO, and hence limit the extent of bandwidth-saturation attacks on (Gi/SGi interface of) the mobile networks or the corporate networks.

The adoption of Software Defined Networks (SDN) facilitates deploying new services in the network. Based on this, we have discussed the technology deployment in the networks and proposed that CES function be integrated with PGW in 3GPP architecture. In combination with other network control nodes, that is, PCRF, HSS and P/S-GW, CES can enhance the traffic management and establish fine-grained security to safeguard the network against ills of the Internet: spoofing, botnets, network scans, DNS floods and DDoS. We have also discussed the potential use cases of CES in mobile broadband, corporate networks, national CERTs and Industrial Internet scenarios from operation, security, scalability and reliability perspectives.

The evaluation of the security mechanisms reveals that CES can protect the networks against classical Internet attacks at the cost of a negligible processing delay. CES can further harden its security offerings by leveraging a commercial firewall solution that uses the best current practices for tackling Internet attacks. Unlike NATs, CES does not use cumbersome NAT traversal protocols for admitting flows into the private network. Instead it offers a solution that scales well to the battery-powered mobile and wireless hosts.

Due to its support for a rich set of security mechanism that can trigger on both source and destination addresses in a scalable manner, as well as many features of host behavior, CES blends the boundary of closed and open networks effectively by executing security at the network edge node. This can potentially relieve the Internet non-default core from supporting numerous Virtual Private Network routing tables that reside in the high-speed memory of the core BGP routers. Effectively, the necessary functionality is executed at the edge nodes and the edge cloud. This can improve the scalability of the non-default core in the Internet.

CES adoption in networks does not require any changes in the existing hosts or protocols and facilitates incremental (i.e. one-step-at-a-time) deployment of the technology due to RGW functions. CES seeks to be mostly independent of the applications, but where this is not possible, it continues to support the traditional methods of NAT traversal.

References

- 1 ITU-T, Global ICT developments (2016) [Online]. Available at: <http://www.itu.int/en/ITU-D/Statistics/> [accessed 30 11 2016].
- 2 5G Security (2015) *Scenarios and Solutions*, Ericsson White paper, June.
- 3 CISCO (2016) *Cisco Annual Security Report*, CISCO.
- 4 SRX Series AS Gi/SGi Firewall for Mobile Network Infrastructure Protection. Juniper Networks, White paper.
- 5 Lauhde, M. (Interviewee) (2016) *SSH Communications Security* [Interview].
- 6 Liyanage, M., Ylianttila, M. and Gurtov, A. (2014) Securing the control channel of software-defined mobile networks. *Proceedings of the IEEE 15th International Symposium on World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, Sydney, Australia.
- 7 Kantola, R., Kabir, H. and Loiseau, P. (2017) Cooperation and end-to-end in the Internet. *International Journal of Communication System*. 30, 1–18.
- 8 Yan, Z., Kantola, R. and Shen, Y. (2012) Unwanted traffic control via hybrid trust management. *IEEE TrustCom*, Liverpool, UK.
- 9 Zhang, L.F., Yan, Z. and Kantola, R. (2017) Privacy-preserving trust management for unwanted traffic control. *Future Generation Computer Systems*, 72, 305–318.
- 10 Daigle, L. (2002) *IAB Considerations for UNilateral Self-Address Fixing (UNSAF) Across Network Address Translation*, RFC 3424.
- 11 Meallling, M. and Daniel, R. (2000) *The Naming Authority Pointer (NAPTR) DNS Resource Record*, IETF RFC 2915.
- 12 Rekhter, Y., Moskowitz, B., Karrenberg, D., Groot, G.J. d and Lear, E. (1996) *e.Address Allocation for Private Internets*, IETF RFC 1918.
- 13 Hinden, R. and Haberman, B. (2005) *Unique Local IPv6 Unicast Addresses*, IETF RFC 4193.
- 14 Llorente Santos, J., Kantola, R., Beijar, N. and Leppäaho, P. (2013) *Implementing NAT Traversal with Private Realm Gateway*, ICC 2013.
- 15 Ed, F.A. and Jennings, C. (2007) *Network Address Translation (NAT) Behavioral Requirements for Unicast UDP*, RFC 4787.
- 16 Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J. et al. (2002) *SIP: Session Initiation Protocol*, IETF RFC 3261, June 2002.
- 17 Eddy, W. (2007) *TCP SYN Flooding Attacks and Common Mitigations*, RFC 4987.
- 18 Mechtri, M., Benyahia, I.G. and Zeghlache, D. (2016). Agile service manager for 5G. *Proceedings of the IEEE/IFIP Network Operations and Management Symposium (NOMS)*, Istanbul, pp. 1285–1290.
- 19 SCAPY (2016). Available at: <http://www.secdev.org/projects/scapy/>
- 20 Llorente, J. and Kantola, R. (2015) Transition to IPv6 with Realm Gateway 64. *IEEE International Conference on Communications (ICC)*, London.
- 21 Kabir, H., Llorente Santos, J. and Kantola, R. (2016) Securing the private realm gateway, *IFIP Networking*.
- 22 Costa-Requena, J., Kantola, R., Llorente, J. and Ferrer, V. (2014) Software Defined 5G Mobile Backhaul. *1st International Conference on 5G for Ubiquitous Connectivity*, Levi.

10

Software Defined Security Monitoring in 5G Networks

Madhusanka Liyanage¹, Ijaz Ahmad¹, Jude Okwuibe¹, Edgardo Montes de Oca², Hoang Long Mai², Oscar López Perez³, and Mikel Uriarte Itzazelaia³

¹ Centre for Wireless Communication, University of Oulu, Oulu, Finland

² Montimage, Paris, France

³ Nextel S.A., Bilbao, Spain

10.1 Introduction

Network monitoring is a crucial network management enabler for mobile networks. Its importance is growing with the continuous growth of network traffic and the adoption of virtualization. A monitoring system gathers network statistics, traffic, application and user profiles, and flow samples at various intervals and granularities to evaluate the status of the network for different management tasks such as application identification, anomaly detection, network forensics, load balancing, traffic engineering, enforcing Service Level Agreements (SLA), QoS/QoE and network maintenance. Therefore, a network monitoring system should be capable of monitoring the network and traffic flows at different granularities and for obtaining measurable metrics such as level of aggregation, time intervals, bandwidth utilization and accuracy.

Traditionally, network monitoring systems are deployed at specific locations of the mobile network to monitor data at network boundaries or ingress or egress ports. Similarly, many of the currently used security monitoring systems rely on physical systems and interfaces. However, mobile networks are evolving and so also is the complexity in network management and security, hence, the monitoring systems used today won't be able to support the dynamically changing topologies and increasing varieties of new technologies such as cloud networks and virtual environments. These challenges are exacerbated for wireless networks since the radio channels are open to jamming attacks, and the access networks are prone to disruption on critical links, MAC abuse and flooding attacks [1].

One of the key transformations in 5G network is the use of two new concepts: Network Functions Virtualization (NFV) and the Software Defined Networks (SDN) [2–4]. SDN decouples the network control from the data-forwarding devices and enable programmability by introducing programmable interface into their networking equipment. The control plane is centralized in high-end servers with the capability of

programming multiple network equipment at run time. The SDN control plane has global visibility and finer control over the packets traversing the network. Since the network is controlled from centralized controllers and the network components have programmable interfaces, network monitoring is augmented to a higher level in terms of efficiency, cost and complexity. NFV is ETSI standardized architecture that separates network functionality from the hardware. NFV means that network functions will be running as a service in commercial off-the-shelf hardware.

On the one hand, the limitations of legacy monitoring systems to secure wireless networks can be overcome by introducing novel monitoring architecture based on SDN and NFV. On the other hand, the use of SDN and NFV bring new challenges to network trouble shooting and monitoring. This chapter investigates the challenges introduced by SDN and NFV in 5G networks and how the 5G operators need to tackle them by using efficient network monitoring solutions. Moreover, we highlight new opportunities that will help achieve efficient SDN- and NFV-based 5G network monitoring.

10.2 Existing Monitoring Techniques

Several network monitoring techniques with different levels of capabilities exist in today's network management space [2]. First, we have router based monitoring protocols which allow gathering information supplied by NEs (Network Elements):

- *Simple Network Monitoring Protocol (SNMP)*: for the management of NEs and high-level information on resource use (e.g. monitor bandwidth usage of routers and switches port-by-port, device information like memory use, CPU load, etc.);
- *Remote Monitoring (RMON)*: for the exchange of network monitoring data; and,
- *Netflow or sFlow*: for collecting information on IP network flows and bandwidth usage.

These protocols are mostly dedicated for performance analysis and network management, but they have also been used for detecting some security problems, for instance NetFlow. Current networks are also using packet sniffing, DPI (Deep Packet Inspection), DFI (Deep Flow Inspection), virus scanners, malware detectors and other techniques for analysing network packet headers, complete packets or packet payloads. They are used by NIDS (Network Intrusion Detection Systems), IDPS (Intrusion Detection and Prevention Systems), firewalls, anti-virus scanning appliances, content filtering appliances, and when combined with different methods (e.g. statistics, machine learning, behaviour analysis and pattern matching), to detect security breaches (i.e. passive security appliances) or prevent/block detected security problems (i.e. active security appliances).

Network monitoring solutions come in different variants, depending on what they measure and how they collect the data:

- 1) *Active Probing*: a service-centric approach that collects data based on synthetic measurements, i.e. ICMP Echo Requests, HTTP GET requests or specially crafted packets. Often these measurements are trying to analyze properties of the network that would be impossible to capture from pure passive measurements and are arguably the only way to measure service availability.

- 2) *Device Polling*: a device-centric approach that queries devices typically using SNMP (Simple Network Management Protocol), collecting interface status information, traffic volumes, device load, CPU, etc.;
- 3) *Flow Collection*: solutions that collect traffic information from network devices such as routers/switches. Here traffic can be aggregated in flows using, e.g. Cisco Netflow and stored on disk for post-analysis. Flow data is easier to analyse and process than packet data, but provides less granular information;
- 4) *Packet Analysis*: usually involves a SPAN port from a switch or a network tap and extracts information from individual packets, including information from payloads through DPI (Deep Packet Inspection);
- 5) *Log Analysis*: are solutions that collect machine generated data typically in the form of log files (e.g. syslog) and present a query interface to correlate events across different types of systems, e.g. routers, web servers, load balancers.

Combining the above-mentioned sources of information, we have what is called “Security information and event management (SIEM)” technologies. SIEM provides, on the one hand, security information management (SIM), and on the other hand, security event management (SEM). SIEM technology aggregates event data produced by security devices, network infrastructures, systems and applications. The primary data source is the log data, but SIEM technology can also process other forms of data, such as NetFlow and packet capture (DPI). Event data is combined with contextual information about users, assets, threats and vulnerabilities. The data is normalized, so that events, data and contextual information from disparate sources can be correlated and analysed for specific purposes, such as network security event monitoring, user activity monitoring and compliance reporting. This technology provides real-time security monitoring, historical/trends analysis and other support for incident investigation (e.g. forensics) and compliance reporting.

10.3 Limitations of Current Monitoring Techniques

The legacy monitoring systems have a number of limitations that could be solved by software defined monitoring systems. These limitations are in two-fold. First, is the limitation of the monitoring systems themselves, these include high complexity and operational costs, as well as delays and overheads. For example, the currently used vendor-specific monitoring systems come with hardwired operational logic in their firmware. This means that changes in the legacy monitoring system either require complex configurations or changes in their firmware. As a result, these systems lack the flexibility needed and cannot cope with the dynamic changes in network conditions.

Similarly, there are inherent limitations in the monitoring systems used today. For example, the required synchronization between observation beacons in passive measurement schemes increases the complexity of the system, as well as the delay in the monitoring process. The active measurement methodology on the other hand, increases network overhead by inducing additional packets. Thus, additional packets in the active approaches influence the accuracy of measurements [5].

The second limitation is imposed by the operational environment, such as stagnant behaviour of the network, bandwidth constraints, and complexity of the environment. The currently used monitoring systems obtain network statistics or packet samples from

some agreed-upon locations and choke points. However, observing the network at different locations or gathering network statistics or packet samples from randomly chosen locations might be required. Since the control plane in legacy networks is co-located with the forwarding plane, changes in the monitoring system would require making changes in the devices, which are physically spread in large networks.

Traditional networks comprise of many autonomous chunks of networked systems (ASs), where a change in some parameters can induce undesirable effects on the overall network state. Moreover, there is no global visibility of the network state in current networks. This leads to localized decision-making at multiple points in large networks. Hence, synchronizing a huge number of monitoring decisions is a daunting task for both network management and monitoring systems.

Existing monitoring systems need to be adapted and correctly controlled, since they were meant mostly for physical and not virtual systems and boundaries, and do not allow fine-grained analysis adapted to the needs of SDN/NFV based 5G network management. The lack of visibility and controls on internal virtual networks that are created, coupled with the heterogeneity of devices used, make many performance assessment applications ineffective. For instance, existing security monitoring applications cannot monitor virtual connections in 5G network elements.

10.4 Use of Monitoring in 5G

Network performance and security monitoring can be viewed as complementary entities. Monitoring can provide the knowledge necessary to assess, and in consequence assure both the network's QoS/QoE (Quality of Service/Quality of Experience) and security. Network monitoring is required for the verification and validation of SLAs, managing performance (QoS) and user experience (QoE), troubleshooting, assessment of optimizations and use of resources. Detection and prevention of security breaches will enhance the performance, for example it can prevent Denial of Service (DoS) attacks. In the context of 5G, monitoring mechanisms need to be rethought to be able to deal with the requirements introduced by virtualization and profit from the flexibility obtained from SDN and NFV to obtain the best balance between costs, reliability and quality.

Future 5G networks will support extremely large amounts of devices with various capabilities and intelligence (e.g. mobile phones, tablet computers, IoT devices, tactile internet and automated vehicles). This requires automated management and security services to assure confidentiality and integrity. This will also lead to high signalling and processing costs and hence would require new strategies for cost-effective adaptive security. For this reason, it is necessary to have a clear view of what is happening in the network, the devices used and how they are used. Monitoring is instrumental for understanding the network traffic and how the services and applications are being used; enabling improved and automated security assurance.

Existing security solutions (e.g. SIEM, IDS, IPS, firewalls.) need to be adapted and correctly controlled since they were meant mostly for physical and not virtual systems and boundaries and do not allow fine-grained analysis adapted to the needs of SDN- and NFV-based 5G network management. The lack of visibility and controls on internal virtual networks created coupled with the heterogeneity of used devices make many security applications ineffective.

On the one hand, the impact of virtualization on these technologies needs to be assessed. For instance, security applications need to be able to monitor virtual connections. Virtualization can help isolate systems, but can also be used to introduce malicious techniques that exploit software vulnerabilities or introduce compromised systems that are difficult to detect. For instance, virtualization creates boundaries that could be breached by exploiting vulnerabilities and bugs in the virtualization code (e.g. hypervisors); and whole systems actually become files that can more easily be stolen or replaced.

On the other hand, the security technologies need to cope with ever-changing contexts and trade-offs between the monitoring costs and risks involved. Here, virtualization, as well as SDN, facilitates changes making it necessary for security applications to keep up with this dynamicity.

SIEM-type solutions are necessary in order to gain security and status awareness. If an incident occurs, the system should be able to determine the source, and recover and protect against it in the future. It should be verified that everything that comes out of the system is logged. Managers have centralized control over the network and it is necessary to log every change and treat it accordingly in a management solution. Log analysis and event correlation in SDN will fast become a “big data” issue. Tools also are needed that can address all the forensics and compliance requirements.

With SDN, it is possible to create network monitoring applications that collect information and make decisions based on a network-wide holistic view. This enables centralized event correlation on the network controller, and allows new ways of mitigating network faults.

To design an effective monitoring system in 5G networks, improvements are needed in the following main areas [2]:

- *Information extraction:* understanding how to deal with virtualization to obtain information on traffic flows, profiles and properties by means of extracted protocol metadata, measurements, data mining and machine learning techniques;
- *Scalability and performance issues:* the design of the monitoring architecture and the location of the observation points need to be done in such a way as to assure scalability, and different monitoring use cases need to be studied to obtain the best balance between performance, cost and completeness of the results. Furthermore, hardware acceleration and packet pre-processing technologies need to be integrated and controlled by applications and functions to obtain highly optimized solutions;
- *Heterogeneity:* analysis of different control and user plane traffic flows over the network domains and new interfaces between SDMN and existing networks and identification of related flows in different network domains;
- *Dynamicity:* changes in virtualized networks and applications become more easy and frequent. Monitoring solutions need to be able to adapt to these changes.

10.5 Software-Defined Monitoring Architecture

In order to solve the above issues, different architectural possibilities were studied and proposed in the SIGMONA [6] project. The Software Defined Monitoring (SDM) architecture was specified for 5G mobile networks. Figure 10.1 illustrates the SDM architecture for 5G Networks.

An extension of the OpenFlow type interface, referred to as the SDN/SDM Control Interface in Figure 10.1, allows obtaining the packet and flow data and meta-data

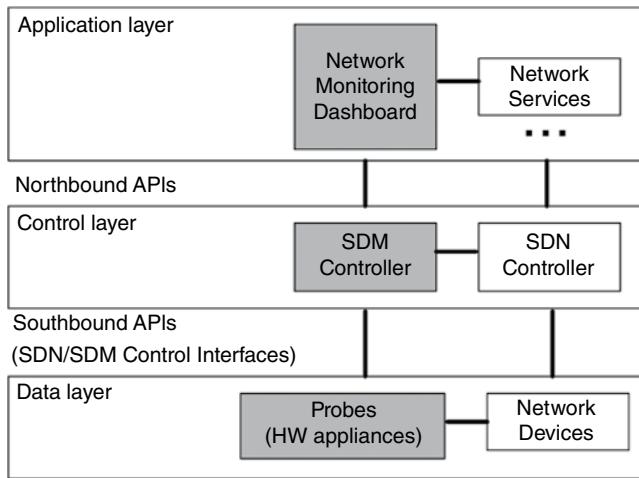


Figure 10.1 Software defined monitoring architecture for 5G Networks.

needed by the security applications (e.g. Management/Monitoring/Security, Applications and Network Services modules) from either the switches or the probes (i.e. agents). The probes can be passive (e.g. Traffic Monitoring and Analysis module that analyse mirrored traffic) or active (e.g. Active Probes acting as a firewall that filters traffic). In Figure 10.2, the SDM-CTRL acts as a controller for the software and hardware security devices and could be integrated into the SDN-CTRL or remain separate. If separate, then it will interact with the SDN-CTRL via an OpenFlow type interface. The architecture of the devices and controllers can be hierarchically organized or distributed (e.g. with peer-to-peer communications between the controllers).

A control layer based on SDN/SDM is inserted between the application and network infrastructure layers. At the network infrastructure layer, an SDN protocol, such as OpenFlow, is used as an interface. Typically, an operator first informs the Orchestrator to deploy a virtualized DPI (vDPI) application, then the application is configured by deploying the rules and/or properties that need to be detected, and finally informs the SDN controller to direct the network traffic to be analyzed to the vDPI function. Such deployed rules will allow detecting the performance properties of the connection and, at the same time, help locate performance problems and verify the conformity with Service Level Agreements (SLAs).

Figure 10.2 illustrates the deployment of the components of SDM architecture in SDN/NFV-based 5G Networks.

The added modules and interfaces of the SDM architecture are [2]:

- **Modules:**

- *Security sensor*: an active monitoring probe for the detection of security and behaviour related information (e.g. security properties and attacks) and mitigation (e.g. filtering). It can be installed on the Network Elements or in network taps (passive network observation points);
- *SDM CTRL*: a new module or extension of SDN CTRL to allow the control of the monitoring function (e.g. management of network monitoring appliances, traffic

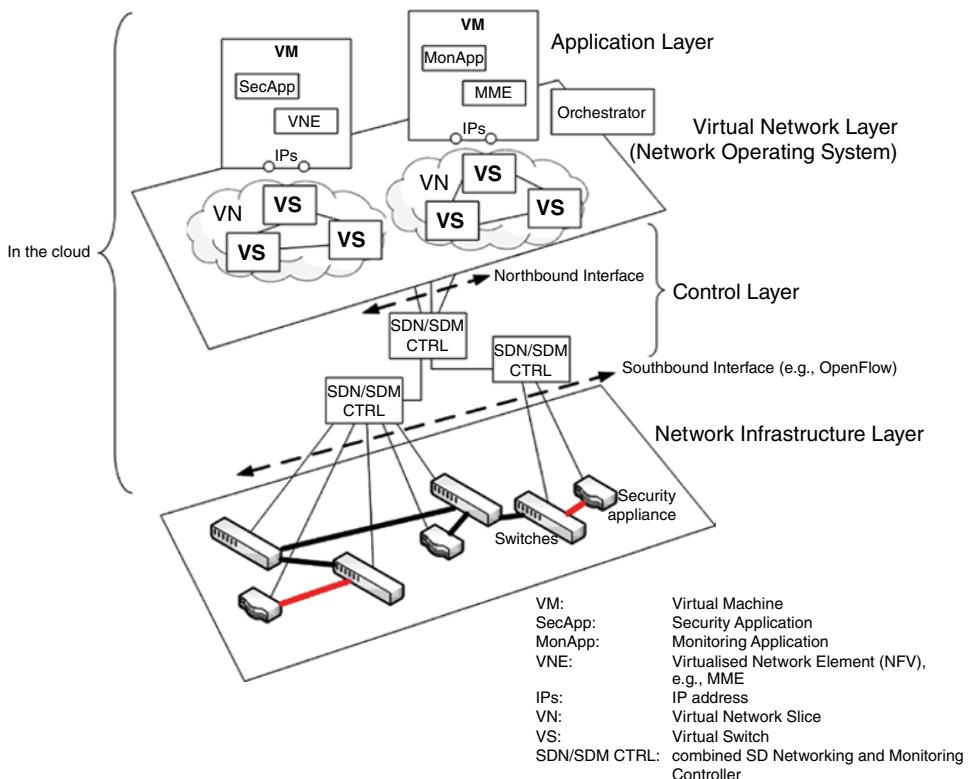


Figure 10.2 The deployment of the components of SDM architecture in a 5G Backhaul Network.

mirroring, traffic load balancing and aggregation) and accept requests from network functions and applications. SDM CTRls are distributed following either a peer-to-peer or hierarchical model. They interact with the management/monitoring/security function and act as distributed analysis or decision points for the defined security policies (i.e. security SLAs);

- *Network monitoring:* a virtualization of the monitoring function (i.e. part of the traffic analysis moved to the cloud);
- *Traffic Mirroring and Analysis:* a passive backhaul traffic monitoring device required by different network functions.

• Interfaces:

- *SDN/SDM Control Interface:* an interface that allows controlling the use of the monitoring resources, recuperating traffic or metadata for analysis. It allows performing monitoring requests and obtaining the status of the network links. In this way, applications and network functions can send requests for monitoring-based information, and receive the status information they need.

By programming flexible switches and other network devices to act as packet interception and redirection points, it becomes possible to detect and mitigate a variety of attacks. By introducing SDN-driven security analysis, or Software Defined Monitoring (SDM), SDN-enabled switches, COTS packet processing and security appliances can

act as packet brokers. Controllers can act to aggregate and correlate distributed metadata (e.g. flow and statistical data). This information can be sent to monitoring and analysis appliances and applications. In this way, it is possible to obtain adaptive and optimized monitoring, analysis and mitigation.

10.6 Expected Advantages of Software Defined Monitoring

The chart in Figure 10.3 shows the outcome of a market study by SDNCentral [7] to ascertain the benefits of network virtualization technologies as reported by different respondents. This outcome foretells an eventual change in the way networks are monitored, from dedicated monitoring systems to software-only systems integrated as part of 5G SDMNs. This change will be spurred by the perceived benefits virtualization technologies offered, for example flexibility, cost saving and scalability.

With the advent of widespread adoption of SDN, the way the network is monitored will change. Dedicated monitoring systems will eventually be replaced by software-based systems integrated as part of 5G, mainly due to the benefits virtualization technologies offer and the needs for flexible and adaptable solutions.

Programmability of networking equipment and abstracting the network complexity from applications are two of the key aspects of SDN, which are very useful for network monitoring systems. The abstractions simplify developing and deploying new network functionalities, whereas programmability of network equipment enables simple and easy-run-time deployment, which is further aided by the centralized control. Therefore, a number of SDN-based monitoring systems are being proposed and evaluated.

The limitations in legacy monitoring systems, for example, lack of flexibility, can be addressed by software defined monitoring and SDN. It can be overcome by the use of software-based monitoring systems implemented on top of the control plane. Hence, the necessary changes would require updating the software modules rather than

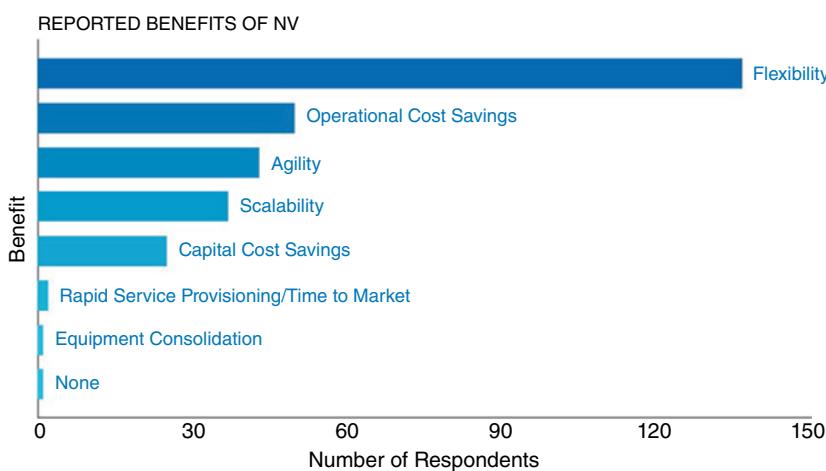


Figure 10.3 Benefits of network virtualization.

Table 10.1 Legacy monitoring techniques vs software defined monitoring.

| Legacy Monitoring Techniques | Software Defined Monitoring |
|------------------------------------------|-----------------------------------------------|
| Difficult to deploy and maintain | Simplifies network management and maintenance |
| Distributed infrastructure | Centralized control |
| Difficult to automate mitigation actions | Automates mitigation actions |
| Independent resources | Sharing of resources |
| Increase CAPEX and OPEX | Reduce CAPEX and OPEX |

making changes in a device firmware or its hardware. Since the networking devices in SDN are programmable, changes in networking parameters are comparatively simple. In SDN, centralization of the control plane would help to avoid inter-domain conflicts, since the autonomous per-domain decisions can be monitored by a centralized monitoring system implemented on top of the control plane, as presented in the previous section.

Table 10.1 presents the limitations in legacy monitoring techniques and the possible solutions proposed by Software Defined Monitoring.

Another main advantage of SDN is that it simplifies network management and facilitates the upgrade of functionality and debugging. SDN-enabled centralized control and coordination makes it possible to deliver the state and policy changes more efficiently, and deploy corrective measures more rapidly. NFV also brings advantages, since it improves scalability of applications such as QoS/QoE monitoring and by introducing virtualized abstraction where the complexity of hardware devices is hidden from the control plane and SDN applications. Furthermore, a managed network can be divided into virtual networks that share the same infrastructure, but are governed by different SLA policies. SDN and NFV make possible the sharing, aggregation and management of available resources, enables dynamical reconfiguration and changes of policy, and provides granular control of network and services through the abstraction of the underlying hardware.

SDM provides higher scalability than legacy monitoring techniques. Cloud infrastructure introduces elasticity and scalability that can benefit the monitoring tasks, but also help to improve the monitored cloud services, the resource utilization, the performance load and the capacity planning. The goal is to guarantee that the end-users have an acceptable performance with minimum resources defined by the Service Level Agreements (SLAs) negotiated between customer and provider.

SDM also allows the monitoring function to maintain high levels of visibility, evasion resistance (even if the host is compromised), and attack resistance (isolation), and even enable the manipulation of the state of virtual machines. Unfortunately, VMI (Virtual Machine Image) based monitoring software depends on the operating system, application type and versions. VMI requires privileged access, meaning that cloud providers need to authorize its access. Nevertheless, VMIs can be proposed as a cloud service by cloud providers.

To be able to perform end-to-end network monitoring, mobile operators need to define and deploy monitoring tools that will measure and analyze the network flows at different observation points that could include the devices of the end-users, as well as

the virtual and physical machines. Setting up several observation points is necessary to better diagnose the problems detected. With SDN, it is possible to create network monitoring applications that collect information and make decisions based on a network-wide holistic views. This enables centralized event correlation on the network controller, and allows new ways of mitigating network faults.

10.7 Expected Challenges in Software Defined Monitoring

The lack of visibility and controls on NFV internal virtual networks and the heterogeneity of devices make many performance assessment applications ineffective. On the one hand, the impact of virtualization on these technologies needs to be assessed. For instance, network monitoring applications need to be able to monitor virtual connections. On the other hand, these technologies need to cope with ever-changing contexts and trade-offs between the monitoring costs and the benefits involved.

In order to monitor virtualized telecommunication network, it should be possible to monitor inter-VNF communication; however, such monitoring may not be possible. Current OpenStack specifications [8] provide blueprints for the solution, but are not yet part of the release.

The OpenStack development version has introduced what is called Tap-as-a-Service. It is being developed as a plug-in of Neutron and provides an API [12]. TaaS is a project developed to provide a network service by mirroring network traffic from the port of virtual machine to another port. The main idea of TaaS is to copy each packet in/out of the virtual machine via the port and transfer to another port.

In the scenario depicted in Figure 10.4, a new virtual machine VM Monitor is created. By using TaaS, one can copy and transfer the traffic in/out of VM1 and VM2 from Port 1 and Port 2 to Port M. Then, the VM Monitor receives every packet coming in/out of VM1 and VM2. Hence, it can monitor the traffic in the Tenant without access to the compute hosts. However, TaaS only provides the mirroring service, so in order to monitor, analyze and secure the tenant, we must integrate a Monitoring Service.

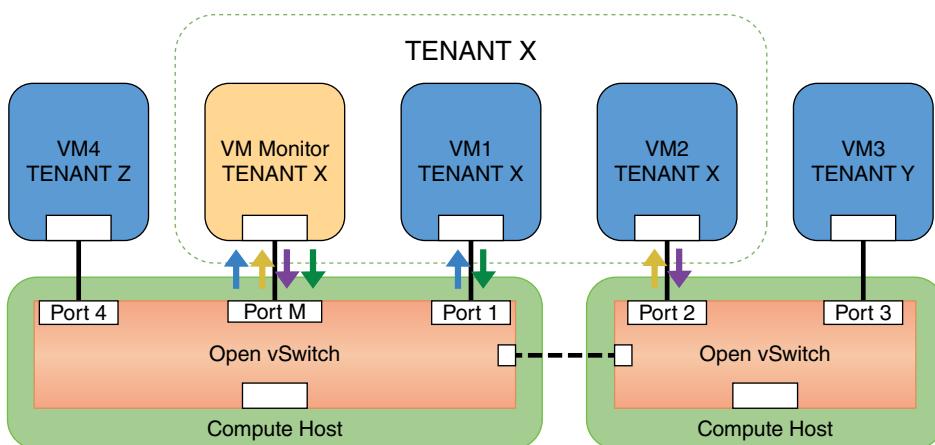


Figure 10.4 OpenStack monitoring based on TaaS.

In order to observe the whole system automatically, it is possible to have a Monitoring Service (based, i.e. on Suricata [9] or MMT [10]), which communicates with the Neutron and TaaS APIs. This service will observe and compare the ports of OpenStack via Neutron API and analyze the traffic going through the ports that are mirrored by the TaaS via the TaaS API. If a new virtual machine is created, a new port will be also created, then the service will create a mirroring service for that port by using the TaaS API. The Monitoring Service can be also implemented directly in a dedicated monitoring virtual machine, which helps simplify the deployment of the Monitoring Service in OpenStack. This technique allows monitoring of all the traffic, but then a major issue is scalability.

From the experiences gained from LTE monitoring, LTE control plane traffic of 2M subscribers is approximately 300 Mbps/300 Kpps when S1-MME, S6a and S11 interfaces are monitored. When using Intel DPDK, poll mode drivers capture to memory is not a problem. DPDK optimized packet capture is capable of millions of PPS. In-house tests show that an SW analyser is able to perform 300 Mbps analysis with 6 cores at 3 GHz, with 64 GB of maximum amount of memory required and even DPI type analysis of 10 Gbps of traffic is also possible using 16 cores (2.4 Mpps, where packets have average size equal to 600 bytes). Similar results using Suricata can be found in [13].

On the other hand, capture to disk is used to store raw packets for drilldown to be used in troubleshooting. Capture to disk performance depends on the virtual storage. The virtual storage must be selected to fulfil throughput and capacity requirements (300 Mbps). Normally, a few days of storage capacity is required. Table 10.2 provides examples of storage capacity requirements. Therefore, storage of monitoring data remains an issue in 5G networks with exponentially increasing traffic load.

In the case of SDN, network topologies are no longer as static as they were when their implementation was only physical. The SDN networks allow a very dynamic configuration of routes, filters, converters, etc. Aware of this, and taking into account the co-existence between legacy network components, software network components and virtualized network functions, it is a necessary tool that is able to show a unified view of the topology. To build this unified view, it is required that the *Network Descriptor* module could collect and normalize network data from a wide range of sources such as SDN controllers, networks emulators and legacy infrastructure (Figure 10.5).

Once the information has been collected and normalized properly, a tool is necessary to build a unified structure to allow a global visualization, the Topology Viewer. This unified structure is built by the builder component of the topology viewer matching the common field's present into the normalized information.

Table 10.2 Packet capture storage capacity.

| Throughput Mbps | storage days | total required storage TB |
|-----------------|--------------|---------------------------|
| 300 | 1 | 3.1 |
| 300 | 3 | 9.3 |
| 300 | 7 | 21.6 |

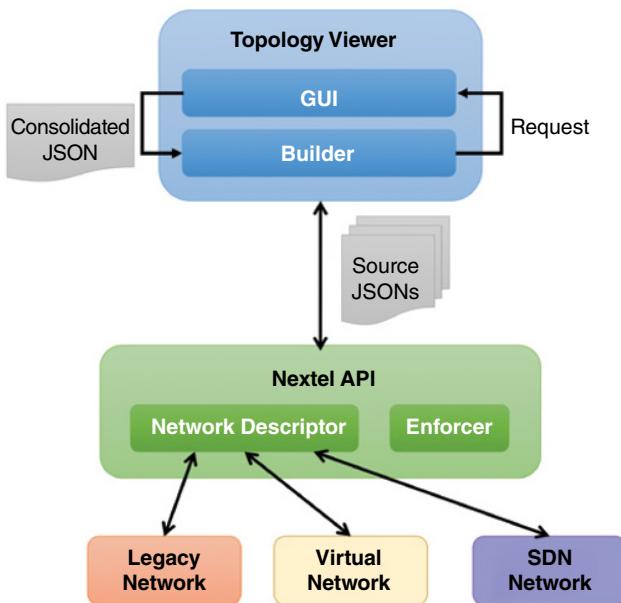


Figure 10.5 Network descriptor.

10.8 Conclusion

Leveraging SDN and NFV for 5G networks will advance the course of its evolution in many different dimensions. Mobile users are looking forward to a more user-centric network that is energy- and cost-efficient, characterized by resource sharing, optimization and dynamic control of network policies; with an overall goal of achieving higher levels of QoS and QoE for end users and more efficient service models for the providers. Network monitoring in 5G will require a more dynamic approach than we have in present-day monitoring techniques. In this respect, approaches for abstracting certain underlying hardware components of the network will become paramount to realizing a more dynamic and granular monitoring that would be much required in 5G networks. SDN and NFV techniques have been seen as promising enablers for realizing this vision. In this chapter, we have presented a description of existing monitoring solutions and their limitations. We further presented a novel monitoring scheme that was proposed in SIGMONA, which is based on SDN principles. We highlighted the proposed benefits of this approach such as high scalability and visibility, simplified network monitoring system, automated mitigation actions and lower CAPEX and OPEX. We also discussed the key limitations of this approach ranging from its inheritance of common vulnerabilities that come with typical software defined solutions, to other performance and scalability concerns.

References

- 1 Huang, F., Yang, Y. and He, L. A flow-based network monitoring framework for wireless mesh networks. *IEEE Wireless Communications*, 14, 5.
- 2 Liyanage, M., Gurtov, A. and Ylianttila, M. (eds) (2015) *Software Defined Mobile Networks (SDMN): Beyond LTE Network Architecture*. John Wiley & Sons, West Sussex, UK.
- 3 Rodriguez, J. (2015) *Fundamentals of 5G Mobile Networks*. John Wiley & Sons, West Sussex, UK.
- 4 Hu, F. (ed.) (2016) *Opportunities in 5G Networks: A Research and Development Perspective*. CRC Press, Florida, USA.
- 5 Van Adrichem, N.L.M., Doerr, C. and Kuipers, F.A. (2014) Opennetmon: network monitoring in openflow software-defined networks. *Network Operations and Management Symposium (NOMS)*, IEEE.
- 6 SIGMONA – SDN Concept in Generalized Mobile Network Architectures [Online]. Available at: <http://www.sigmona.org/>
- 7 SDX Central Webinars. *Network Virtualization Report*, 2014 Edition [Online]. Available at: <https://www.sdxcentral.com>
- 8 Blueprint: *Neutron-Services-Insertion-Chaining-Steering* [Online]. Available at: <http://stackalytics.com/report/blueprint/neutron/neutron-services-insertion-chaining-steering>
- 9 Suricata: *Open Source IDS/IPS/NSM Engine* [Online]. Available at: <https://suricata-ids.org/>
- 10 Montimage. Montimage Monitoring Tool (MMT) [Online]. Available at: <http://www.montimage.com/products.html>
- 11 Intel® Open Network Platform Server Reference Architecture (Version 1.1) [Online]. Available at: https://01.org/sites/default/files/page/intel_onp_server_release_1.1_solutions_guide_v1.1.pdf
- 12 Tap-As-A-Service API [Online]. Available at: https://github.com/openstack/tap-as-a-service/blob/master/API_REFERENCE.rst
- 13 Suricata Study [Online]. Available at: <https://github.com/pevma/SEPTun/blob/master/SEPTun.pdf>

Part III

5G Device and User Security

11

IoT Security

Mehrnoosh Monshizadeh^{1,2} and Vikramajeet Khatri¹

¹ Nokia Bell Labs, Finland

² Aalto University, Finland

11.1 Introduction

Mobile network operators should meet connectivity requirements for new applications, which will be released in coming years. Comparing to Long Term Evolution (LTE) networks, 5G will offer increased data rate, reduced end-to-end latency, and improved coverage, which are essential factors for many Internet of Things (IoT) applications such as unmanned cars, smart cities and intelligent transportation systems. Figure 11.1 shows the 5G use cases and requirements [1].

The existing technologies of cloud computing, IoT, and wireless controlled robots combine to produce a new variant called Cloud Robotics. Cloud Robotics uses the big data techniques and computing power of the cloud along with the connectivity provided by 5G, LTE and other wireless technologies to control the actions of wireless robots. We shall refer to this technology as Mobile Cloud Robot (MCR), where the robot-cloud connectivity is provided by mobile networks. This technology has the mutual attention of telecommunication vendors as well as different industries, including manufacturing, medicine and agriculture.

Because of the operational characteristics of resource sharing and a centralized controller, MCR can act as a potential cost saver. On the other hand, with big data techniques such as data mining and knowledge reuse, its efficiency and security could be improved. Using all computing power has risks and it is critically important to secure the communications path and efficiently utilize the wireless bandwidth. Big data and the use of data intelligence can be applied to MCR to provide both services and security support to an MCR network. Examples such as fault detection, service issue resolution, network segmentation and security are common. For this purpose, we propose a distributed security platform including robots, Local Robot Controller (LRC), mobile cloud, IoT anomaly detection module and IoT orchestrator. Here, we aim to reach high efficiency from different perspectives to perform data analysis that will identify improvements for an MCR environment, for example processing time, data analysis accuracy, energy saving for different applications such as robot programming, monitoring, security and fault management [2].

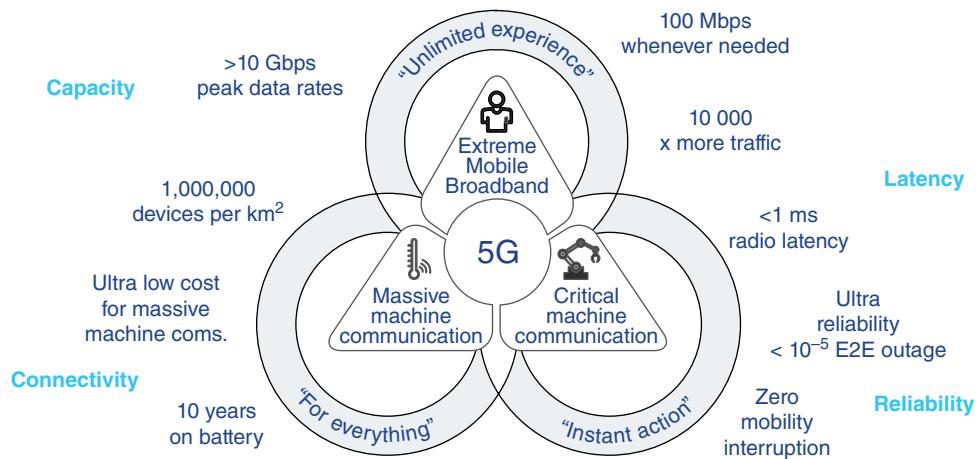


Figure 11.1 5G use cases and requirements [1].

The rest of this chapter is organized as follows. Section 11.2 briefly reviews related work. Section 11.3 discusses IoT services, IoT threats, MCR and its security challenges. Section 11.4 discusses the distributed security architecture for MCRs that detects anomalies and distributes preventive actions to mitigate identified threats. Further, in Section 11.5, the data classification and connectivity scenarios for security platform are discussed. Finally, the conclusion is presented in the last section.

11.2 Related Work

Ian *et al.* [3] have proposed an automated drone security system for surveillance purposes, where on-board sensors and an imaging device are used to capture surveillance data. Drones execute flight operation, store surveillance data and connect to a server for location update and transmission of encrypted surveillance data. In this study, Global Positioning System (GPS) is used for positioning. A user device is connected to a server to receive and display surveillance data. The flight operation of a drone is controlled by a server or a user device. It also covers drone docks that are used for launching, landing and charging drones, as drones are battery powered. Their proposed architecture does not address authentication hijacking scenarios for drones and the server. A malicious drone may get authenticated to a server; or a server may get hacked. In such cases, an attacker gains access to data and misguides the drones for surveillance.

Daniel [4] proposed a system for a drone docking station to deliver goods. A drone resides at a docking station and goods from a storage facility are attached to the drone for delivery. The communication between the drone docking station and the drone is via GPS, Wi-Fi, Bluetooth and satellites. The drone can be traced via GPS; however, its flight path cannot be controlled. When the drone reaches the docking station, it may inform the next delivery to the storage facility via Wi-Fi or Bluetooth. Then the storage facility exchanges information about the next delivery, such as what and where to deliver.

However, the proposed system is only about flight location update and not applicable to control the drone on the flight.

Shriram *et al.* [5] have discussed drone delivery assurance, where a drone delivers a package of goods to a destination. Here, assurance refers delivering to a correct destination after verifying the recipient and its location while taking pictures as evidence. When a drone arrives near the delivery location, a notification is sent to the receiver. The receiver then sends the goods purchase code to the drone using Wi-Fi or Bluetooth. The drone receives the purchase code, authenticates it and after successful authentication, lands and delivers the goods. In case of failed authentication on the purchase code, the drone will not land. Global Navigation Satellite System (GNSS) is used for positioning purposes. In addition, cellular communication is used for establishing connection with other devices. The proposed architecture lacks a security mechanism to verify or authorize the drone; a malicious drone may arrive and take the goods using a fake purchase code to authenticate.

Anthony *et al.* [6] proposed a system and a method for managing communications in robot competitions. The communication takes place over a wireless network between a network arena controller and a robot controller. The network arena controller provides security keys and firewall policies to robot controllers; and robot controllers execute those firewall policies to secure communications between the brand or operator stations and robots. The network arena controller may also monitor and log communication traffic to verify connectivity and monitor battery level, signal strength and robot status. The brand or operator stations may include any software, hardware and/or firmware, which are configured to monitor and control robots associated to it. These stations give commands to robots to perform an action. The proposed system prevents one or more intentional or unintentional security risks, such as Denial of Service (DoS) attack and spoofing attack via bandwidth monitoring. Upon detection of intentional flooding, the system either imposes bandwidth limitation for the robot or totally blocks the robot using a firewall.

Studies are mostly focused on robot applications, such as automated drone security system for surveillance, drone delivery assurance, and drone communication management with security keys and firewall policies. Yet there has not been specific research that concentrates on MCR security for Mobile Virtual Network Operators (MVNOs); none of the current studies in the relevant area have utilized data-mining techniques for security threat analysis. This study applies data-mining mechanisms in an orchestrated security platform to detect intrusion in robots belonging to MVNOs. The proposed platform could detect and prevent different types of attacks on both robot and LRC, which are explained in Section 11.4.

11.3 Literature Overview and Research Motivation

With an increased data rate, larger number of users, lower latency, higher reliability and wide network coverage, 5G network would be promising to meet the needs of mobile communication systems and end users. The 5G network would be capable of several Gigabit per second (Gbps) per user data rate to satisfy growing demand for ultra-HD and 3D video content [7]. However, there are still devices that need lower bandwidths, such as remote health monitoring. In remote health monitoring, patients have wearable

devices and sensors to collect and send patient information periodically to a hospital server or cloud. The devices are sending patient information with low data rate; therefore, with using non-orthogonal multiple access mechanism, several of these IoT devices can be squeezed into the same time slot [8].

In addition, it is expected that the total number of mobile connections, including IoT or machine type communication, reaches several hundred times the world population. 5G would serve many IoT devices across multiple domains, for example, self-driving cars, package delivery, medicine, smart grids and home automation. 5G network will be heterogenous and will have many new technologies and services incorporated into it. Therefore, security risk assessment must be considered by examining security threats and their mitigation mechanisms before and after their integration into 5G network. The security threats arising on the 5G network can be divided into following two categories:

- 1) *Security threats arising from earlier generations of mobile networks*: since a 5G network will be a heterogenous network, the known threats from legacy mobile network architecture will be part of 5G network. Therefore, a revised security architecture is needed to resist and retaliate to such threats;
- 2) *Security threats arising from new technologies to the 5G network*: these security threats should be analyzed and their mitigation mechanisms should be applied to 5G security architecture [7].

IoT applications can be classified into two categories:

- 1) *Monitoring based IoT applications*: these applications collect data from connected sensors or devices periodically and transmit it to the cloud. Examples include home automation, patient monitoring and smart metering. These applications also offer remote monitoring and data analytics;
- 2) *Control-oriented IoT applications*: these applications use the sensor data to control the connected actuators in real-time. Examples include self-driving cars, industrial robots and remote surgery. Depending on the use case, the latency, reliability and availability requirements can vary [9].

11.3.1 IoT Devices, Services and Attacks on Them

The demand for IoT connectivity will be increased with 5G, with estimates of up to 46 billion IoT connected devices by 2020 [10], for applications such as smart cities, smart environment, smart utilities, consumer market, logistics, industry 4.0, smart agriculture, home automation and ehealth. On the other hand, the Internet of hacked Things is also on the rise; as we connect more devices, and more value is created from the data generated, the risk for abuse and security breaches increases. In October 2014, millions of smart meters in Spain were found to be vulnerable [11]. In February 2015, 2.2 million BMWs were impacted by a bug in ConnectedDrive software, which allowed remote unlocking of cars [12]. In July 2015, 1.4 million Chryslers had been recalled due to vulnerability in Uconnect dashboard computers, which allowed hackers to control dashboard functions, steering, transmission and brakes for Chryslers [13]. Some of the 5G IoT use cases are illustrated in Figure 11.2 [14] and IoT services provided in different areas, IoT devices used and attacks on such services are described here [15]:

- 1) *Public safety*: while services such as surveillance, tracking and emergency management will be managed by law enforcement and the military, yet attacks on CCTV cameras and speeding trackers will compromise public safety;
- 2) *Digital health*: services include preventive health, patient care, remote surgery, connected hospitals and health workers. The patient will have wearable IoT devices such as a wrist watch and sensor that will monitor their health parameters and send them to hospital if required. The system at the hospital will monitor, detect health problems at an early stage and alert the hospital to take the next steps. An attack on such devices and services could be stealing patient health information which violates privacy; and modification in sensor data that will compromise the patient's health, for example parameters for a healthy patient may be changed as unhealthy parameters or vice-versa. An attack on remote surgery platform will have a critical impact on patient health, if surgery is not properly instructed;
- 3) *Mobility*: services will offer connectivity between transport vehicles, including connected cars, connected trains, connected ships and connected airplanes. Self-driving cars also fall into this domain. IoT devices will be embedded within transport vehicles that will interact with other peers and services will reduce accidents. An attack on such services may cause a transportation delay or failure, if attackers exploit one of the vehicles or the platform and issue misleading directions, commands or generate fake warning messages. Apart from this, if the cars are under control of an attacker, an attacker can spy on the car owner and the vehicle may also be used for illegal purposes;
- 4) *Industries*: services include smart production, smart agriculture and smart buildings. Industries have a lot of automation, and sensors are utilized to enhance production. In smart agriculture, sensors can be used for diagnosing water levels, humidity, need for pesticides, harvesting times, etc. An attack on these services will impact their production;
- 5) *Smart cities*: services include smart infrastructure, traffic, parking, tolls, air quality monitoring, tourism, virtual reality and advertising. Smart parking services will help a driver to find a free parking slot nearby and tolls will collect toll charges automatically, for example, with car number plate scanning or smart payment mechanisms. Tourism can be improved with virtual reality (VR), which will enhance customer experience. An attack on smart cities' infrastructure can have big impact, as tolls will not be functional, air quality will not be monitored properly and tourism revenues will be impacted because of VR and advertising failure;
- 6) *Utilities*: services include smart grid, smart metering, water and waste management. An attack on the smart grid may impact the electricity distribution and the underlying areas may experience blackouts due to electricity failure. An attack on smart metering can have significant impact on revenue and customer service, as an attacker may compromise the smart meter to report less or excess consumption to energy providers;
- 7) *Smart homes*: services include smart consumer, lighting, temperature, music and video. IoT devices will consist of various sensors and electrical appliances, such as temperature sensor, fire sensor, humidity sensor, toaster, refrigerator, coffee machine, etc. If coffee beans are running out, the coffee machine will place an order for beans and will work as a smart consumer. IoT devices will report their activity and would be configurable as to when to execute an action such as start, stop, order, status

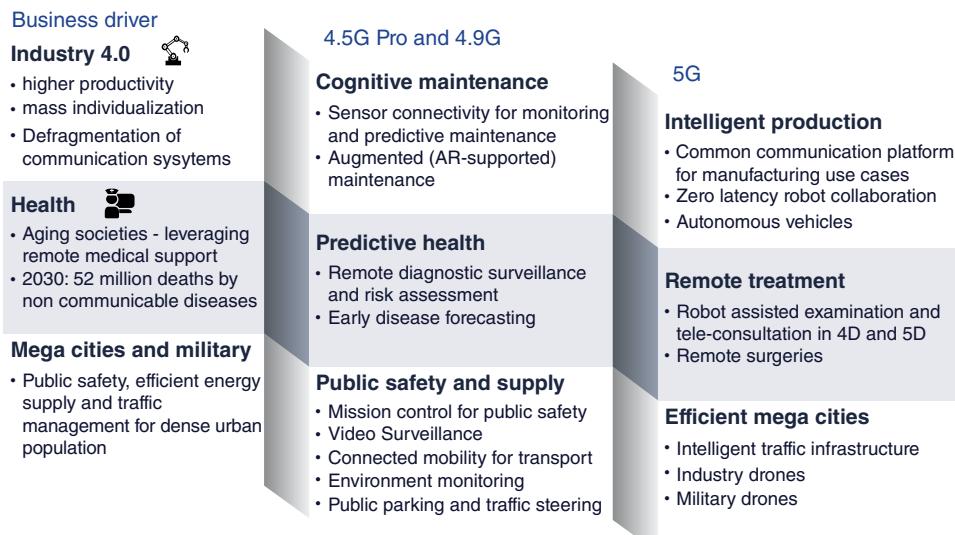


Figure 11.2 5G IoT use case evolution.

- update, etc. An attacker can exploit IoT devices and place orders, modify their functionality, modify their operations or inject a malicious code and devices that can act as a bot and participate in carrying out a distributed DoS (DDoS) attack;
- 8) *Retail sector:* services in the retail sector are smart payment and smart stores. The stores will be automated and tailored to meet customers' demand based on consumption (collected analytics) and will offer smart payment systems. The stores will have the benefit of satisfied customer service. The potential target for attacker will be smart payment systems for financial gains. The attack on smart stores will impact availability of products, as either there would be no products or more products than customer demand.

IoT devices will be deployed in both low-priority and critical applications, such as military, emergency services, remote surgery in healthcare, self-driving cars and many others. In general, the attack on IoT service will start with an IoT device initially. IoT devices may be in remote areas without any human supervision and may be stolen to modify their functions or steal their information to learn vulnerabilities. The attacker may exploit vulnerabilities, inject malicious code or steal data, which will result in full or partial loss of services or revenues for end users or networks. If an attacker makes multiple IoT devices vulnerable, they can be used to launch a DDoS attack, like the recent Mirai botnet attack [16]. User awareness is another important element; typically default passwords are not changed and the attacker can easily enter IoT devices with default passwords. Therefore, the IoT devices must have strong authentication mechanisms to prevent unauthorized access.

The IoT threats, as mentioned in Figure 11.3, include service disruption, information theft, device theft, system manipulation and gaining control of the system. Service can be disrupted by disabling IoT devices or carrying out a DDoS attack on the service platform. Information theft happens via vulnerable IoT devices or stolen IoT devices.



Figure 11.3 IoT threats [17].

If the system is manipulated, it may provide incorrect sensor information (data), which will impact analytics significantly. If an attacker gains control of a system, IoT devices operating under that system may become a member of a botnet and perform operations including email spams, carrying out DDoS attack, bitcoin mining, click fraud, or gain access to information and issue misleading commands.

11.3.2 Research Motivation

Diverse industries have been using sensors and robots for years, but their control systems often remain deliberately isolated to avoid any attacks. Performing data mining on data collected from sensors and robots in the industrial environment helps to detect the anomalies before they occur. The collected data should be made available across each site to perform data mining and classification. Therefore, we introduce a mobile network cloud platform that collects data from IoT robots, and performs data mining and data analysis to find malicious robots and isolate them.

In order to perform efficient intrusion detection, we introduce a platform comprising of two main parts: LRC and IoT service provider that is a MVNO. Once the attack is detected in a MVNO, an IoT orchestrator is used to distribute the malicious pattern and information on a malicious robot to other MVNOs, so they can also retaliate to the malicious intent without re-analyzing the same kind of robots. This capability enhances the scalability of the proposed platform and makes it an efficient distributed security framework. Therefore, the proposed platform is also a distributed mitigation strategy for malicious robots to prevent its malicious intent to all nearby MVNOs. LRC would directly interact with all connected robots for various functions such as local privacy, security and monitoring, local hardware control, local software manager and local application, etc.; while in the proposed architecture, LRC concentrates only on local security and monitoring.

Apart from robots that are working in an industrial environment, the flying robots, so-called drones, are increasingly used in the service industry, for example, postal delivery, vaccine delivery, remote surveillance and inspection of terrain for radio network planning, etc. The regulatory authorities have been allocating bands for their operation, but who controls where the drones can fly, what they can do and what forbids them are

yet challenges to be addressed. In addition, if a drone breaks down while performing an assigned task such as delivering vaccines, the control platform for drones should guide another drone to replace it and recover the failed drone so it is not stolen to be used later to gain unauthorized access to the platform of the drone. One can attack, gain unauthorized access, and divert the mobile robot to an improper location or steal the data collected. Therefore, security for such an environment (MCR) is extremely important. However, the overall scope of the proposed architecture is on security-related functions, which covers three main security requirements: Authentication, Authorization, Accounting (AAA), and integrity and availability of data. To fulfil these security requirements, the data should be collected and sent from LRC to a service provider for further analysis.

11.4 Distributed Security Platform

Various communication systems may benefit from an improved security. For example, communication systems that include machine-type communications may benefit from improved security and fault detection. This study proposes a platform for enhancing robots and MVNO security, where robots refer to both fixed robots and drones. The platform includes detecting an attack based on the robot information and determining a preventive action. Furthermore, the architecture includes sending an indication of an attack to LRCs or MVNOs via an IoT orchestrator. An MVNO receives data at an IoT anomaly detection module from an LRC. The data includes robot user plane and control plane information.

11.4.1 Robot Data Classification

Collected data from robots could be on different protocols, such as HyperText Transfer Protocol (HTTP) and User Datagram Protocol (UDP). The chosen protocol is robot vendor specific and based on the robot's application. In particular, a model could be tuned based on robot vendor proprietary protocols when their protocol specifications are known. But, in general, and for all protocols, data is categorized into two groups: control plane data and user plane data.

Control plane data includes signaling data such as Non-Access Stratum (NAS), authentication-security, bearer request, paging, location area update, etc. User plane data covers telemetry and command-control data. Telemetry data is related to application of the robot and geographical location, for example, images, video, measurements (temperature, speed, etc.) and HTTP traffic. Command-control data is related to control of the robot such as status check, software update and task management. All user plane data should be normalized before being processed for anomaly detection.

MVNOs can deny access to their networks for vendors that do not open their data specifications to the operator and export only normalized data using a commonly agreed data model, thus hiding vendor proprietary protocol design from other operators and vendors. However, legislation mandates that vendor protocols are open as a perquisite for approval for license to enter markets. Another way for tuning the model to process different types of robots' data is reverse engineering the vendor protocols, in case the specifications are not communicated.

11.4.2 Robot Attack Classification

The victim is either a mobile network or a robot. In case a robot attacks another robot, as shown in Figures 11.4 and 11.5, the attack can be done either through the LRC, where a malicious robot and a target victim robot belong to the same brand, or they belong to different brands and the attack is issued through MVNOs. Examples of attacks include corrupting and tampering other robots' data and mission, stealing other robots' data, and overloading other robots with consecutive requests. The attacks on mobile network include any kinds of malicious behaviors that cause unavailability or degradation of network services or stealing network information. A concrete example is the DoS attack on Mobility Management Entity (MME) through a burst of signaling messages (paging, status transition, etc.), configuration corruption, and stealing Home Subscriber Server (HSS) information.

In Table 11.1, different types of attacks that may target robots and LRC are classified based on the three main security requirements [18–19].

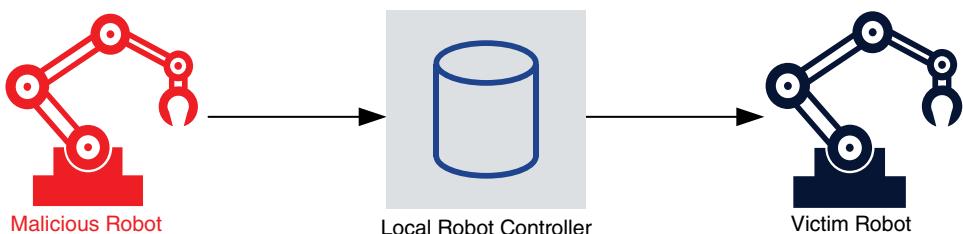


Figure 11.4 Attack scenario for robots belonging to a same brand.

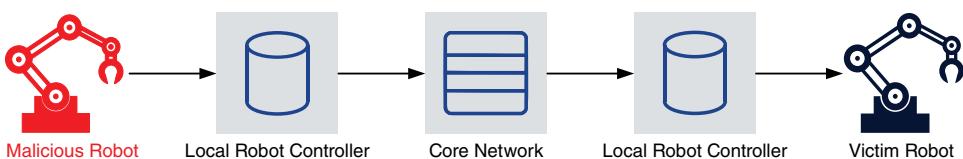


Figure 11.5 Attack scenario for robots belonging to different brands.

Table 11.1 Threats classification.

| AAA | Availability | Integrity |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Unauthorized access and privileged access: <ul style="list-style-type: none"> ● Probing ● Remote access ● Man-in-the middle ● IP spoofing ● Spyware ● Service injection | Data loss and resources unavailability: <ul style="list-style-type: none"> ● Unexpected system failure ● DoS ● Image loss ● Configuration loss ● Misconfiguration | Data corruption, tampering and leakage: <ul style="list-style-type: none"> ● Botnet ● Malware ● Application corruption ● Ransomware |

Some of the threats listed in Table 11.1 are described here:

- *Probing*: an attempt to monitor a robot and steal information such as open ports and IP addresses of robots connected to the network;
- *Remote access*: this threat tries to access the robot without permission. The access is made possible by exploiting a vulnerability;
- *Man-in-the middle*: occurs when an attacker gains access to the communication channel established between robot and LRC, or between two robots. The attacker can perform unauthorized activities, such as intercepting robot data or modifying communications to change the robot mission;
- *IP spoofing*: in this attack, an attacker creates IP packets with a false source IP address to hide his/her identity, to be introduced himself/herself as another robot and steal the data;
- *Spyware*: all kinds of software that monitors and steal robot's information;
- *Service injection*: the attacker targets a robot service and injects malicious code to corrupt the service;
- *Botnet*: a group of connected vulnerable robots in a network, which are remotely controlled by a master computer (hacker). Like robots, they automatically perform some functions that are predefined by the botmaster and forward the information like viruses to target LRCs or robots, which could cause denial of service;
- *Malware*: refers to all kinds of software codes, i.e. viruses, worms, Trojans. These attacks are programmed to perform malicious operations on a robot;
- *Ransomware*: any kind of software that locks a robot and demands some form of payment to make the robot unlock;
- *Denial of Service (DoS)*: an attacker occupies the network resources by flooding it with consecutive requests and cause denial of services to robots;
- *Unavailability of robot or LRC*: based on unexpected failure or any attack, robot's and LRC's firmware or configuration could be lost or corrupted and make the robot unavailable.

11.4.3 Robot Security Platform

As shown in Figure 11.6, the proposed Internet of the robot security platform has two sections: robot section and mobile network section.

The robot section includes robots and Local Robot Controller (LRC). LRC is a central node for all robots and responsible for connectivity between robots and the mobile network. The first level of data monitoring and security analysis is done in LRC and before data is forwarded to mobile network. The IoT service provider can be either a mobile network operator or a Mobile Virtual Network Operator (MVNO). The MVNO carries out advanced analysis of collected data from various LRCs, through data mining mechanisms in an IoT anomaly detection module. The analyzed data would be used for intrusion detection in the same or different MVNO. While a robot has been identified as malicious in an MVNO, a notification will be sent to all MVNOs by an IoT orchestrator to block the malicious robot and prevent re-analysis of the malicious robot in future, since the malicious robot may try to attack the cloud platform using other MVNOs. Therefore, the distributed security platform reduces the computation processing for a malicious robot that has been already identified as malicious in another MVNO.

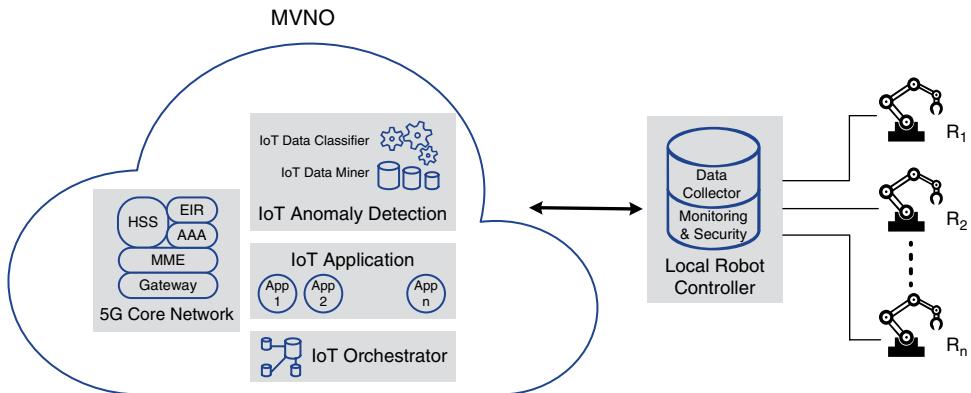


Figure 11.6 Internet of robot security platform.

11.4.3.1 Robot Section

The robot section includes robot, LRC, Local Security & Monitoring, and Data Collector, which are described below:

- a robot with connectivity to a mobile network through an LRC;
- LRC will be a central node for all robots and sensors present in an industrial environment. LRC will be responsible for hardware, software and security policy for all robots and sensors connected to it. The hardware domain includes controlling engine and its power, whereas the software domain includes drivers and its applications:
 - *Local Security & Monitoring*: all hardware and software Key Performance Indicators (KPIs), as well as some other parameters of interest for all robots and sensors connected in an industrial environment, will be monitored via a module called Local Security & Monitoring. This module will send the data to a node called Data Collector;
 - *Data Collector*: will act as a gateway for the industrial environment, collect the data from the modules connected to it and continuously communicate with mobile network for data mining and classification.

In the proposed architecture, there is no direct communication between robots. If robots belong to the same brand, LRC will take care of trust management among the robots and if they belong to different brands, then the mobile network takes care of trust management (by correlating information between two LRCs).

11.4.3.2 Mobile Network Section

IoT anomaly detection module compromises of IoT Data Miner and IoT Data Classifier:

- *IoT Data Miner*: the collected data will be heterogeneous in nature, that is, data will be from different IoT devices and with different profiles. Therefore, the collected data should be properly combined, processed and correlated. The IoT Data Miner performs data mining on the collected data, and labels the collected data in an efficient way so that it can be easily classified by the IoT Data Classifier. The collected data is normalized in this module;

- *IoT Data Classifier*: big data algorithms are applied to label the data. Data classification refers to association of collected data to different classes based on the purpose of anomaly detection. The classes are defined based on different features. The features are either predefined or extracted based on training data or they would be dynamically defined during analysis processing.

In the IoT anomaly detection module, different linear and learning algorithms are combined in a wide range to investigate a hybrid model for achieving high detection performance and yet relatively low detection time, and will be deployed to detect intrusion attempts on IoT robots. The anomaly detection could be either for attack detection or other applications, such as traffic safety, etc. In principle, it would be possible to use the described data-mining mechanisms to detect all types of anomalies and to feed the information, for example to Public Warning System (PWS) [20], to send a warning as a cell broadcast to alarm any robots in an affected area of a potential threat.

As shown in Figure 11.7, robot traffic, after being analyzed in the local robot controller, reaches the IoT anomaly detection module in the related MVNO's cloud. At first, a vulnerable protocol, such as HTTP, would be filtered and packets carried over vulnerable protocols would be sent to the next module for analysis. In this stage, traffic would be clustered and suitable features would be extracted in order to label the attack. Through an error check mechanism, the actual result would be compared with the expected result to evaluate the detection accuracy. As malicious traffic is detected, the mitigation mechanism would be applied to block the malicious robot and inform the rest of the network.

- *IoT Application*: refers to an application layer and provides interfaces for various IoT domains, i.e. such as smart parking, smart home automation and security;
- *IoT Orchestrator*: distributes the information about malicious robots to MVNOs using orchestration module. Once the attack is detected, the prevention action should be performed. Prevention includes sending information to Local Robot Collector (LRC) to identify the malicious robot. LRC traces back the malicious robot and blocks it from authenticating to MCR. The above-mentioned mitigation strategy can be safely applied if the malicious robot is authenticated either to the same IoT service provider (MVNO) or other cloud service providers (MVNOs). Therefore, an IoT



Figure 11.7 Detection dimensions.

orchestrator is used to distribute the malicious patterns and information on malicious robots to all cloud service providers (MVNOs), so they can also retaliate to the malicious intent. This enhances the scalability of the proposed platform and makes it a secure distributed platform.

11.5 Mobile Cloud Robot Security Scenarios

Machine-type communication is increasingly important to both government entities and enterprises, where the private LTE networks are used to support the application of drones and robots for military, transportation and energy industry services. Government entities deploy private LTE networks for emergency and strategic situations, where an anomaly in the original function means the difference between life and death, such as in military, national security and emergency services. Therefore, a secure communication is a major consideration for governments as well as some enterprise entities, such as energy, transportation, etc. [20–22].

In this chapter, we propose an orchestrated security platform for Internet of robots that is applicable to entities that use private LTE networks as well as commercial service providers. On the other hand, based on the robot connectivity to the network, two scenarios: robot-with-SIMcard and robot-without-SIMcard, are considered for the proposed platform.

11.5.1 Robot with SIMcard

As shown in Figure 11.8, we consider “ $n*m$ ” robots, which belong to “ n ” different brands, while robots belonging to brand 1 and brand 2 are connected to MVNO1 and robots belonging to brand “ n ” are connected to MVNO2. In this scenario, each robot has a Subscriber Identity Module (SIM) card. If robots are belonging to a commercial network (and not a private LTE network), LRC is installed at Base Transceiver Station (BTS) nodes for cost savings.

As shown in Figure 11.9, when a robot tries to connect to a mobile network, first the robot is authenticated internally by LRC and information is analyzed. Based on the policies or rules that are locally defined in the LRC, the traffic would be either dropped or passed to related MVNO. If the traffic is dropped at LRC, a notification message would be sent to the parent MVNO. Later parent MVNOs will inform other MVNOs about malicious robot through an IoT orchestrator. If the traffic is safe and traffic arrives at the MVNO1, an IoT anomaly detection module in MVNO1 will analyze whether the traffic is safe or not.

If the traffic is not safe, it will be dropped and a message sent to associated LRC, and through an associated IoT orchestrator, to other MVNOs. If the traffic is labeled safe by an IoT anomaly detection module, the MVNO will forward the robot attach request to MME and HSS/AAA for the authentication procedure. After the robot is authenticated, the robot User Plane (UP) data would be forwarded to the mobile network for further analysis. The same anomaly detection procedure will be applied to UP data, and safe traffic would be forwarded to mobile network. The above-mentioned procedure is illustrated in Figure 11.10.

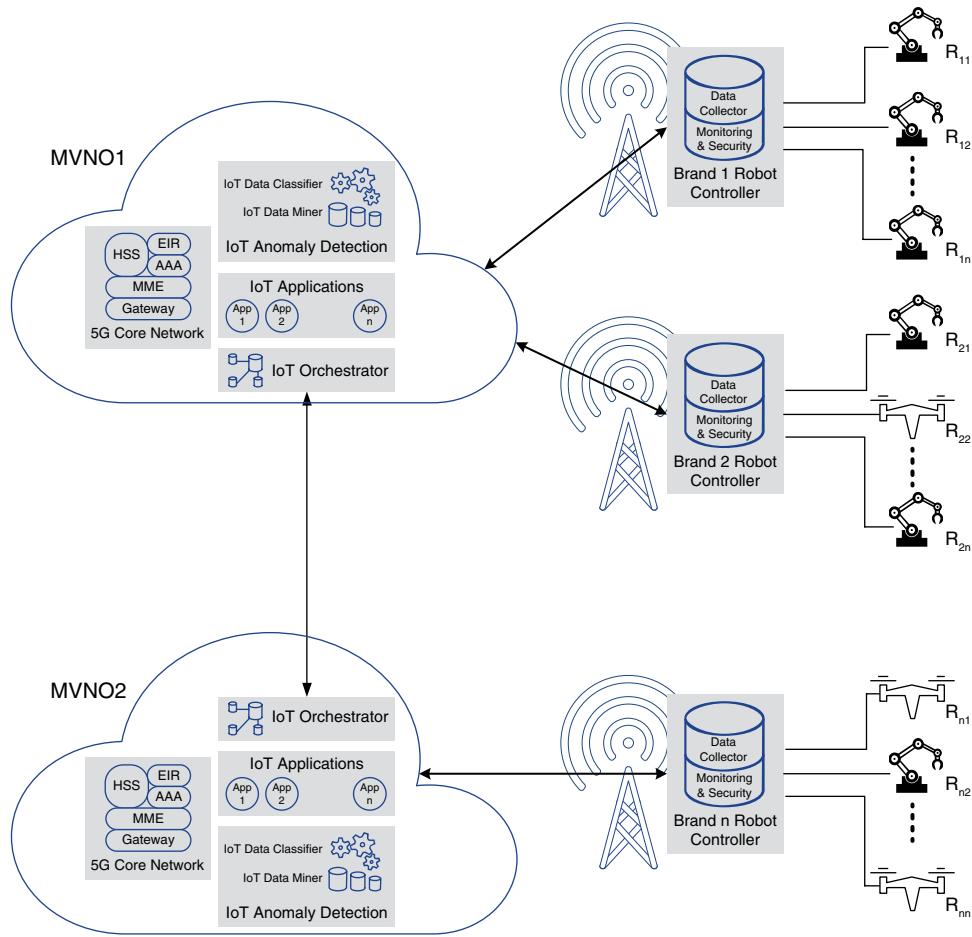


Figure 11.8 Distributed security platform for robot with SIM.

11.5.2 SIMless Robot

In the SIMless robot scenario, robots do not have any SIM cards; they are connected to LRC via Wi-Fi, and LRC has a SIM card to connect to the mobile network. LRC acts as a gateway for all the robots connected to it. The communication via the mobile network and anomaly detection procedures are as in the above-mentioned scenario.

For a drone communication case, as shown in the Figure 11.11, each cell may have multiple LRCs that are either fixed at a location (Private LTE) or integrated into a BTS. LRCs are authenticated with the nearest BTS of the mobile network operator in that area. The LRC is the primary contact for drones operating in the area; and every drone will be first authenticated to the LRC in the associated cell of an area. After authentication with the first LRC, the drone will be assigned with a task. In addition to task assignment, task cancellation, and task replacement processes, the LRC periodically collects information from drones and sends it to BTS. The received information will be forwarded to an IoT anomaly detection module for further analysis. Each LRC has

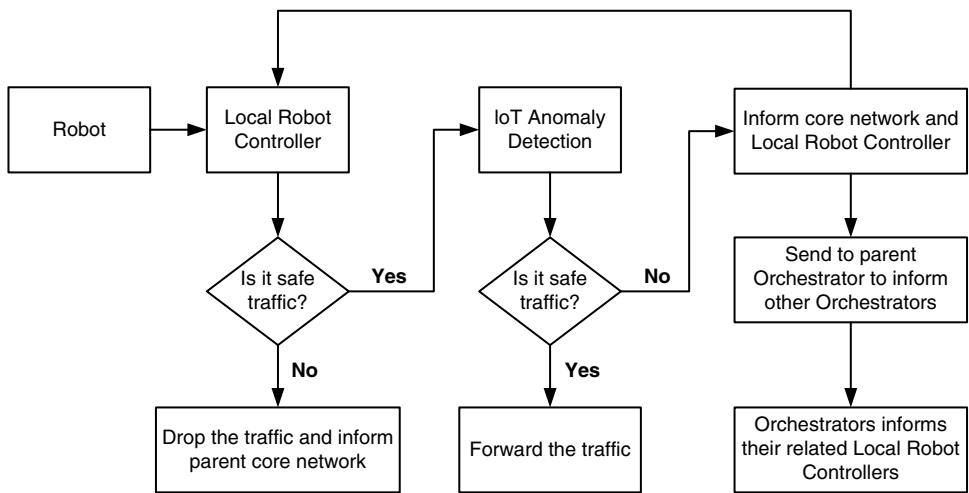


Figure 11.9 Robot anomaly detection and mitigation procedure.

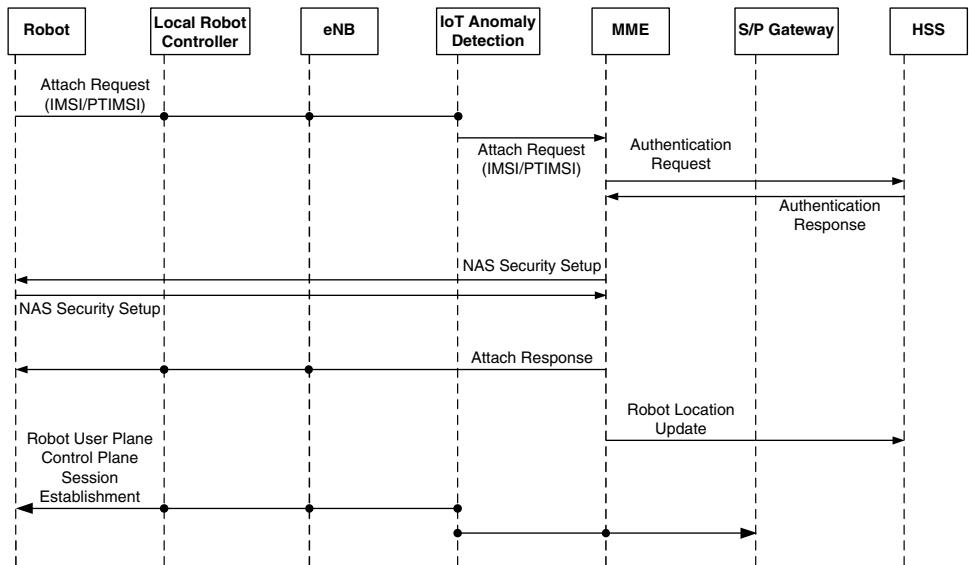


Figure 11.10 Robot authentication and sessions establishment.

a coverage area and while a drone moves from one area to another, the new LRC will be responsible for connectivity and related procedures.

If the IoT anomaly detection module detects a suspicious pattern in the data received from any drone, it will inform the associated LRC to initiate a mitigation strategy. The mitigation strategy starts with cancellation of the task, triggering automatic clean-up of malwares or malicious content. If the automatic clean-up activity fails, LRC sends a message to the service center to perform manual clean-up and de-authenticate malicious drones from the network. In case the suspicious drone does not belong to an

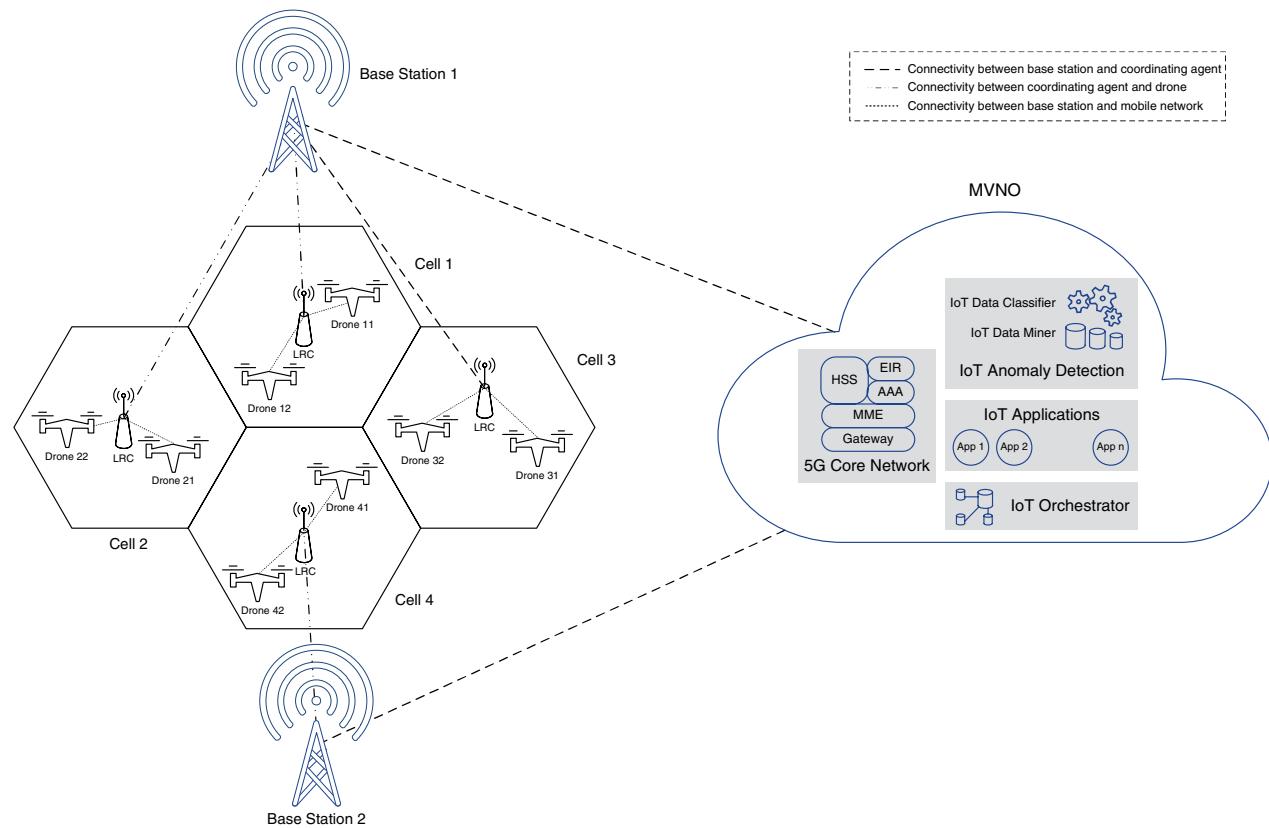


Figure 11.11 Distributed security platform for SIMless robots.

associated MVNO, it will be de-authenticated, blacklisted and a blacklist will be updated to all MVNOs and LRCs in the network. Afterwards, that malicious drone will not be authenticated anymore to harm the services or network elements.

11.5.3 Robot Attack

Figure 11.12 shows robot attack scenarios, which were discussed in Section 11.4. The attack detection considering the whole MVNO has been illustrated here. Consider there are 3 drones; drone 2 is an attacker, whereas drone 1 and drone 3 are victims. Drone 1 and drone 2 are in one cell (cell 1) and belong to the same LRC, but drone 3 belongs to a different cell (cell 2) and belongs to a different LRC. Drone 2 can attack in two ways:

- 1) drone 2 may either attack drone 1 directly using LRC; or
- 2) drone 2 may attack drone 3 indirectly via MVNO.

11.5.4 Robot Communication

There are three types of communication between robots:

- 1) *Intra MVNO communication*: a robot communicates to other robots inside its MVNO;
- 2) *Inter MVNO communication*: a robot communicates to other robots inside other MVNOs;
- 3) *Roaming between MVNOs*: a robot moves to different MVNOs and tries to communicate to other robots. As shown in the Figure 11.13, a robot has been arriving earlier at a certain location and later has been physically moved to another location. The authentication request will be to send the robot to the nearby LRC of the service provider.

A robot sends a handover request through LRC and the source base station, for example an evolved Node B (eNB) to an IoT anomaly detection module. If the received data is safe, the IoT anomaly detection module will forward the request to a source MME. Then a source MME sends the relocation request to a target MME, and a source MME receives a relocation response. The source MME returns the handover response to the source eNB, which then forwards the handover response to the robot through LRC. The source eNB sends an eNB status transfer notification to source MME, which indicates to source MME of the robot handover to target eNB. Target MME then informs target eNB of an MME status transfer and a handover confirmation will then be sent to the target eNB via LRC.

11.6 Conclusion

In this chapter, we reviewed 5G network features for machine type communication, Internet of Things (IoT) services, IoT threats, mobile cloud robot security challenges and finally proposed a distributed security platform for mobile cloud robots.

The proposed platform has two sections: robot section and mobile network section. The robot section includes robots and Local Robot Controller (LRC). LRC is a central

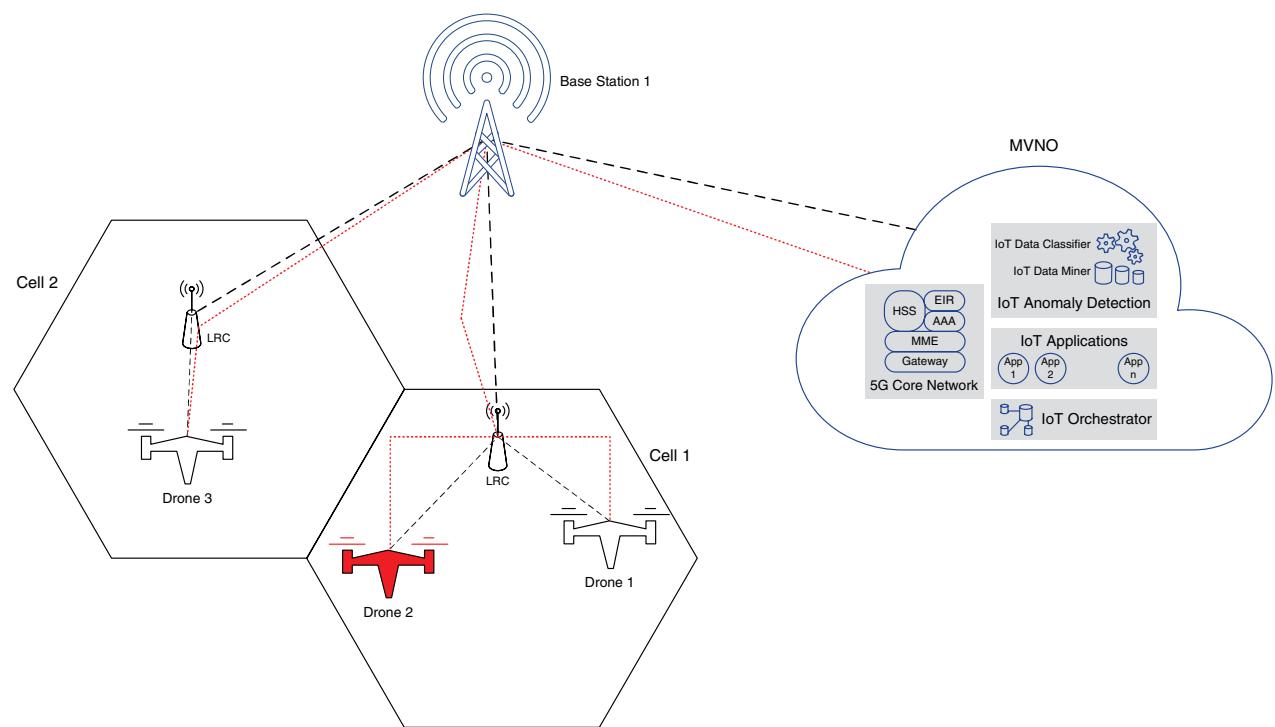


Figure 11.12 Robot attack.

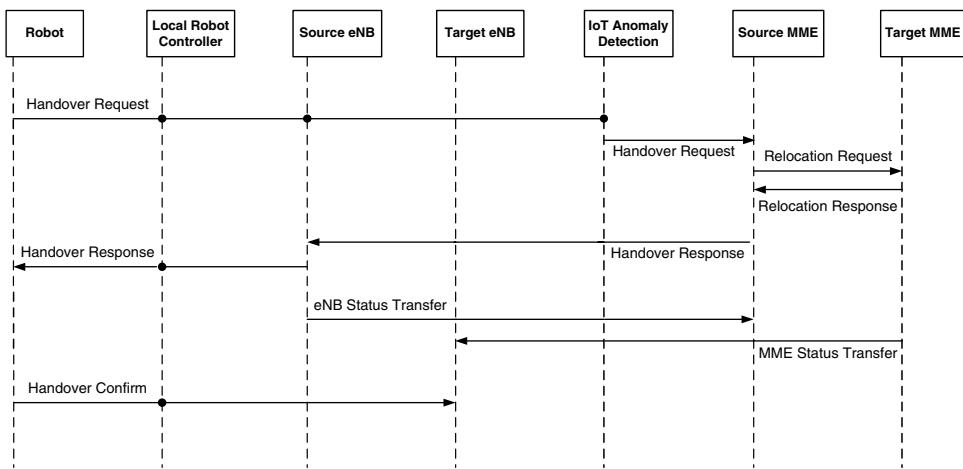


Figure 11.13 Robots roaming.

node for all robots and responsible for connectivity between robots and mobile networks. The first level of data monitoring and security analysis is done in LRC before data is forwarded to mobile network. The IoT service provider can be either a mobile network operator or a Mobile Virtual Network Operator (MVNO). The MVNO carries out advanced analysis of collected data from various LRCs through data-mining mechanisms in an IoT anomaly detection module. The IoT anomaly detection module includes linear and learning algorithms. The linear data-mining algorithm will extract and provide proper attributes to the next process that uses a learning algorithm. This mechanism would decrease the load of input data for the learning algorithm, which is the most time-consuming part (because of the algorithm complexity). So, the proposed algorithm could label new attacks and detect known ones (e.g. botnet, malware) faster than the existing solutions. The analyzed data would be used for intrusion detection in the same or a different MVNO. While a robot has been identified as malicious in an MVNO, a notification will be sent to all MVNOs by an IoT orchestrator to block the malicious robot and prevent the re-analysis of the malicious robot in future, since the malicious robot may try to attack the cloud platform using other MVNOs. Therefore, the distributed security platform reduces the computation processing for a malicious robot that has been already identified as malicious in other MVNOs.

References

- 1 Nokia Networks (2016) 5G masterplan – five keys to create the new communications era. White Paper, C401-011949-WP-201601-1-EN.
- 2 Monshizadeh, M., Khatri, V., Kantola, R. and Yan, Z. (2017) An orchestrated security platform for internet of robots. *Proceedings of The 12th International Conference on Green, Pervasive and Cloud Computing (GPC)* [accepted].
- 3 Ian, C. et al. (2016) *Automated Drone Systems*. Patent US2016266579, September 15, 2016.
- 4 Daniel, O. (2016) *Drone Docking Station and Delivery System*. Patent WO2016094067, June 16, 2016.

- 5 Shiram, G. and Roberto, M.J. (2016) *Methods, Systems and Devices for Delivery Drone Security*. Patent US2016068264, March 10, 2016.
- 6 Anthony, N.D. et al. (2008) *Managing Communications between Robots and Controllers*. Patent WO2008156910, December 24, 2008.
- 7 Hidano, S., Pečovský, M. and Kiyomoto, S. (2015) New security challenges in the 5G Network. *Proceedings of the Computational Intelligence and Intelligent Systems: 7th International Symposium (ISICA)*, pp. 619–630.
- 8 *Why IoT needs 5G*, IEEE Spectrum, May 2015. Available at: <http://spectrum.ieee.org/tech-talk/computing/networks/5g-taking-stock>
- 9 Zhang, Q. and Fitzek, F. (2015) Mission critical IoT communication in 5G. *Proceedings of the First International Conference (FABULOUS) on Future Access Enablers for Ubiquitous and Intelligent Infrastructures*, pp. 35–41.
- 10 Nokia Bell Labs Consulting (2016) Who will satisfy the desire to consume. Mobility Report, PR1603018674EN, April.
- 11 *Smart Meters in Spain can be Hacked to hit the National Power Network*, Security Affairs, October 2014. Available at: <http://securityaffairs.co/wordpress/29353/security/smart-meters-hacking.html>
- 12 *BMW Update Kills Bug in 2.2 Million Cars that Left Doors Wide Open to Hackers*, Forbes, February 2015. Available at: <https://www.forbes.com/sites/thomasbrewster/2015/02/02/bmw-door-hacking/#435cfa346c92>
- 13 *After Jeep Hack, Chrysler Recalls 1.4M Vehicles for Bug Fix*, Wired, July 2015. Available at: <https://www.wired.com/2015/07/jeep-hack-chrysler-recalls-1-4m-vehicles-bug-fix/>
- 14 Dropmann, U. (2016) 5G Technology Aspects. *Workshop on Forward Thinking for Spectrum – Getting ready for 5G, GMSA*, November.
- 15 Nokia (2016) An Internet of Things blueprint for a smarter world. Strategic White Paper, PR1509014445EN.
- 16 Mirai DDos attack a wake-up call for IoT industry, Nokia Networks Blog, November, 2016. Available at: <https://blog.networks.nokia.com/mobile-networks/2016/11/01/mirai-ddos-attack-wake-call-iot-industry?hootPostID=068bdf3d2df0e8cd251248fb5917293e>
- 17 Ahrlich, N. (2016) Cybersecurity protecting and securing utilities. *Workshop on CyberSecurity, European Utilities Telecom Council (EUTC)*, September.
- 18 Monshizadeh, M., Yan, Z., Hippeläinen, L. and Khatri, V. (2015) Cloudification and security implications of TaaS. *Computer Networks and Information Security (WSCNIS), 2015 World Symposium on*, pp. 1–8.
- 19 Monshizadeh, M. and Zheng, Y. (2014) Security related data mining. *Proceedings of the 2014 IEEE International Conference on Computer and Information Technology (CIT)*, pp. 775–782.
- 20 3GPP TS 22.268 V13.0.0. *Public Warning Systems*. Available at: http://www.3gpp.org/ftp/Specs/archive/22_series/22.268/22268-d00.zip
- 21 Nokia targets industrial IoT with private LTE, *RCRWirelessNews*, September 2016 (Online). Available at: <http://www.rcrwireless.com/20160929/internet-of-things/nokia-targets-industrial-iot-with-private-lte-tag4>
- 22 *Private 4G/LTE*, Duons. Available at: <http://www.duons.com.au/private-tactical-4g-lte/>

12

User Privacy, Identity and Trust in 5G

*Tanesh Kumar¹, Madhusanka Liyanage¹, Ijaz Ahmad¹,
An Braeken², and Mika Ylianttila¹*

¹ Centre for Wireless Communications (CWC), University of Oulu, Finland

² Industrial Engineering INDI, Vrije Universiteit Brussel VUB, Nijverheidstraat, Brussels

12.1 Introduction

5G systems are the next major transition in the way of future mobile communications. 5G technology promises to provide higher bandwidth and lower latency. Unlike the traditional mobile technologies, which are mainly meant for voice and data communications, 5G ensures to provide much more. 5G technology has the great potential to enable services for new use cases and vertical industries, for example, in the healthcare, transportation and smart homes. It provides opportunities for companies to build new business models to deliver novel services to consumers in more improved and efficient ways, as well as to increase their revenue. This rise in new business models, architecture and technological changes in 5G will bring new challenges to the user's privacy. The privacy requirements is one of the crucial elements to consider in the discussion of 5G technology as it is of utmost importance to balance the privacy requirements of users with respect to the services offered.

The on-going mobile networks mainly consider four security aspects, that are; authentication, integrity, confidentiality and availability. However, the privacy requirements are not at all (not only) taken into account from the infrastructure and but architecture's perspective as well. As 5G will produce novel and critical applications, it is therefore vital to consider the privacy characteristics from the architecture's point of view, such as observability, anonymity, unlinkability and pseudonymity. This will also ensure the strong trust relationship of the consumer with mobile operators and with the third parties, which are providing the various services. Also, not all privacy aspects can be included while addressing the architecture of the network, due to the lawful and privacy regulatory policies [1,14,15].

5G technology predicts the vision of "always available", where the services are available to users anytime and anywhere. This 24/7 connectivity with other devices may originate a number of attacks such as impersonation, Denial-of-Services (DoS), and replay attacks among others. 5G technology is also considered the key enabling technology for providing ubiquitous connectivity for smart objects. The immersive experiences, such as context

aware services, augmented reality, and concepts of anything as a service and user personalization will be a major vital force behind the massive adoption of 5G technology. 5G is also the main driver for Internet of Things (IoTs)-based applications, where things are connected through this technology and services will be delivered by more efficient and faster means. This means that 5G requires special consideration on privacy requirements from various perspectives of technologies and services [14,15].

Moreover, due to recent advancements in sensing and communication technologies such as smartphones and wearables, the general awareness of privacy in current society has increased, and thus this encourages higher protection of user's metadata and communications. With the kind of capabilities 5G would possess, it is expected that novel use cases and applications will come into the real-time actions. Service-oriented privacy mechanism would be a more preferable way to protect the privacy. Also, in the case of 5G, security- and privacy-based solutions need to be focused from scratch. Therefore, it must add the security and privacy features built in to the system design from the start.

The continuous improvements in mobile communication technologies also require enhancement in identity management techniques. 5G technology will bring an enormous number of users and devices together and they will be connected in a ubiquitously manner, therefore it is crucial to protect the identities of subscribers as well as of devices. It is important to make sure that no any adversary or third party can steal the subscriber's real identity without his/her consent. Similar kinds of secure approaches are needed for building and maintaining the strong trust relationship among subscribers and various stakeholders, such as service provider, enterprises, etc.

This chapter mainly highlights the potential privacy, identity and trust challenges for future 5G technology from the user's point of view. Section 12.2 gives some background knowledge regarding security and privacy issues about previous generations. Section 12.3 elaborates on the user's perspective on privacy, which is further expanded into three sub-parts, that are; data, location and identity privacy. Identity management mechanism and its related challenges are explained in Sections 12.4, and Section 12.5 presents the trust issues and elements required for developing the business models in the 5G system. Finally, we discuss the overall aspects in Section 12.6 and conclude in Section 12.7.

12.2 Background

Over the last two decades, smart devices such as smartphones and tablets have provided more ubiquitous and persuasive types of services to consumers. The initiation of mobile communication systems, starting from the second-generation Global System for Mobile Communications (2G/GSM) and heading towards the third-generation Universal Mobile Telecommunication Systems (3G/UMTS), has expanded widely into all parts of the world. The next major transition is this evolution was the latest generation, "Long Term Evolution" (4G/LTE) systems, which are being broadly deployed.

Right from the start, there have been numerous threats faced by 2G systems, such as no mutual authentication mechanisms available between mobile phone users and the networks. It means that with these limited resources, it was easy for an attacker to launch a fake base station and assure the mobile devices that it is a valid base station and that it can connect to it. Fake base stations can also act as International Mobile Subscriber

Identity (IMSI) catchers, due to lack of authentication mechanisms and thus can be used to trace and monitor users. The next major transition is the 3GPP (Third Generation Partnership Project), which increased the level of security as compared to the 2G systems. The security specifications in 3GPP also included the mutual authentication mechanisms [2,3]. Furthermore, with the increase in the amount of mobile data, along with the evolution of new applications, the motivation grew to move from 3GPP towards the fourth-generation. LTE is designed to allow strong cryptographic, encryption and mutual authentication mechanisms [3,4].

The techniques to tackle the identity management challenges are an essential part of 5G, due to the fact that the security requirements will be high in this case. Threats such as International Mobile Subscriber Identity (IMSI) catching were also discussed during the standardization of 3G and 4G and thus it is also considered as a focal point in 5G systems [5]. There is not yet any complete or exact document/specification available (at least not in the technical specification for 3GPP) regarding trust models for on-going mobile networks (2G-4G). But considering the trend of security requirements, which has evolved from 2G-4G, the current trust model for mobile networks can also be analyzed. However, in the case of 5G networks, the trust model among various stakeholders would be even more complex, because additional entities will also become involved.

12.3 User Privacy

5G technology will enable several novel applications that will potentially open doors for a large number of vertical industries. This leads us to the fact that a large amount of personal information will be carried out over the 5G networks. With the introduction of data-mining techniques, it is easier to retrieve the data privacy information and thus the data is at huge risk. The 5G system must provide security mechanisms for protection of a variety of trusted information, regarding humans as well as for machine-users (e.g. identity, subscribed services, location/presence information, mobility patterns, network usage behavior, commonly invoked applications, etc.).

5G technology would also offer customized network services for consumers by realizing the characteristics of particular services. Thus, the privacy requirements in the 5G network may vary from service to service. 5G technology will also enable service-oriented privacy requirements. For example, health information of the users in certain healthcare applications will require a higher degree of privacy. Also in the case of some critical industrial tasks, equally higher level of privacy protection is required. But applications like searching for some kind of location information may require a smaller degree of privacy. For more focused understanding, we have split the user privacy concepts into three parts, that are; data, location and identity privacy, as shown in Figure 12.1.

12.3.1 Data Privacy

There will be heaps of smart and heterogeneous devices connected through 5G technology, thus the chances of leakage of the user's personal data is very high. Service providers/companies store and use the private information of consumers without their permission. In some cases, the service provider stores the user data for their own product, but later shares them with other companies so that they can analyze the data and find some trends that, which of their own product is more suitable for that particular

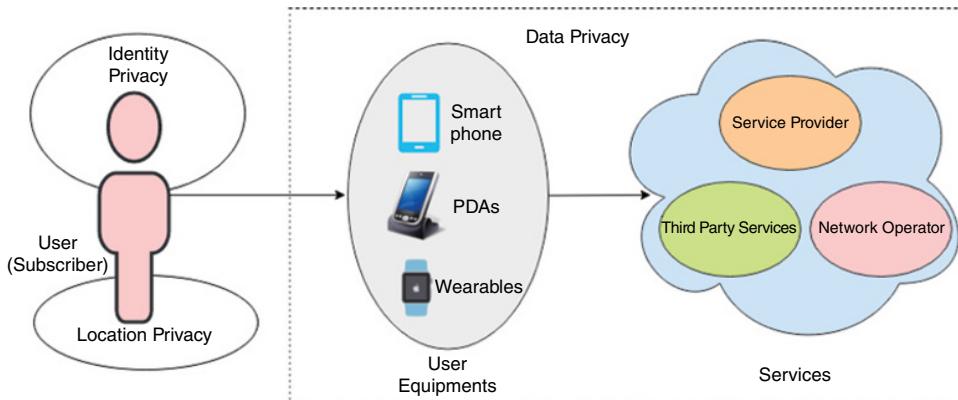


Figure 12.1 Various elements in user privacy.

user. In certain cases, it is even useful to take some of the user's personal data, and based on that, the company can build new products and services. But companies/service providers need to provide a clearer explanation regarding for what purpose their data have been used. They must also answer questions like what data has been taken and how and where they stored it.

Several smartphone applications (Apps) for example in android, ask for certain information before the installation. Mostly, the information for which the App wants permission does not have any direct relation to service of that particular App. This data can be used for other purposes, which are undefined by the Apps developers [6]. Nowadays, social media sites are the most common ways to share public and private information among various users. These are the frequent ways of updating others about your current activates, share/upload personal pictures and even can have live text, audio and video conversations. 5G is supposed to enable this kind of communication seamlessly and continually. But still many people raise doubts over the leakage of their personal information by various means, which are of real concern for our current society.

5G-based IoT systems are the crucial part of future technologies for providing numerous digital services. This will eventually generate massive amounts of data all the times. Since IoT is becoming omnipresent, large amounts of data will be coming into action. 5G will assure the increase in data transfer speeds and thus have a higher risk of malicious attacks. In a similar way, wearable devices produce an explosive quantity of data, because sensors/chips attached to wearables are continuously monitoring and gathering user personal information such as fitness, pulse rate, etc. This data may be analyzed by a third party, who can extract other features from them without asking permission of the user.

Data privacy risks arise when the third party/service provider or any malicious attacker wants to access the user's personal data without their consent. For example, by monitoring the activities of someone using its personal data, one can easily predict the daily routine of that particular consumer. This can be harmful in some cases, because if somebody wants to observe/guess the activities of a person, they can easily do so. The other critical example can be that of healthcare, where the medical data is very sensitive.

In many cases, the patient wants to restrict some particular information to certain people such as doctors, specific family members or friends. But malicious users or unauthorized persons may access the information and use it for unethical purposes. Another such example of privacy breaches can be the purchasing of anything on a consistent basis, like any particular type of food that may reveal religion or health information.

In 5G networks, for many cases, the privacy protection requirements are also dependent on the usage of the particular access technology. Because the element of heterogeneity will be available at much higher rates, along with multiple access technologies it will be used to obtain the required services. User data will be traversed in various access networks in 5G and different vendors will provide the functional entities for the network. As a possibility, by using data-mining methodologies, a third party can derive user personal information by analyzing the user disperse data, which might be available on any part of the network. Because of the risk of such scenarios, more rigorous data privacy protection schemes are required for 5G networks [7].

It is important to formulate strong data protection mechanisms while discussing standardization and policy-making for 5G technology. The service providers must also explain ways of data collection and its use for various services. There should be a balance between the user privacy and data used by the service providers, so that companies can build novel and useful applications for the user and at the same time, user privacy should not be affected. Accountability mechanisms should be involved so that monitoring of each action by various entities will be easy. Data minimization techniques should also take into account such that companies/service providers/third parties limit the data they collect and retain, and dispose it once they no longer need it.

12.3.2 Location Privacy

Nowadays, several smart devices such as smartphones, tablets and wearables, which have powerful computational and storage capabilities along with positioning technology, can request services at anytime and anywhere. Location Based Services (LBS) are popularly used with respect to the development of future wireless technology. With the introduction of 5G, which will enable seamless and continuous availability of services, the location of the user is also continuously monitored in such cases. In order to provide improved services, various companies have also started to track the current location of the user. From this information, they are constantly monitoring the habits and routine of the user. At one hand, this kind of tracking service helps companies to improve their services and to build new user-friendly services. However, on the other hand, it raises serious concerns over user privacy.

Also, many online apps on mobile devices require location information along with their personal information. In some cases, location information of the user is taken, no matter whether it is to be used or not. These online apps want more and more information with each of their updates. Nowadays, the social media applications like Facebook also have the option “check-in”, where users are sharing their current locations. This will raise concerns over tracking of user movements by observing the location information constantly. Recently, wearable devices are also actively used for tracking purposes, such as tracking children and pets. These wearable devices are tracking the respective user every second and that also raised huge privacy concerns.

There are few existing techniques available for preserving the location privacy that can also be useful in the context of location privacy protection in 5G applications/technology. Common methods used to protect the location privacy of the user may include anonymization, pseudonym change and path perturbation [8]. Regulatory approaches are also needed, so that strong rules, regulations and legislation can be designed for proper usage of the network, putting the awareness over of the specification of internet and network security [9]. Encryption-based techniques are among other available ways to protect the user's location privacy. The message is encrypted by the user before sending it to the LBS provider. Once the message is received by the LBS provider, it will be decrypted. This approach includes relatively more intensity of anonymity, but it has high computational and communication costs, which is one of the disadvantages of using this approach [9].

Anonymity-based approaches hide the user's real identity and replace it with pseudonyms. In this case, a trusted middleware is used to generate the fake information (being the pseudonym), which is then sent to the LBS provider for particular location service. There is also another approach, in which the quality of user location information downgrades in order to preserve the location privacy; this is known as obfuscation. For example, one spatial cloaking-based technique is formulated that enlarges the location point of the user to a region called ASR which contains the location of the user, where a typical k-anonymity technique is mostly utilized. Then the trusted third party sends the ASR to the LBS provider to complete the LBS query. And finally, some privacy policy-based approaches are required to ensure that it can restrict the misuse of location information in certain ways, for example, by protecting the user privacy through information retrieval methods [9].

12.3.3 Identity Privacy

While acquiring digital services on certain occasions, consumers do not want to reveal their original identity to other users or to service providers. For example, when asking any queries online or giving feedback on the website of companies, users prefer to remain anonymous. In some situations, users might use temporary or fake identities and discard them when the required task is completed. Knowledge of permanent identity of a user may permit an adversary to track and amass comprehensive profiles about individuals. The trend of stealing online identities is more common nowadays. There are numerous online applications such as shopping and banking that would require online ways of payment through credit cards. This information may lead to revealing the real identity and can cause possible risks to the user's privacy. Usually, anonymity-based approaches are used to hide the real identity. Identity privacy can be further divided into subscriber and device identity privacy:

- *Subscriber Identity Privacy:* In this case, threats can arise when users are tracked or monitored by the subscriber's identifier or may be through a temporary identifier. Users also generally do not want any kind of linkage/connection between their subscriber's identity and device identity. The possible solution to protect the subscriber's identity privacy would be through encryption of the IMSI and usage of enhanced pseudo-identifier. In order to guarantee the unlinkability of the subscriber and device identifier, an anonymization system might be one of the potential approaches to consider [10].

- *Device Identity Privacy:* The vulnerabilities that could exist concerning device identity privacy are that, subscribers do not wish to be tracked by their UE identifiers. Likewise, as with subscriber identity privacy, users also do not want linkage between their subscriber's identifiers with device identifiers. This can be resolved by studying the possible end-to-end anonymization approaches that provide a guard against the unauthorized tracking of devices and preserve the disclosure of device identity. 5G also ensures that only through a confidential protected message, the International Mobile Equipment Identity (IMEI) should be sent [10].

12.4 Identity Management

The mechanism of handling the identity in a 5G system would be a crucial task for certain reasons, such as keeping a high degree of security to ensure mutual authentication along with maintaining the user's friendliness and privacy. As 5G technology will become a more multi-vendor environment and comprise of various stakeholders, a strong identity management mechanism is needed to protect the identity and network from unauthorized access of users. During the standardization process of 3G and 4G, when equipment such as mobile devices show their specific identities, there is a serious threat regarding International Mobile Subscriber Identity (IMSI) catching. At the moment, there is no protection mechanism proposed, because that particular threat has not caused any serious trouble to the access network. It is not completely clear yet that whether this attack is still valid for 5G technology and needs any further consideration [8].

The core reason behind an IMSI catching attack is that while in the unavailability of Temporary International Subscription Identity (TMSI), User Equipment (UE) might be unable to use IMSI as its identifiers. IMSI catching attacks can be both active and passive. In passive attacks, all IMSI can be captured and gathered by the IMSI catcher when it eavesdrops in the neighborhood of the wireless traffic. On the other hand, in active attacks, a fake base station is established that has strong signal strength and might be considered as a legitimate base station. Mobile devices usually prefer the base station with the highest signal strength. A message requesting the identity is sent to all mobile devices within the specific area through this fake base station. Mobile devices send their IMSI, as they consider it a valid base station, which have lost the connection to TMSI. Catching IMSI is often supposed to be a starting point for more detailed eavesdropping attacks in GSM [12]. An enhancement technique is mentioned in [12], which allows home network operators to place their trust less on serving networks and guard against IMSI catchers. The idea is to enhance the handling of identifiers and protocols, so that the UE and home network can see IMSI in clear text.

The on-going privacy preserving mechanisms do not guarantee protection against the threats on air interface, because they act as a valid network that has lost temporary identity and also when the request for IMSI is made, there is no proper protection for passive eavesdroppers that possibly could be available there [12]. Several privacy related attacks have been reported, such as in some cases when a fake base station is plotted leading to the derivation of user's personal information. IMSI attacks are mainly focused on stealing IMSI of a mobile subscriber. The IMSI catcher requests the subscriber's identity from the mobile subscriber for the longer term. This is considered as a normal routine request and in reply, mobile devices send its IMSI using standard

security mechanisms. Hence the IMSI catchers are also used to monitor and track the locations of specific subscribers.

The traditional cellular systems are usually dependent on Universal Subscriber Identity Module (USIM) cards, to manage user's identities and keys. In the case of 5G systems, devices and equipments such as smartphones, wearables and smart sensors would be comparatively smaller in size and too economical to accommodate USIM. Therefore, novel methods are required in this case to manage the device identities [7]. There is possibility of a framework that can comprise the device and service identities together. During the manufacturing phase, the device identity can be allocated and is considered as a unique global identity. On the other hand, the service provider offers the service identities. Device identity is also referred to as the physical identity and may address single or multiple service identities. It allows users to make their decision regarding which particular device can be permitted to access the network and utilize the assigned services [7].

The 5G ecosystem would need more flexible, general and open identity management infrastructure, which should have the scope for various alternatives and be able to give permission for that. One way of proceeding for companies is that, they might allow the existing ID management mechanism to reuse it for 5G access. In 5G, there would be an immense number of hand-held devices that may include smartphones and tablets, as well as wearables; therefore it is important to find ways on how to handle these devices and at the same time maintain the subscribers' identities. It is also crucial to think about novel trust models for 5G networks. There are also some key concepts such as network slicing and virtualization, which must be considered when proposing the secure identity management solutions for such scenarios [11].

12.5 Trust Models

Trust models change with respect to the time and improvements in technology. Previously, for companies, the mobile devices of all users are considered to be more trustworthy, because they were issued and managed by their own IT department. But recently, every user in the company/enterprise wants their own personal gadgets, which can eventually lead to a number of security threats to firewalls of the company [11]. Trust includes capabilities such as security, identity management and privacy. Even with to date, there are no standard trust models available for current networks (2G-4G). The current (2G-4G) trust model mainly comprises of the entities, such as user (subscriber), service provider, network operator, Virtual Mobile Network Operator (VMNO), and equipment manufacturer among others, as shown in Figure 12.2 [1]. Usually, the subscribers keep their trust over the service provider and assume that the rules and regulations written in the service subscription contract/billing must be followed, and this trust is then further developed based on experience, reputation and the legal framework. It is assumed that the end-to-end path between subscribers and service providers are secured during any communication (may be voice), and users trust that their critical data is secure with operators.

The service provider is responsible for providing the required services to subscribers through some kind of user device or equipment, such as a mobile phone or tablet. The trust which service providers mostly seek from the subscribers is that they must

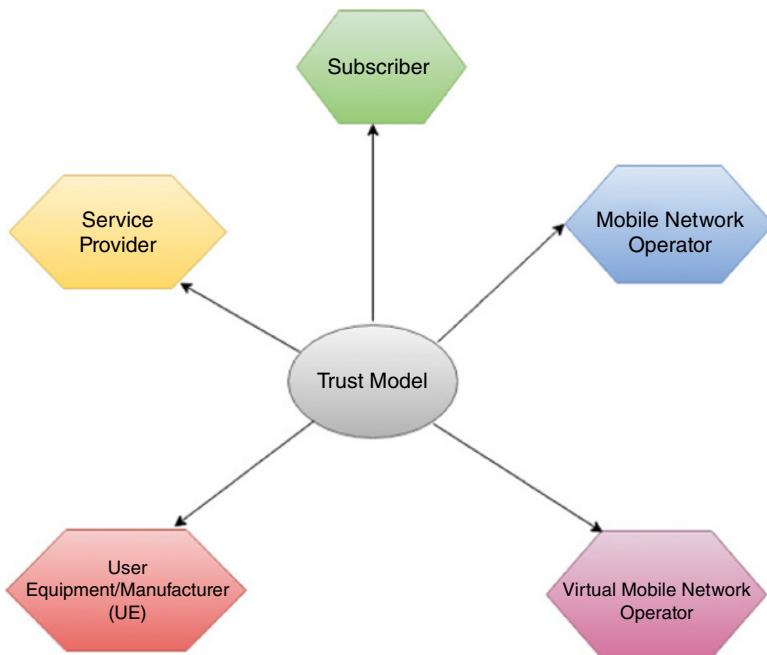


Figure 12.2 Actors/Entities in the trust model.

be able to pay the billing/charging or subscription price well within the predefined time. Although service providers do not have much trust that the subscribers will maintain strong passwords for their authentication to the services and therefore, in order to authenticate securely, it offers the subscriber an UICC. In some scenarios, the two terms, provider and network operator are used in the same context. The network operator is known to be the central element and provides a trust relationship with various other elements in the network. For example, Mobile Network Operators (MNOs) or Satellite Network Operators (SNOs) perform operations such as deploying, maintaining and managing the network (Satellite). Up to now, there are no such standardized security procedures available, which can highlight the network operators sharing certain information. In the current scenario, the trust relationship among various network operators is strong and regulated by contract. However, there can be untrustworthy network operators, who can misuse the personal data, and that can be a serious threat for such networks [1].

VMNO is based on one of the special forms of network operator. It does not contain the mobile network, but instead borrows some virtual space from the database of that network operator. Therefore, it follows almost the same trust model and entities as assigned for the network operators. It keeps the trust between VMNO and its infrastructure elements and can utilize the various resources as agreed in the contract between both of them. Regarding the (UE), it was assumed that this entity does not need to be included in trust models, because the network operator is the one who chooses which manufacturer is trustworthy and which equipment should be used. But there are some cases such as for USIM manufacturer, where it is

required to consider the higher level of trust because of the special requirements for USIM/UICC [10].

From the network operator perspective, Next Generation Mobile Networks (NGMN) [13] presented three types of business models, that is the Asset provider, Connectivity Provider and Partner Services Provider. For Asset, XaaS and network-sharing models are the most important. The Connectivity provider relies on two business models, that is basic (projection of current 4G business) and enhanced. There are also two business models for partner service providers. The first is “operator offer enriched by partner”, which deals with the services provided by the mobile network operator using the unique capabilities of third-party resources. The second one is “partner offer enriched by operator”, where unique capabilities of an operator is utilized to deliver the services directly to subscribers.

The existing trust models might not be fully applicable for 5G networks, as in the case of 5G, where there will be additional entities and actors coming into action to provide and support numerous novel services. Thus, building the trust model will not be straightforward and will include far more complexities than one could ever have imagined. For example, one of the possibilities could be to introduce a new entity, such as the external cloud infrastructure. This is because through virtualization, network operators can run some operations and applications of the network on the external cloud. Furthermore, these external clouds might also have their own data centers at different jurisdictions. Other possibilities to improve the services delivery in 5G are by in sourcing some of the network functionalities, which can be performed by the third party. As Content Distribution Network (CDN) providers integrate catches in the network operator, it is crucial to consider the fact that the network should not become affected by the addition of these new functionalities. In traditional mobile communication architectures, telecom authorities are responsible for giving access to valid users for specific networks only. There is no trust model available between the authentication of users with their services. However, in 5G networks, this shortcoming would also be resolved, as networks can authenticate the service providers to have even more secure and efficient identity management, as shown in Figure 12.3 [7].

Service providers do not actually trust subscribers to authenticate themselves; instead they are dependent on IMSI stored in the USIM. Hence, IMSI and IMEI are used to

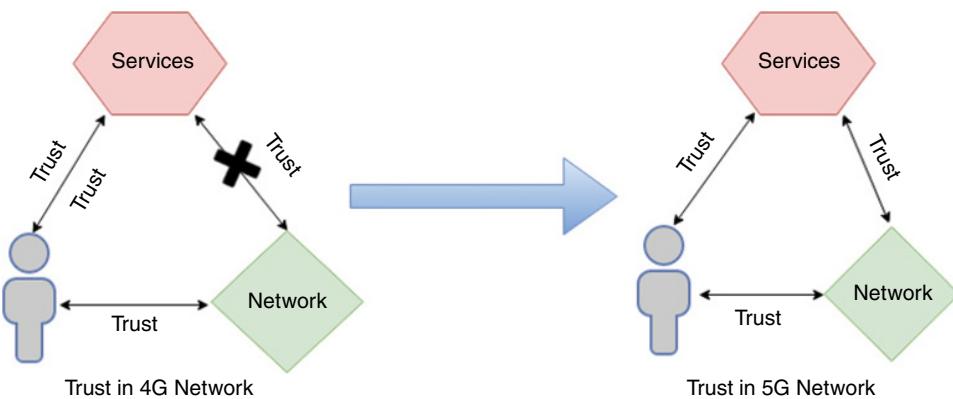


Figure 12.3 Evolution in the trust model.

Table 12.1 Potential threats between stakeholders.

| Type of threat | Detail |
|--------------------------------|--------------------------------------------------------------------------------------------------------|
| Malicious stakeholders | One stakeholder might work against another's interests |
| Non-malicious actions | Caused by actions of stakeholders technological proxies, user errors |
| Malicious attacks | External attacker may subvert technology of stakeholders, which lead to act against other stakeholders |
| Internal failures | Caused by internal faults in system leads to detriment of the stakeholders |
| External disasters | Caused by external sources, such as natural disaster |
| Threat to stakeholder trust | Stakeholder continues trusting and using system |
| Threat to stakeholder distrust | Stakeholder loses trust and withdraws from system |

detect the attempts of attacks by or against stakeholders. Therefore, subscribers and service providers put their trust in the manufacturers of USIM and Mobile Equipment (ME) domain equipment. Equipment operators are mainly responsible for its behavior, whereas the manufacturer has less responsibility. In some cases, contracts reflect the trust between various stakeholders, such as a roaming agreement of the service provider with other roaming providers, which can permit users to connect to their respective domains. Both the service and roaming providers must have agreements with other providers (services) to establish a communication path for subscribers. In the 5G (4G) network, trust relationships exist between various stakeholders, but there may be certain attacks in which trustworthiness of the equipment may not stay according to expectations [1]. Table 12.1 highlights such kind of threats.

12.6 Discussion

5G technology will dramatically change the current way of acquiring digital services. At this moment, it is hard to elaborate on the complete picture as to how 5G will benefit different vertical industries. However, 5G has promised to do a lot for industries such as automation, healthcare and transportation among others. It is assumed that services will get delivered in a continuous and seamless manner. However, this vision can only be achieved by giving proper attention to security, privacy, identity and trust issues for 5G technology.

In the beginning, we focused from the user's privacy perspective and divided this idea into three key elements; data, location and identity privacy. Huge quantity of personal data is generated by various devices, which are being processed by a service provider/ third party. The main issue is that personal information of a subscriber is being used without their permission. In some cases, one company shares the personal information with other companies, so that they can analyze the data and make the relevant new product to generate higher revenues. Hence, strong accountability and transparency is required in such cases. Strong data protection and privacy regulations should be considered in 5G systems. Similar in the case of location privacy, companies through various location-based applications track the user's location without informing them. This raises

privacy concerns for the subscribers. The subscribers may even not want to reveal his/her real identity in environments that are unfamiliar to them. An anonymity-based solution is preferable in such scenarios.

Then identity management mechanisms are discussed, which is another key area to focus while designing the standards and regulation for 5G. New 5G radio access technologies for certain applications, such as industry automation, can be beneficial in terms of low cost and high quality of services. In such cases, it is required to have device identity management for 5G access in various industries. Also, similar kinds of identity mechanisms are needed in situations involving satellite networks and dual satellite and terrestrial 5G access.

Finally, trust is as equally important as privacy and should be taken care of properly in parallel. The trust must cover both human and machine perspectives and trustworthiness by design methodologies. We have explained the entities and elements involved in current 4G networks and highlighted the trust aspects among them. There are no complete standardized trust models available for 4G networks, but various concepts from them can be useful in building the trust among various stakeholders in 5G. The role of privacy would be considered vital in defining the actors for modelling the trust in 5G networks. The ideal trust model for 5G network should be able to answer the questions such as “for what one does on trust?”, “how much should one trust?”, and “how much anyone can trust?”

12.7 Conclusion

This chapter primarily focused on privacy, identity and trust challenges of the user in future 5G systems. It is undoubtedly agreed by all entities and actors involved in 5G technology that without handling the privacy properly, 5G would face larger obstacles in the way of complete acceptance, adoption and success among their users. From the user's point of view, data, location and identity privacy are the basic-key elements to consider. In the 5G system, privacy features must be considered right from the designing phase and some of them should be built into the system. Furthermore, the system should be intelligent enough and can adopt the privacy accordingly to the degree of importance of services. The context aware applications and services would also require more focused privacy solutions.

5G will be the driving force of many other technologies, such as IoT and therefore an enormous number of users and smart devices would come into action. It is necessary to have a secure identity management mechanism for both subscriber and device. The privacy sometimes has a conflicting relationship with trust, as more trust on service provider can increase the risk of privacy violations. Future 5G systems will introduce new business models that will eventually increase the number of stakeholders and therefore trust association among each of them will be crucial. 5G technology might use some similar concepts to existing trust models, along with the addition of few new actors and entities.

References

- 1 Deliverable D2.2 Trust model, 5G-ENSURE. Available at: http://www.5gensure.eu/sites/default/files/5G-ENSURE_D2.2%20Trust%20model%20%28draft%29_v1.1.pdf
- 2 Shaik, B.A. et al. (2015) Practical attacks against privacy and availability in 4G/LTE mobile communication systems, Computing Research Repository, October.
- 3 Gindraux, S. (2002) From 2G to 3G: a guide to mobile security. *Proceedings of the Third International Conference on 3G Mobile Communication Technologies*, p. 308–311.
- 4 Seddigh, N., Makkar, N.R. and Beaumont, J.F. (2010) Security advances and challenges in 4G wireless networks. *Proceedings of the Eighth Annual International Conference on Privacy, Security and Trust*, pp. 62–71.
- 5 Rannenberg, K. (2004) Identity management in mobile cellular networks and related applications. *Information Security Technical Report*, 9(1), 77–85.
- 6 Sørensen, L.T., Khajuria, S. and Skouby, K.E. (2015) 5G visions of user privacy. *Proceedings of the IEEE 81st Vehicular Technology Conference (VTC Spring)*.
- 7 5G Security: Forward Thinking Huawei, White paper. Available at: http://www.huawei.com/minisite/5g/img/5G_Security_Whitepaper_en.pdf
- 8 Sadegh, F. et al. (2015) PHY-layer location privacy-preserving access point selection mechanism in next-generation wireless networks. *Proceedings of the IEEE Conference on Communications and Network Security (CNS)*.
- 9 Yu, R. et al. (2016) A location cloaking algorithm based on combinatorial optimization for location-based services in 5G Networks. *IEEE Access*, 4, 6515–6527.
- 10 Deliverable D2.1 Trust model, 5G-ENSURE. Available at: http://www.5gensure.eu/sites/default/files/Deliverables/5G-ENSURE_D2.1-UseCases.pdf
- 11 5G Security, Ericsson, White paper, June 2015. Available at: <https://www.ericsson.com/res/docs/whitepapers/wp-5g-security.pdf>
- 12 Norrman, K. et al. (2016) Protecting IMSI and user privacy in 5G networks. MobiMedia, Xi'an, China, June 18–20, pp. 159–166.
- 13 NGMN 5G, White paper. Available at: https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf
- 14 Ijaz, A., Namal, S., Ylianttila, M. and Gurtov, A. (2015) Security in software defined networks: a survey. *Proceedings of the IEEE Communications Surveys and Tutorials*, 17(4), 2317–2346
- 15 Securing the future of mobile services. An analysis of the security needs of the 5G market: a SIMalliance 5G Working Group marketing white paper. Available at: http://simalliance.org/wp-content/uploads/2016/02/SIMalliance_5GWhitepaper_FINAL.pdf

13

5G Positioning: Security and Privacy Aspects

*Elena Simona Lohan¹, Anette Alén-Savikko², Liang Chen³, Kimmo Järvinen⁴,
Helena Leppäkoski¹, Heidi Kuusniemi³, and Päivi Korpisaari²*

¹ Tampere University of Technology

² University of Helsinki, Faculty of Law

³ Finnish Geospatial Research Institute

⁴ University of Helsinki, Department of Computer Science

13.1 Introduction

Positioning has so far been an add-on feature to mobile phone standards. Wireless positioning typically relies on power-hungry technology and has traditionally been designed and optimized separately from the communication part. With the advent of 5G communications, this is likely to change and joint communication-positioning architectures based on 5G are expected to be implemented. The official 3GPP target for future cellular networks is to support 1 trillion devices [3]. The 5G concept is based on dense access point deployment and very large bandwidths, and thus it has an intrinsic capacity to achieve very accurate positioning, at extremely low energy consumption in mobile device. However, this requires a careful design of the 5G network in order to utilize fully this positioning potential without a negative impact on the communications features. The top candidates in future 5G positioning are likely to rely on Time of Arrival (TOA), Time Difference of Arrival (TDOA), Direction or angle of Arrival (DOA), and Received Signal Strength (RSS) information. 5G white papers [3,81] specify that the enhanced 5G services should support network-based positioning capabilities, “with accuracy from 10 m to more than 1 m on 80% of occasions, and better than 1 m for indoor deployments”, and even up to 0.3 m in automotive applications. Such a high resolution in locating the user terminal can trigger, on one hand, enormous benefits for both the network operators and end users, for example in terms of user-personalized Location Based Services, location-aware and context-based optimized Radio Resource Management, power and latency optimized end-to-end communications, and location-aware interference mitigation. On the other hand, it can create important privacy concerns from the users’ point of view and it can prove to be sensitive to intentional interferences and security breaches in the mobile localization. Moreover, with the advent of cloud 5G positioning [85], the Location Information Service Providers will likely have to cope with hacker attacks into the databases and malicious or erroneous data inputs.

The motivation of addressing the security and privacy aspects in the 5G positioning comes from the fact that a secure and privacy-preserving positioning architecture can lead to many other applications, which are now not possible because of the lack of appropriate security and privacy mechanisms. All actors in the positioning chain, for example the end users, the network operators, the location service providers and so on, can benefit from the availability of new security and privacy solutions. If the security and privacy of current positioning systems can be increased, this will also enhance the usability of location as a security parameter for digital interactions in more general contexts, for example, social media, wellbeing/health monitoring or surveillance systems. Figure 13.1 illustrates the main players or stakeholders in the 5G localization chain:

- *The Location Information Service provider (LISP)*: also called “location aggregator” [6], is the provider that either performs the user localization in a network-centric approach, with information or measurements from the user (e.g. cellular/5G networks) or transmits relevant information to the user, which enables the user device to compute its own position (user-centric positioning approach). LISP also provides access to their databases to third parties for location-based application development and advertising.
- *The Location-Based Service Provider (LBSP)*: this is the actual provider of a location-based service, such as smart shopping, physical activity detector, tourist information, etc. LBSP processes the user location information and creates suitable location-aware content to the user.
- *The end-user with 5G connection*: this is the user device with broadband access through the 5G spectrum, which requires a certain location-based service. The user device can either position itself with inputs from LISP (e.g. mobile-centric approach), or can obtain its position from the 5G network (network-centric approach).

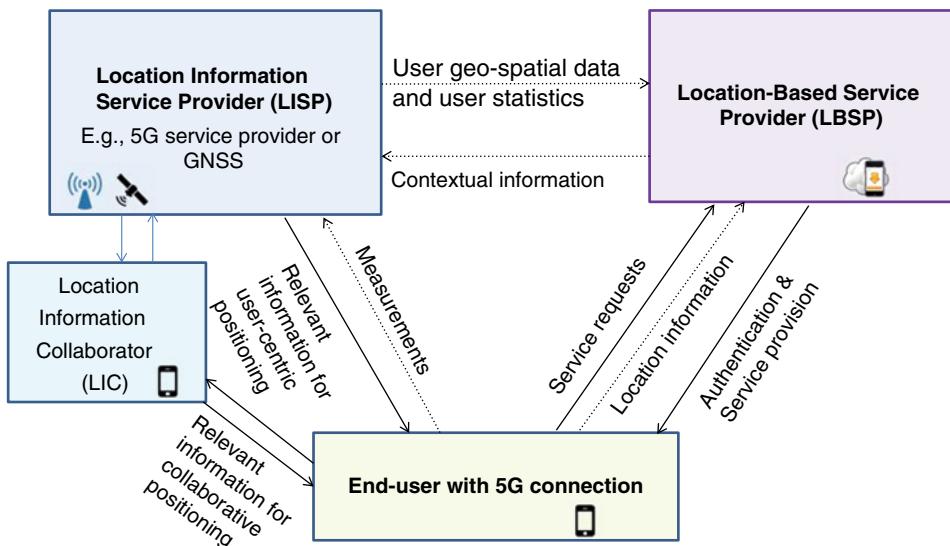


Figure 13.1 Illustration of the localization chain and players in the Location Based Services in 5G.

The simplest example of a mobile-centric approach is when the user has a GNSS engine on his/her mobile device and has geo-maps downloaded into the device memory; thus, the position is calculated entirely based on the memory maps and the GNSS signals, and such a position estimate can fully preserve the user privacy if not sent further to the LBSP. An example of a network-centric position is cell-ID positioning, when the network identifies first the serving “cell” or the serving base station of the user, and then estimates the user location to be within a certain radius (a few tens of meters to a few tens of kilometers) from the identified cell. In this situation, the user position is no longer private, as it is already known by LISP.

- *The Location Information Collaborator (LIC)*: can be present or absent and refers to any other mobile user in the network with whom the desired end-user can collaborate. Indeed, 5G standard supports Device-to-Device (D2D) communications and collaborative communications, and such collaboration can also serve in positioning phase.

In all the interactions between the localization chain players shown in Figure 13.1, there are various threats and weak points that can affect the security and privacy of the users’ position.

13.2 Outdoor versus Indoor Positioning Technologies

While many positioning technologies exist nowadays, there is no winning standalone technology able to offer good coverage and high accuracy in both indoor and outdoor scenarios. There are several differences between indoor and outdoor positioning. First, the outdoor areas, unlike indoor areas, can typically receive satellite signals at a power sufficiently high to allow communications and positioning, while satellite signals are highly attenuated by walls and windows, and barely penetrate the indoor spaces. Second, the outdoor maps are nowadays highly available and highly accurate, while indoor 3D mapping is still an area with many unsolved challenges, such as proprietary map information, privacy issues, non-standardized reference systems, etc. The main positioning technologies available nowadays are summarized in Table 13.1 and their main underlying positioning mechanisms are briefly overviewed in Section 13.3.

13.3 Passive versus Active Positioning

The dichotomy between positioning and communication architectures in 5G is illustrated in Figure 13.2, where the upper plots (a,b,c) explain the positioning-related concepts, and the lower plots (e,f) explain the communication-related concepts. In both cases, we talk about cell- or network-centric versus device-centric architectures, but the terminology is slightly different. In positioning, the unit-centric terminology refers to the unit that actually computes the location. The other unit (network or mobile device) can provide measurements or other signaling sequences to the unit that has the location engine. In addition, when several mobile devices interchange data useful for position estimation (e.g. various measurements or other signaling sequences), we talk about

Table 13.1 Summary of main positioning technologies and their main positioning-related features.

| Technology | 5G | Older cellular (2G, 3G, 4G) | GNSS A-GNSS | WiFi | BLE | UWB | Inertial sensors and other sensors (magnetic, optical, etc.) | RFID |
|-------------------------------------------------------|----------------------------------|-----------------------------------|-------------------------------------------------------|------------------------------------------------------------|------------------------------------------------------------|------------------------------------------------------------|----------------------------------------------------------------------|--------------------------------------------------------|
| Outdoor suitability | High | High | High | Low | Low | Low | Low; they work well as complementary system to some other technology | Low |
| Indoor suitability | High | Moderate | Low | High | High | High | Moderate | Moderate |
| Expected accuracy | sub meter | tens of meters | few meters (mass-market) and sub-meter (professional) | few meters | few meters | cm | few meters | few meters |
| Expected positioning availability | 99% both indoors and outdoors | 80–90% outdoors 60–80% indoors | 99% outdoors 30–50% indoors 99% indoors | 30–50% outdoors 70–99% indoors if devices are installed | 30–50% outdoors 70–99% indoors if devices are installed | 20–50% outdoors 70–99% indoors if devices are installed | 30–50% outdoors 70–90% indoors | 0% outdoors 50–80% indoors if devices are installed |
| Positioning mechanisms (see Section 13.3 for details) | TOA, TDOA, AOA/DOA, AGNSS, CGNSS | Cell-ID, TOA, TDOA, AOA | TOA | RSS, RTT | RSS, AOA | TOA, TDOA | dead reckoning, etc. | RSS |
| Main references | [10][15][18][55][70] | [17][98] | [8][42][52][74] | [63][119][54] | [89] | [89] | [89][119][97] | [61] |

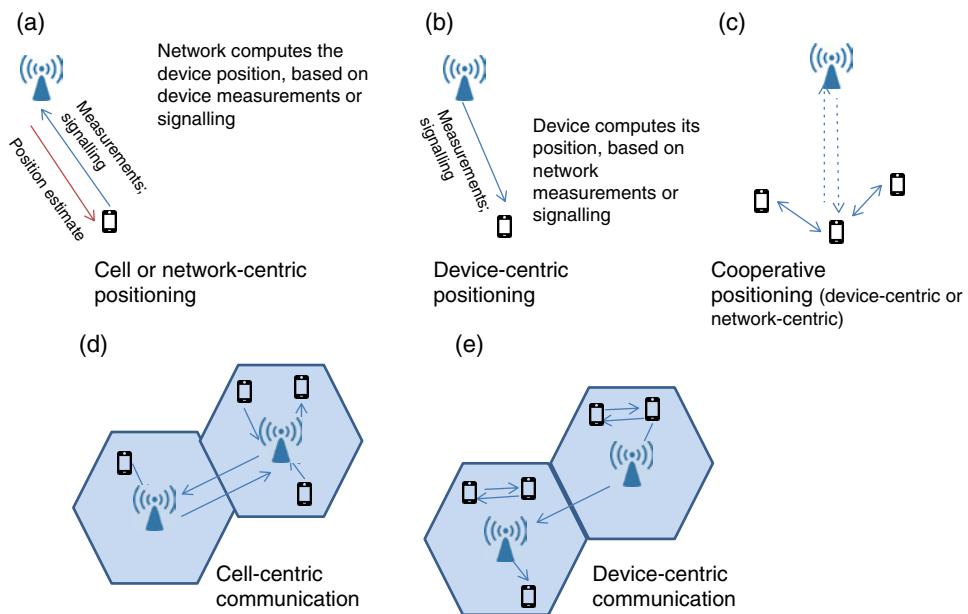


Figure 13.2 Positioning-communication dichotomy in 5G: (a) cell- or network-centric positioning; (b) device-centric positioning; (c) cooperative positioning (device-centric or network-centric); (d) (network) or cell-centric communication; and (e) device-centric communication.

cooperative positioning. This type of positioning can be implemented in both network-centric and mobile-centric approaches.

On the other hand, in cell-centric communication architectures (Figure 13.1 (d)), the mobile devices always communicate with one or several Access Nodes (ANs) that serve their geographical area. In a device-centric architecture, (Figure 13.2 (e)), the mobile devices can communicate with either only between themselves or both between themselves and with the ANs. Thus, in communication architectures, device-centric or cooperative communications refer to the same architecture; also such architecture is also referred to as D2D or Machine-to-Machine (M2M) architecture.

13.4 Brief Overview of 5G Positioning Mechanisms

The key features of the 5G physical layer include a ultra-dense network of ANs, large receiver bandwidths, receive beamforming with large antenna arrays (massive MIMO), and support of device-centric architectures [10,52,75]. It is also expected that the 5G receivers will operate to high carrier frequencies, such as mm-wave communications ranging from 30 GHz to 300 GHz, and will always be under Line Of Sight (LOS) conditions with at least two transmitters or ANs, due to the expected high density of the ANs [52,70]. The ubiquitous LOS presence, the high density of the ANs and the large available bandwidths make 5G an ideal system to achieve sub-meter level positioning accuracy, as the positioning accuracy is directly related to the size of the available bandwidth and the probability of LOS conditions [10,70]. The current 5G white papers do not

Table 13.2 Summary of positioning mechanisms in 5G.

| Positioning mechanism | Time of Arrival | Time difference of arrival | Direction of Arrival | Received Signal Strength | Assisted and cloud Global Navigation Satellite System |
|------------------------|-----------------------------------------|--------------------------------------------|----------------------|----------------------------------------------|-------------------------------------------------------|
| Abbreviation | TOA | TDOA | DOA | RSS | AGNSS and CGNSS |
| Underlying idea | Trilateration (intersection of circles) | Trilateration (intersection of hyperbolae) | Phase differences | fingerprinting, path-loss statistical models | GNSS plus cellular-based or cloud-based information |
| Illustration, if shown | Figure 13.3 a) | Figure 13.3 b) | Figure 13.4 b) | Figure 13.5 | Figure 13.6 |

specify any particular positioning technology that has to be used with 5G devices; they only focus on the target accuracy of less than 1 m in all suburban environments where 5G is available [4].

However, the 5G research papers dealing with positioning focus mostly on TOA [10,55], TDOA [17] and DOA [10,55,91]. Very few papers also mentioned the RSS in the context of 5G positioning [43], while some others also talked about Assisted-GNSS (AGNSS) in 5G [60]. In addition, Phase Difference of Arrival (PDOA) has been mentioned in the context of wireless localization [15,61].

These various positioning mechanisms are summarized in Table 13.2 and described briefly below.

In TOA and TDOA, the position estimate is based on the distance between the receiver and at least 2 (TDOA) or 3 (TOA) transmitters or Access Nodes (AN). TOA measurements require that the clocks of the ANs and the mobile device are synchronized. In many cases, it is difficult to synchronize accurately mobile devices to the time of ANs. However, if the ANs can be synchronized together, the problem of clock difference between the mobile and the ANs can be circumvented by adding one more AN and TOA to the system. This allows estimation of the clock difference together with the position coordinates. TDOA does not require the clock of the mobile to be synchronized with these. With TDOA, we obtain the difference between the distances to two ANs.

An example of the Cramer Rao Lower Bound (CRLB) for the delay error standard deviation in Additive White Gaussian Noise (AWGN) channel, as a function of the available bandwidth, is shown in Figure 13.4 (a). Clearly, the higher the bandwidth, the better delay tracking accuracy we can achieve. In 5G, as the bandwidths are expected to be of the order of tens or hundreds of MHz, the TOA-based estimators have the ability to reach very fine accuracy, even at ns level for standard deviation of error. One problem in TOA and TDOA estimation is the presence of Non Line of Sight (NLOS) situations. There are many solutions to perform NLOS detection and to eliminate NLOS paths [18, 46, 50, 88]. In addition to the measurement accuracy of TOA/TDOA, the positioning accuracy is also affected by the measurement geometry, that is, how the ANs and the mobile are geometrically located with respect to each other.

The DOA is estimated by measuring the difference in the received phase at each element of the antenna array. Antenna arrays can be configured into various types of geometry.

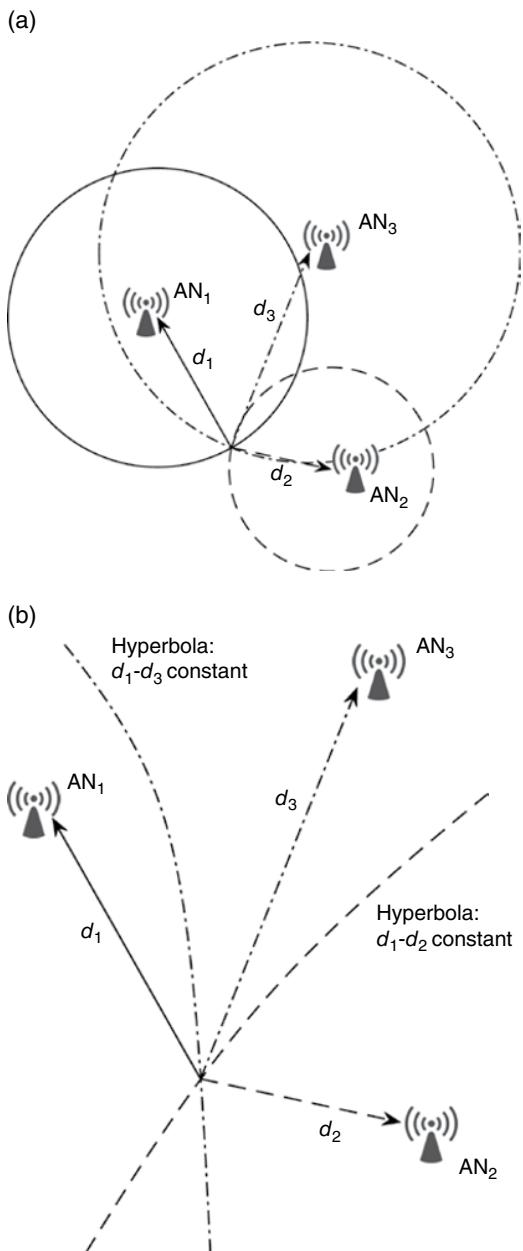


Figure 13.3 Principles of TOA, TDOA and AOA- based positioning: (a) Time of Arrival (TOA) trilateration; (b) Time Difference of Arrival (TDOA) hyperbolic positioning.

The DOA estimation has been an active area of interest in the field of radar, sonar, electronic surveillance and seismic exploration, and more recently of course in mobile radio communication [65,116]. The estimation methods for DOA include spectral estimation, minimum-variance distortionless response estimator, linear prediction,

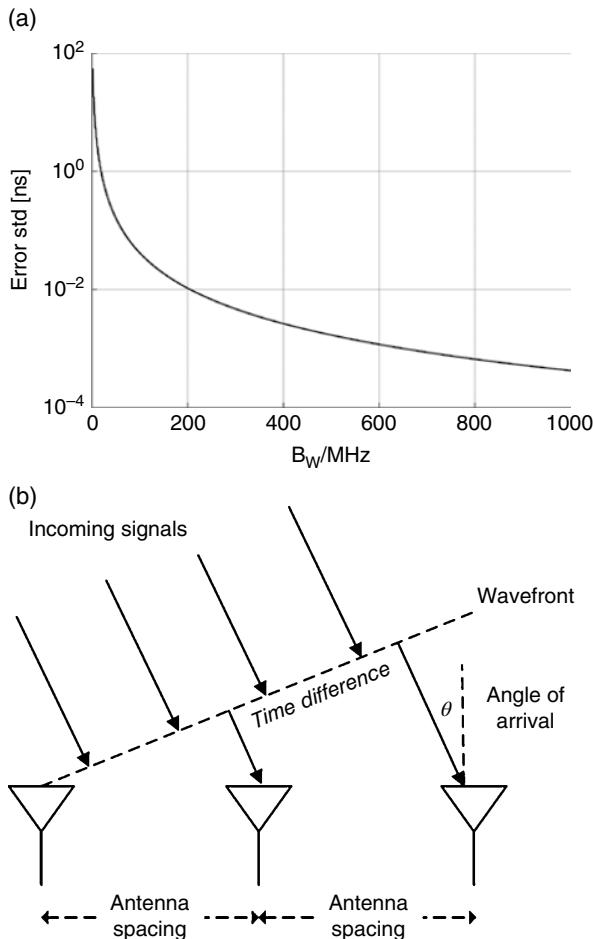


Figure 13.4 Example of TOA accuracy and principle of AOA-based positioning: (a) Example of TOA accuracy versus bandwidth, CRLB bound; and (b) a linear antenna array in measuring the angle-of-arrival of an incoming signal.

maximum entropy, maximum likelihood, as well as various Eigen-structure methods such as many versions of the MUSIC (Multiple Signal Classification) algorithms, minimum norm methods, the ESPRIT (Estimation of Signal Parameters via Rotation Invariance) method, and the weighted subspace fitting method, as discussed in [66].

RSS-based positioning is another range-based positioning that relies on the fact that the RSS is proportional to the distance between the transmitter (e.g. AN) and receiver (e.g. mobile device). The relationship between RSS and communication distance is known under the name of the path-loss model and there are many path-loss models to characterize the signal propagation. The simplest and most generic one is the one-slope path-loss model below, which shows the RSS in dB in terms of the logarithmic distance [110]:

$$\text{RSS}[\text{dB}] = P_{1m}[\text{dB}] - 10n \log_{10}(d) + \eta \quad (13.1)$$

where $P_{1m}[\text{dB}]$ is the transmit power at 1 m away from the transmitter (in dB), n is the path-loss coefficient and η is a noise term, typically modeled as zero-mean Gaussian,

which models the shadowing and fading effects over the wireless channel. If the RSS from several transmitters is measured and if the path-loss coefficient n and the transmit power $P_{1m}[\text{dB}]$ are known or estimated, then the distance to several transmitters can be estimated, and the position can be computed via trilateration, in a similar way as with TOA trilateration. However, due to typically high shadowing variances, the n and $P_{1m}[\text{dB}]$ estimates are not accurate. Moreover, as 5G are likely to operate at high carrier frequencies such as mm-wave bands, and as such bands have not yet been measured or understood properly, path-loss models for 5G are still not well known. A few 5G path-loss examples can be found here [11,104].

A higher accuracy solution can be obtained by so-called *fingerprinting*. In fingerprinting, it is first necessary to measure the RSS in various geographical points and to store such measurements in a database, called the training database. Such a database is typically formed and maintained by the LISPs from Figure 13.1. Then, in the estimation phase, the new RSS measurements are compared to the training database, according to a selected similarity measure, such as Euclidian distance or log-Gaussian likelihood, etc., and the position estimate is taken as the position of the training point that has the most similarities with the new RSS measurements. Figure 13.5 illustrates the steps involved in a fingerprinting process, from offline training database collection and forming to the online positioning estimation based on mobile measurements and relevant parameters from the training database.

The advantage of fingerprinting is that no path-loss models are required and the accuracy of the position is typically higher than in path-loss approaches. The main drawback is the need of creating and maintaining in a timely fashion a training database in environments that are highly changing by their nature. Another drawback is that large-scale location estimates (e.g. at country level or continent level) would require huge databases and huge computational power, and thus large-scale fingerprinting becomes easily unsuitable for mobile-centric positioning solutions, which are the ones more suitable for privacy preservation.

AGNSS refers to a technology where time, frequency, orbit and clock parameters, as well as navigation data bits, are provided as assistance via a mobile network to a GNSS

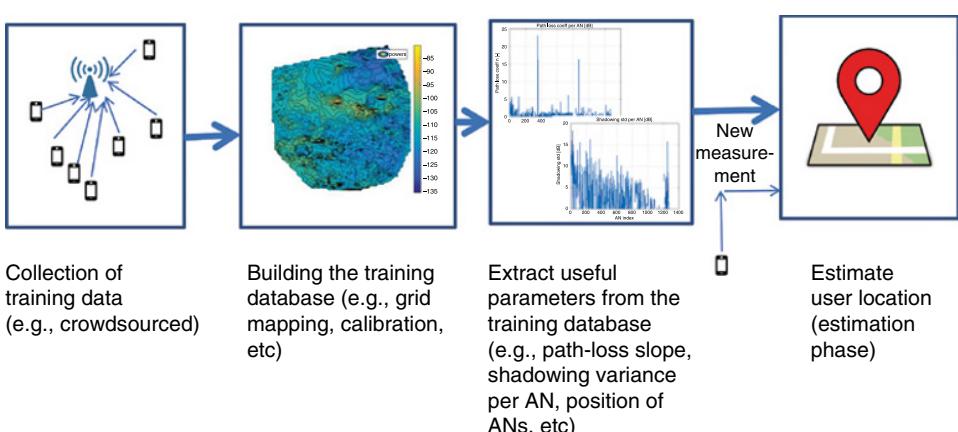


Figure 13.5 Fingerprinting principle in 5G positioning.

receiver, leading to significant improvements in the time-to-first-fix and signal sensitivity. In AGNSS, high sensitivity receivers rely on assistance data, including time, approximate position, satellite ephemerides, and possibly also code differential GNSS corrections to increase availability and accuracy [53,74]. In principle, as brought up in [42], AGNSS works by giving the receiver a hint of which frequency bins to search for when acquiring the signal. AGNSS and its standardization as well as harmonization are discussed more deeply in [64].

CGNSS is a recent and new paradigm used in conjunction with modern wireless receivers [8,112], and refers to the situation when most of the computations regarding the GNSS-based position estimation are no longer computed with the mobile resources, but rather in a remotely located cloud. The cloud thus undertakes the energy consuming tasks and the receiver can access the cloud-based solution via a web portal. In CGNSS, the device itself often does not even know its position. The cloud server (i.e. LISP is the cloud server in this case) collects measurements (e.g. GNSS observables) from the mobile devices, processes the measurements in a conglomerate manner, for example, allowing indoor users to benefit from the measurements collected by the nearby outdoor users, and it computes the mobile location. The mobile is continuously tracked at the cloud side, but this architecture does not impose the position to be sent back to the mobile. The simplified block diagram of the cloud GNSS concept is shown in Figure 13.6 [8,112]. The users send GNSS raw data to the cloud server (or LISP) and the LISP computes the user location. All the computationally intensive processing takes place at the server side.

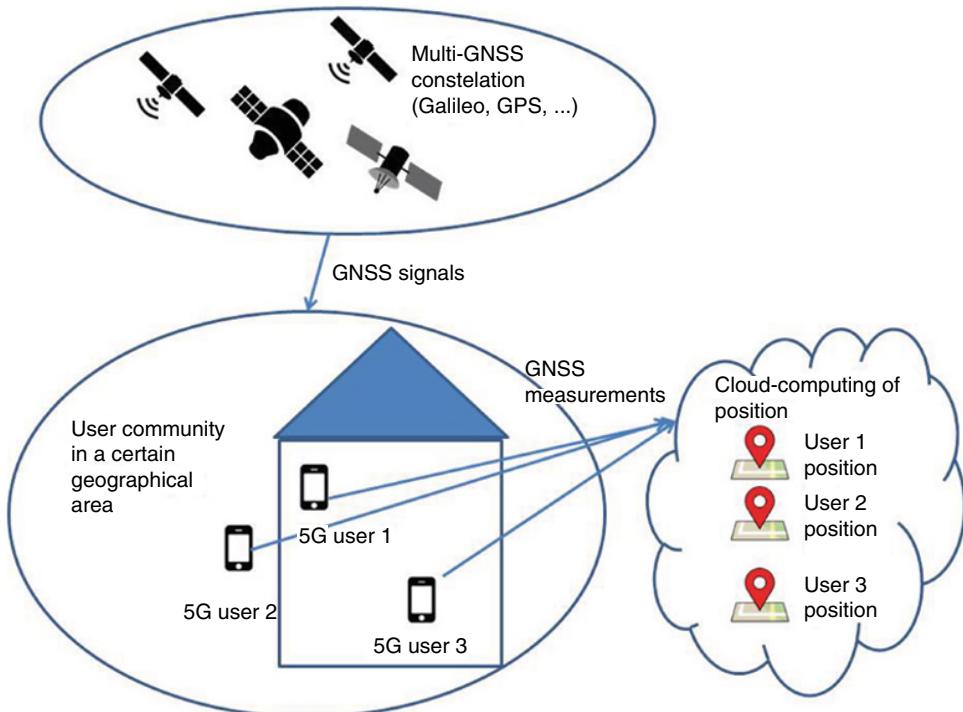


Figure 13.6 Cloud-GNSS positioning.

13.5 Survey of Security Threats and Privacy Issues in 5G Positioning

We have seen so far the main players in 5G positioning and the main positioning technologies for 5G. In what follows, we will address the security and privacy threats according to each of the 5G positioning players, as shown in Figure 13.1. We group here the security-related vulnerabilities in 5G into two main classes:

- vulnerabilities related to the reliability and integrity of the positioning solution in the presence of interferers, attacks or unintentional errors. We will refer to this class under the generic name of “security threats” and they are addressed in Section 13.3.1.
- vulnerabilities related to the privacy of the users’ location solution. We will refer to this class under the generic name of “privacy concerns” and they are addressed in Section 13.3.2.

13.5.1 Security Threats in 5G Positioning

Security threats can be further divided according to the 5G positioning player in the block diagram of Figure 13.1. LIC and the end-user are treated jointly, as these security threats are common to both. Also, some of the security threats are common to several players.

13.5.1.1 Security Threats Affecting Several or All Players

These kinds of security threats are also common to other wireless networks, not only to 5G, and both from the communication and localization aspects. In general, the information threats to wireless devices can be classified into passive and active attacks. The passive attacks consist of situations when attackers in the wireless networks attempt to grasp information via exploiting the network vulnerability, while the active attacks are those when the attacker is attempting to disrupt the network communication and also affect the user productivity in a network. Listed below are some of the most common types of security threats as discussed in [83,103]:

- A *Denial of Service attack (DoS)*: these are active attacks attempting to inhibit or prevent legitimate use of the wireless navigation or communication services;
- *Distributed DoS (DDoS)*: a distributed DoS is another active attack that occurs when multiple systems are used to flood the resources or bandwidth of a group of servers or one single server, such as LISP or LBSP from Figure 13.1. The main purpose of this attack is to saturate a resource so that it is no longer available for its legitimate use. It is often used as a decoy to hide a more malicious attack, which attempts to steal sensitive information or other data;
- *Man in the Middle (MiM) attack*: this is an active attack where an attacker intercepts the path of communications or positioning signaling between two legitimate parties, thereby obtaining authentication credentials and other data. A subclass of MiM is the *Message Modification (MM)*, when an attacker actively alters a legitimate message by deleting, adding to, changing or reordering it. The message can be, for example, the positioning signaling message between LISP and the end user;

- *Physical Attack/Firmware Replacement (PA)*: in a PA attack, also of the active type, an attacker has physical access to the device and can replace firmware or steal credential information such as static keys;
- *Eavesdropping (ED)*: in an ED attack, which is a passive attack, the attacker passively monitors the network communications for capturing communicating data and authentication credentials.

Here, we add a few notes about how security aspects have evolved from the first generation (1G) of mobile phones to 5G, and these again apply to both communication and navigation aspects. The first generation of mobile networks (1G) basically did not have any mechanism of encryption [58]. In early 2G systems, the mutual authentication between mobile users and the network did not exist, which left the possibility for an attacker to set up fake base stations and convince legitimate mobile devices to connect to [98]. In order to minimize exposure of user identifiers (known as International Mobile Subscriber Identifier or IMSI) in over-the-air signaling messages, 2G systems introduced the use of temporary mobile subscriber identifiers. However, in the absence of mutual authentication, fake base stations were used as “IMSI catchers” to harvest IMSIs and to track movements of users [16].

The security has significantly been strengthened by 3GPP (Third Generation Partnership Project) in 3G specifications, which introduced mutual authentication and the use of stronger and well-analyzed cryptographic algorithms [2]. LTE specifications further improved signaling protocols by requiring authentication and encryption (referred to as “ciphering” in 3GPP terminology) in more situations than was previously required. Previously known attacks, such as the ability to track user movement, were thought to be difficult in LTE [99], although recent work [106] proved that this problem can be solved. However, radio jamming, which generates an attack through a jammer by transmitting energy to disrupt reliable data communication, can still be a potential attack to LTE or 5G [72].

13.5.1.2 Security Threats Affecting LISP

From the LISP’s point of view (Figure 13.1), there are several potential sources of vulnerabilities in the location solution, which could hinder the robustness of the location estimate:

- 1) *The presence of malicious nodes in the system*: the malicious nodes are those nodes (fixed or mobile) sending fake or erroneous information to the LISP. Examples are the spoofing and meaconing nodes when AGNSS is used, fake ANs in 5G network or malicious mobile devices sending wrong or erroneous measurements to the 5G network. Spoofing here refers to the situation when a malicious node broadcasts a synthetic GNSS signal in order to try to trick the mobile AGNSS receiver into using the false signals and obtaining an incorrect position or time. Mecaoning here refers to the situation when a malicious node re-broadcasts real satellite signals after a brief delay, in order to create errors in the AGNSS unit on the 5G receiver;
- 2) *The intentional and unintentional interferences*: examples here are the jamming (i.e. broadcast of a narrowband interference signal) of the GNSS signal when AGNSS solutions are used for 5G positioning or narrowband and wideband interferences in 5G band, which may affect the quality of the measurements and signaling needed for positioning;
- 3) *Network-centric database deterioration*: this vulnerability applies to positioning techniques relying on a training database, such as RSS-based approaches.

13.5.1.3 Security Threats Affecting LBSP

From the LBSP's point of view, the main security threats come from:

- 1) *Unauthorized use of the Location Based Service*: for example, a user who did not pay the service would try to use it by accessing fraudulently the LBSP;
- 2) *Location leakage or theft*: user location information can leak accidentally or due to hacking of LBSP and such location leakage can adversely affect the user and its trust in LBSP. For example, knowing when a family is on holidays (based on their location) can create opportunities of house burglary if such information gets into malevolent hands. Or stealing the location identity of another user can allow one to ride freely on automatic toll highways, as the bill would be sent to another user;
- 3) *Lack of transparency in privacy policies*: an LBSP uses location information or content to provide a service to the user. Often, such a service is a web-based service, meaning that it requires access to the Internet. Often the LBSP developers rely on mixed third parties' data sources. For example, a location-based advertising application may use data about various shop offers in a certain shopping center, combined with customers' loyalty cards to that particular shop. Such a service could offer discounts to loyal users or to users passing in a certain time interval around that shop. The third-party unit can also require the user location, for example for storing up statistics about a particular user's shopping habits and these requirements might be into conflict with LBSP policy that claims that location data is only used anonymously. The LBSP should make it clear to what extent and what kind of user location is collected by the third parties (e.g. floor or building level versus meter accuracy position) and if such data can be associated with individual user profiles, and this should be made visible in the LBSP policies to the users. Also, a best approach would be when the user is given the possibility to choose how his/her location data is used and there are mechanisms to verify and reinforce the correct usage of the location data.

13.5.1.4 Security Threats Affecting the 5G User Device or LIC

From the users' point of view, some of the threats encountered at LISP and LBSP are also affecting the users, and some new threats appear. The main security threats on the users' side can be grouped into:

- 1) *The presence of malicious nodes in the system*: this affects both the LISP and the users in the mobile-centric positioning, as the location estimate relies on information collected from various nodes in the system;
- 2) *Low trustworthiness level in LISP or LBSP or both*: this affects user devices using both device-centric and network-centric localization. This may happen when the user relies, for example, on a cloud LISP or LBSP or on solutions involving crowd-sourced data. Trustworthiness is critical in some location-based applications involving emergency help, road assistance or billing (e.g. road tolling). The trustworthiness levels are often defined with respect to a certain target accuracy or availability. For example, if an LISP is trusted to provide a location accuracy of less than 5 m in 80% of cases, we cannot say if the same LISP can be trusted to provide a location accuracy of less than 0.5 m in 99% of cases;
- 3) *The intentional and unintentional interferences*: this affects both the LISP and the users in the mobile-centric positioning, as a low-quality measurement or low-quality signaling would deteriorate the location estimation or in extreme cases prevent the

LISP totally. Indirectly, these interferences also affect LBSP: the deteriorated location estimation may affect the service quality or deny it in the case that the access to the service is based on location;

- 4) *Location leakage or theft* [76]: this is when a user device reports a fake location; such location information can be used to wrongly identify the user and may contribute to identity thefts;
- 5) *Erroneous database transfer from network to the user*: this threat is valid for the location methods relying on a training database. Such a training database is typically collected by the 5G network and relevant parts of it are then transferred to the mobile for the positioning purpose (Figure 13.5). Errors on the communication line can affect the accuracy of the transmitted database, and thus the accuracy and robustness of the location;
- 6) *Location-tracking malware applications on the mobile device*: a recent market analysis by Symantec showed that more than 3 million applications on mobile phones could be classified as malware in 2016, which almost tripled the figures from 2015 [108]. Malware applications on mobile devices are thus increasing at a worrying pace. Such applications can steal various vital data from the user mobile phones, including location data. There is also the problem as to whether tracking a user location is ethical or not [12].

13.6 Main Privacy Concerns

According to [95], “privacy” is the ability of a person to control the personal information about oneself. Privacy threats refer to the limited capability or inability to control the usage of personal data, in this case location data. The risks of losing one’s location privacy can range from mild discomforts to serious dangers of burglary, thefts or even loss of life. Location data and places most frequented can disclose significant personal information, such as work and home places, attended schools, family plans (e.g. did you visit a fertility clinic?), health problems (e.g. did you visit a unit specialized in cancer treatment?), religious and sexual orientation (e.g. how often have you visited a certain worship place or a certain club), and so on: [95]

- 1) *Potential misuse of location information by LBSP or LISP or both*: when the users agree to certain distribution terms of using location-based services and localization engines on their mobile device; the terms are sometimes not very clear or too broad and may allow legal misuse of user location data [87];
- 2) *Unauthorized user tracking and other unauthorized use of location information*: when the user is giving permissions to his/her location data to LISP and LBSP, such location data sent, for example over Internet, has a huge potential to be misused, especially when the LISP or LBSP are not trusted players [111];
- 3) *“Right to be forgotten”*: the user location data is many times stored by the LSBP and LISP for indefinite periods of time, which are often not transparent to users. The right to be forgotten rule in Europe states that a person has the right to ask for the removal of his/her personal data, when such a data is no longer necessary for the purpose for which it had been initially collected [23,33,35]. However, the technical solutions to reinforce such a right to be forgotten are still vulnerable to unauthorized copying and dissemination of user data [90];

- 4) *Sending user location information to a remote location without users' consent:* this situation is also related to the trustworthiness of an LISP, LIC or LBSP; the users generally agree to certain distribution terms when using location engines and location services, but sometimes the distribution terms are not respected and the user location is distributed to third parties who were not initially pre-agreed with the users. This situation also refers to any inappropriate use of the location information by LBSP, LISP or LIC [76];
- 5) *Conflictual location policies:* depending on the application and service, the location policy can vary from being user-defined to provider-defined or institution-defined. In some cases, one or several of these location policies can enter in conflict with each other [111]. For example, an institutional policy can state that all workers in a certain building have to be located with a certain precision within the working hours and such information has to be kept visible only to authorized users, such as the building management, but the LBSP policy can state that the user himself or herself can either give permission to the location data to all or to no one. Such situations can usually be solved if there are common and flexible agreements on the location policies between different players in the system.

According to the study in [68], while user surveys usually show that users are highly concerned about their privacy, and in particular their location privacy, the users are typically not taking active steps to protect their location privacy.

13.7 Passive versus Active Positioning Concepts

The network-centric and user-centric localization concepts from the upper plots of Figure 13.2 are also closely related to the concepts of “passive” and “active” positioning. The question addressed in this section is about the feasibility of passive positioning schemes in 5G.

There are two different definitions encountered in the literature of passive versus active positioning:

- 1) *Definition 1 (adopted here)* [63]: Passive positioning schemes refer to the mobile-centric positioning schemes when the user device only receives positioning data, but it does not send full positioning data or data that could reveal the position of the user in a way that the user identity could be linked to that position. Such schemes can be used, for example, in conjunction with off-line downloaded maps to help the user navigation or user positioning on a map. In contracts, the traditional active positioning schemes in cellular communications rely on the bi-directional exchange of location data between the network and the mobile terminal. Such active positioning can take place either in a network-centric or a mobile-centric positioning architecture, but the network is always aware of the user location (with a certain accuracy and time granularity) in the active schemes. For example, a passive positioning scheme based on RSS has been recently proposed in [63]. The method, applied for WiFi-based positioning, relies on passively listening to WiFi beacon frames in monitor mode, without disclosing the user device ID (or MAC address). The RSS from several transmitters is collected from those passive beacons and a probabilistic approach, which relates the RSS to distances, is then employed to compute the user location.

While such a scheme can be directly applied on a 5G mobile device with a WiFi chipset by relying solely on the WiFi signal, a direct application of it to the RSS of 5G signals is not obvious. To enable it, the 5G networks should support some passive beacon modes, which are currently not found in 5G white papers.

- 2) *Definition 2 (mostly encountered in e-health related research)* [109,122]: Passive positioning may also refer to the device-free positioning schemes, where the user is not required to do anything in order to be positioned by the network. In contrast, active positioning means a positioning mechanism where the user carries some positioning device with him/her (e.g. mobile phone, wearable devices, etc.) and may be required to take some active steps to perform the positioning, such as turning on the GNSS or WiFi engine on his/her mobile device. A video-based positioning and a tactile floor are typical examples of passive positioning according to this second definition. This definition is in fact the opposite to the previous definition of passive positioning, and is not the one adopted in this chapter.

Clearly, the passive positioning schemes according to the first definition above can fully preserve the user location privacy. Another example of passive positioning, this time with TDOA, is discussed in [20]. The mobile device computes the TDOA from at least four ANs in range and computes its position based on some hyperbolic equations. The signaling sequences are not discussed in [20]. Again, such a downlink signaling only solution in 5G is highly unlikely, as the 5G network has to first verify the user identity for allowing access to the network and it is likely to use various location-aware mechanisms where the knowledge about the user location is a must [55,70,91].

In conclusion, passive positioning schemes in 5G remain are in contrast with the 5G targets of location-aware communications, mobility management in 5G and detection and tracking of Primary Users (PU). Thus, the passive positioning schemes are not likely to be a gaining technology in 5G.

13.8 Physical-Layer Based Security Enhancements Mechanisms for Positioning in 5G

This section focuses on the description of the main methods proposed so far at the physical layer to mitigate or eliminate the vulnerabilities and security threats in 5G positioning. Such methods include the reliability monitoring and outlier detection algorithms, the methods for detecting and locating interference signals, and backup systems.

13.8.1 Reliability Monitoring and Outlier Detection Mechanisms

From the reliability monitoring point of view, 5G positioning based on TOA has similarities to GNSS based positioning. Due to the dense network of ANs, it is likely that the TOA from the user is measured by 5 or more ANs, which makes the system of positioning equations over-determined. As the typical range of TOA estimation errors is also small compared to the TOA measurements, the redundancy in the over-determined equation system can be used to detect measurement errors in the positioning processing similarly as is used in integrity monitoring with GNSS signals [37]. However, there are also differences between 5G based positioning and GNSS. With GNSS, the positioning geometry is significantly different, as the distances to the satellites are much longer

than the distances to the ANs in 5G. This changes the positioning geometry and therefore also capability of the monitoring to detect errors. In 5G, it is assumed that there are also AoA estimates available, which provide even more redundancy, but also have different geometry effects to the fault detection function compared to TOA. If the number of available measurements is smaller than 5 or the measurement geometry does not allow positioning level integrity monitoring, the 5G positioning engine has to rely on quality indicators produced by the measurement process in the fault detection.

In positioning mechanisms relying on a training database, such as the RSS-based positioning or cloud GNSS (Section 13.4), the “health” of the training database, which is continuously updated, is of utmost importance. Thus, outlier detectors can be employed to detect malicious nodes or other spurious effects in the database. Outlier detection techniques have been widely studied by the statistics and signal processing communities [69] and similar approaches can be used to increase the training database robustness in 5G positioning. A good survey of temporal data outliers can be found, for example in [69]. Following the classification in [69], we can divide the outlier detection schemes into:

- 1) *Unsupervised discriminative approaches*: which rely on a certain similarity metric to identify the erroneous points in the database. Examples from this category are the clustering methods, the rank-based similarity methods (e.g. comparing the number and identity of the transmitters heard in a certain location), cosine similarity methods, correlation-based methods, etc.;
- 2) *Unsupervised parametric approaches*: which are based on building a statistical model for the available databases and computing the probability that a certain pattern, sequence or value belongs to the created model. Hidden Markov Modeling (HMM), for example, belongs to this category;
- 3) *Supervised approaches*: which are valid in the presence of pilot data or some data, which is highly reliable (e.g. data manually collected by the LISP may have more reliability than the data collected in a crowd-sourced mode). Examples here include rule-based classifiers, naive Bayes approaches or Support Vector machines (SVM).

For the estimation phase of fingerprinting, methods for detecting and removing erroneous RSS measurements have been presented, for example in [115], where non-iterative RANdom SAmple Consensus (RANSAC) is adopted to detect faulty RSS, and in [28], where an integrity monitoring method based on the “leave-one-out” approach is proposed. Fingerprint database management is considered in [118], in order to keep the crowd-sourced database consistent and scalable. In addition of snap-shot type of fault detection, where only current measurements and probably the database are considered at a time, it is also possible to use the time series properties of the positioning problem to detect anomalies in the positioning process. For this approach, the statistical prior knowledge of the user motion and position changes are used to formulate robust filter or change detector [48].

13.8.2 Detection, Location and Estimation of Interference Signals

Interference signal sources can, according to [105], be categorized into:

- 1) *Malicious interference*: defined as radio frequency interference (RFI) intentionally transmitted to prevent the use of the signals at hand or make the use hazardous for as many users as possible;

- 2) *Uninformed interference*: that comprises intentional transmission of signals at or near the signal frequencies in question, but without the desire to cause harm; and
- 3) *Accidental interference*: that includes unintentional transmissions appearing at or near the signal frequencies in question, typically from malfunctions.

Jamming is defined as the blocking of the reception of radio frequency signals, by deliberately emitting electromagnetic radiation to disrupt user receivers by reducing the signal to noise level [49]. In severe cases, jamming can lead to loss of signal tracking altogether.

Detection of interference signals is typically done by monitoring the received signal or automatic gain control (AGC) levels, with advanced signal processing in the radio front-end, by monitoring signal strengths, cross-checking against other signals and by monitoring the digitized signal levels [73]. Localization techniques of interference signals can be divided into four groups, according to the type of technology used [9]:

- 1) received signal strength techniques;
- 2) time of arrival techniques;
- 3) frequency techniques; and
- 4) phase and interferometry techniques.

Angle of arrival (AOA)-based geolocation techniques are suitable for all RFI types, but have high implementation complexity since they require phase/gain calibration of antenna array elements. AOA-based localization performance depends on the RFI bandwidth [9]. Time difference of arrival (TDOA)-based localization is suitable for wideband RFI and has low implementation complexity, but requires precise timing synchronization between sensor nodes [9]. Frequency difference of arrival (FDOA)-based techniques are best for narrowband RFI and require either the RFI or the detecting and locating sensor node to be moving, as well as both precise timing and frequency synchronization between the sensor nodes. Received signal strength (RSS)-based localization is suitable for all types of RFI and has very low complexity. However, RSS-based methods work poorly in sparse networks since performance degrades with distance.

13.8.3 Backup Systems

As now being included in 5G techniques, the digital broadcasting systems are able to be considered as one of the backup systems to enhance the security in wireless communication or radio navigation systems. Recently, digital broadcasting systems, such as the Digital Video Broadcasting (DVB), Digital Audio Broadcasting (DAB), and the Advanced Television Systems Committee (ATSC) standards, have been widely suggested to be used as an alternative information transmission technique [120]. The transmission power of digital TV (DTV) is high and the frequency band of DTV signals is wide from 400 MHz to 900 MHz [62], which occupies over 40 channels with one channel having a bandwidth of 8 MHz. It is common for one city to broadcast the DTV transmissions in 4 to 5 different channels, while the channels are sparsely allocated within all the DTV channels. Considering the security aspect, such properties of the DTV signals make the system more robust due to a wide bandwidth and multiple channels available in one local area. Since DTV facilities are already in use and no more infrastructure

investment is required, except on the receiver side, DTV would be a promising backup positioning system and thereby also a means to improve the security of a 5G system. Cloud and AGNSS, as discussed earlier, are also possible backup systems.

13.9 Enhancing Trustworthiness

Several trustworthiness metrics can be used in order to check the trust level of an LISP, LBSP or a mobile user, by any other of the localization players shown in Figure 13.1. These are:

- *Proximity metrics*: only a certain geo-spatial region could be allowed for a LISP, LBSP or end-user when receiving information from the other actors in Figure 13.1. For example, a user device located in Helsinki, Finland could reject all database information or all service providers that are not tagged in the Helsinki region (or in a smaller defined geographical region). Both distance proximity and temporal proximity metrics could be envisaged [57];
- *Authentication metrics*: the signals used for positioning could have authentication keys embedded within them, so that unauthorized or fake signals are automatically rejected. Similarly, the LBSP can offer services only to authorized users, identified through certain authentication procedures. More about authentication is discussed in Section 13.8;
- *Similarity metrics*: if the LISP or LBSP have access to the location information of many users, some similarity patterns between users located in close proximity to each other can be checked and the outliers can be thus detected and removed. Also similarity patterns with past user geo-location information, when such information is available, can be used for an increased trustworthiness of the user location data from the network side;
- *Privacy metrics*: such as the location uncertainty related to a particular user or the linkability of location information to the user who generated it [123].

13.10 Cryptographic Techniques for Security and Privacy of Positioning

Cryptographic solutions that are required for ensuring security and privacy of 5G positioning depend on the application and adversary models. We discuss three main scenarios:

- 1) The end-user, the network (LISP and LBSP), and the devices in their possession are trusted and only security against malicious outsiders is required;
- 2) The location information provided by the end-user (or in certain cases LISP or LIC) is not available or cannot be trusted (e.g. because the end-user itself has the incentive to provide a false location or it may be subject to malicious middle men (e.g. mafia fraud)) and, thus, the location (or the distance from a verifying node) must be verified; and
- 3) The network (LISP, LIC and LSBP) cannot be trusted and, thus, privacy of the end-user's location must be ensured.

All these scenarios lead to different security and privacy goals and require different cryptographic solutions. The four primary goals of cryptography are confidentiality, integrity, authenticity and non-repudiation [19]:

- 1) *Confidentiality*: (often referred to as secrecy) ensures that the information can be accessed only by authorized entities. Typically, this means that data is encrypted with a strong encryption algorithm (e.g. with AES [80]) that allows decryption only by an entity who has the secret key;
- 2) *Integrity*: protects data from unauthorized manipulation and allows the authorized entity to notice any manipulations. This can be achieved, for example, by computing a (cryptographic) check sum with a cryptographic hash function (e.g. with SHA-256 [79]);
- 3) *Authenticity*: provides proof that an entity is the one that it claims to be (e.g. with cryptographic challenge-response protocols) or that data originates from an entity it is claimed to originate from (e.g. with digital signatures);
- 4) *Non-repudiation*: prevents an entity from denying earlier commitments or actions. For instance, it prevents a person from later denying signing of a document. Non-repudiation can be achieved, for example, with digital signatures (e.g. with (EC) DSA [78]).

Traditional secret-key cryptography uses the same key for encryption and decryption and requires key exchange via secure channel prior to communication between the entities. Public-key cryptography [114] uses an asymmetric key pair where only the decryption key needs to be secret, but the encryption key can be public. This allows everyone to encrypt, but only an entity with the secret key can decrypt. The disadvantage of public-key cryptography is that it is significantly more computationally complex than secret-key cryptography. The following discusses certain cryptographic techniques that have been proposed for solving security issues in the aforementioned three scenarios or that may play a role in finding such solutions, although specific solutions are still missing

13.10.1 Cryptographic Authentication in Positioning

A repeating problem in the aforementioned 5G positioning scenarios is how to ensure authenticity and integrity of positioning signals and position information. If an entity, who knows its location, wants to share its location or (e.g. some other information which allows the recipient to derive its own location) in a reliable and secure manner, then it must ensure that any malicious party cannot tamper with the communication. The same problem is omnipresent in data communication and, in principle, integrity and authenticity of positioning can be solved with similar cryptographic techniques used elsewhere in communication. If two entities can share a secret-key via a secure channel, then they can use this key with standard cryptographic techniques to ensure authenticity and integrity of their communication, for example, with cryptographic message authentication codes (e.g. with HMAC [77]). If a key cannot be shared before deployment, then public-key cryptography can be used for key agreement or for providing integrity and authenticity with digital signatures. These cryptographic techniques are typically enough to avoid the threat of “unauthorized use of the location based service” (with standard user authentication)

and “location leakage or theft” (with techniques providing confidentiality and authentication), discussed earlier in Section 13.5.1.3.

Unfortunately, the above solutions can be problematic in positioning schemes because of limited communication or computation resources. Consider, for example, AGNSS systems, where communication is unidirectional (from a satellite to a receiver) and the bandwidth is very limited, preventing straightforward use of public-key digital signatures [44]. Also, a system-wide secret key is not a secure option, because building a tamper-proof device is extremely difficult (confer, side-channel attacks [121]). If such a system-wide secret key would leak from any of the millions of AGNSS receivers, then the whole system would be compromised. Solutions to this problem are available in the open literature, in particular, for the Galileo European GNSS system which is planned to offer authentication even for civilian use. Many of the proposed solutions are based on the TESLA scheme [14], which replaces the requirement to compute expensive digital signatures for each communicated message with considerably cheaper cryptography. Specifically, they authenticate messages with a secret-key message authentication code with a frequently changing key so that a key disclosure is delayed until the end of transmission with that key. Only the authenticity of the delayed key disclosure must be verified with expensive digital signatures and the rest is verified with the cheaper message authentication code. A solution based on TESLA that was derived specifically for Galileo is given in [93]. Similar solutions could be applied also in the case where a network includes trusted beacon nodes whose location is known (Section 13.6). In that case, the end-users’ devices can use the above techniques to ensure the integrity and authenticity of the messages from the beacon nodes, in a similar way to the satellites in the AGNSS setup.

13.10.2 Cryptographic Distance-Bounding

If an entity, called here the verifier (e.g. LISP), needs to verify the location of an untrusted entity, called here the prover (e.g. an end-user’s device), then the above schemes are no longer adequate, because the location provided by the prover cannot be trusted. Instead, the verifier needs to have the means to obtain undeniable proofs about the prover’s physical location.

Cryptographic distance-bounding protocols give an upper bound for the distance between two entities: a prover P and a verifier V . Typically, such distance-bounding protocols have been used successfully in RFID door access, road tolling, prisoner tagging or some wireless sensor network-based applications [47,67].

Here we present the cryptographic distance-bounding protocol introduced by Brands and Chaum [101]. P generates k uniformly distributed bits m_i , $i = 1, \dots, k$, and commits to them using a cryptographically secure commitment scheme. The commitment prevents P from changing m_i and allows V to verify this in the end of the protocol. Also V generates the uniformly distributed bits a_i with $i = 1, \dots, k$. After this, the actual distance-bounding phase takes place by repeating the following steps for $i = 1, \dots, k$:

- V sends the bit a_i to P ;
- P sends the bit $b_i = m_i \oplus a_i$ to V immediately when it receives a_i ; where \oplus stands for the exclusive-or (*xor*) operation; and
- V measures the time t_i between sending a_i and receiving b_i .

Because signals cannot travel faster than light, the distance between P and V is at most $d = tc/2$, where $t = \max(t_i)$ and c is the speed of light. In the end, P opens the commitment by sending a , b and m to V (signed with P 's secret key using a secure digital signature scheme to ensure integrity and authenticity). V accepts d as the maximum distance if m matches the commitment and the signature is valid. The fact that b_i depends on a_i prevents P from sending b_i before it has received a_i , the commitment to m prevents P from fabricating an appropriate m' afterwards, and the signature prevents an imposter P' from acting as P . The above distance-bounding protocol is depicted in Figure 13.7.

The distance-bounding is very sensitive to processing delays. The delay $t = t_s + t_p$, where t_s is the time that the signal travels from V to P and back and t_p is the time spent in processing (e.g. computing the xor and delays in radio transceivers). The distance-bounding protocol gives meaningful results only if $t_s >> t_p$. The signal travels at about 30 cm in one nanosecond, so a very small t_p is needed in order to ensure the accuracy required by 5G. This forms major challenges for implementing accurate distance-bounding protocols and requires specific hardware solutions [5].

Distance-bounding provides only the distance between two entities, but triangulation allows several (mutually trusted) verifiers to derive the exact position of P (see, for example [102]). Distance-bounding can be also integrated into cryptographic user authentication protocols and it is a central part also in securing cooperative positioning systems where peers locate each other without trusted parties [102]. Cryptographic distance-bounding could complement the techniques discussed in Section 13.2 by preventing, for example, various man-in-the-middle attacks.

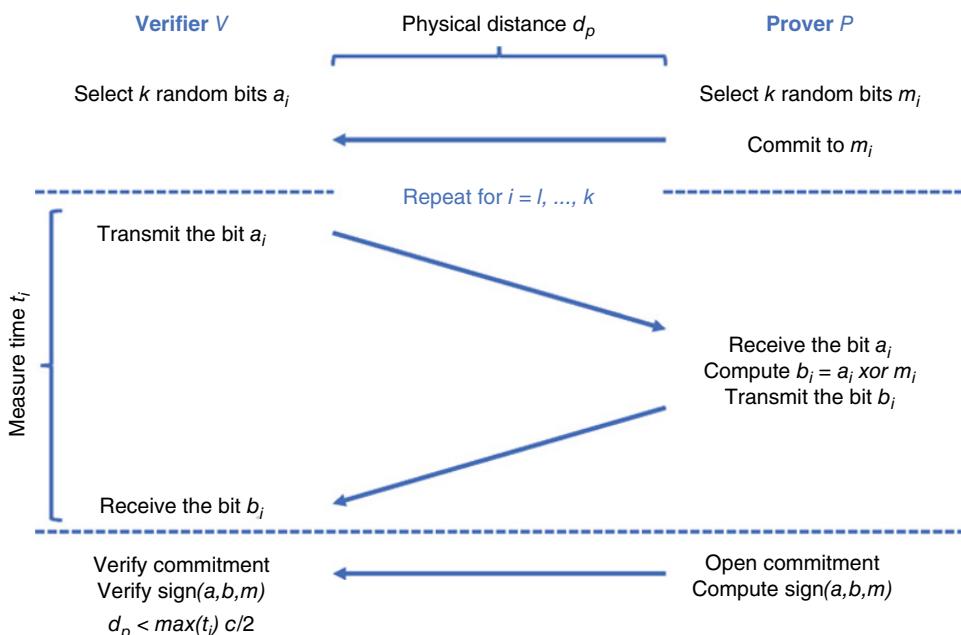


Figure 13.7 Brands and Chaum's distance-bounding protocol [101].

13.10.3 Cryptographic Techniques for Privacy-Preserving Location-based Services

Privacy is a major concern in location-based services. Current services set a lot of trust on the LBSP, because they typically require full disclosure of the end-user's location. Cryptography also has potential to improve the privacy of location-based services. Fully Homomorphic Encryption (FHE) [22], the greatest triumph of recent theoretical cryptography, offers a promise of privacy-preserving cloud computing and, consequently, also privacy-preserving LBSP. Traditional encryption requires data to be decrypted before it can be processed, but FHE allows computing arbitrary functions with encrypted data so that results are correct after decryption. If the end-user uses FHE to encrypt its location, then the LBSP (and LISP) can do computations with this data so that the result can be decrypted only by the end-user, as shown in Figure 13.8.

Unfortunately, FHE is computationally too demanding for practical purposes. In practice, the capability to perform arbitrary computations can be traded for better performance. Certain cryptosystems are partially homomorphic so that they allow, for example, only additions with encrypted data (e.g. the Paillier cryptosystem [92]). Other (more expensive) alternatives are somewhat homomorphic encryptions, which allow both additions and multiplications (similarly to FHE), but so that the number of consecutive operations (multiplications) is limited, thus limiting the complexity of possible computations. Other related concepts, such as multiparty computation (e.g. [94]), which allows two (or more) parties to jointly evaluate a function without revealing their own inputs to each other, or functional encryption [29], which allows setting keys that allow decrypting a certain predefined function of the encrypted data but not the data itself, may also have a role in solving the privacy aspects of location-based services. Differential privacy techniques ([21]) that protect individual records in statistical datasets can also improve privacy of certain location-based services.

Because large-scale use of the aforementioned general privacy-preserving cryptographic schemes (and FHE in particular) can be too heavy for most practical applications, privacy must be ensured by other means. Typically, this means that schemes are carefully tailored for a specific use case. Examples of such can be found in the academic literature. For example, [51] presented techniques for privacy-preserving electronic toll pricing combining multiple cryptographic techniques (digital signatures,

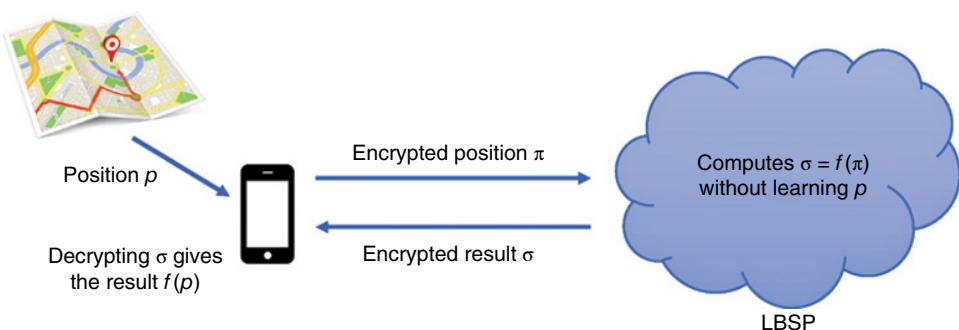


Figure 13.8 A privacy-preserving LBSP based on FHE, which computes a position-related function $f(p)$ on fully homomorphically encrypted data without learning the end-user's position p .

commitments, and zero-knowledge proofs). Another kind of application is the privacy-preserving proximity testing for social media (e.g. Facebook) presented in [13], which allows users of this social media to check if they are close to each other without learning the exact locations. Cryptographic techniques for privacy-preserving location-based services can be expected to be an important and active research topic in the near future.

13.11 Legislation on User Location Privacy in 5G

Legal and policy issues around 5G technology and 5G positioning range from spectrum requirements and standardization to privacy and data protection. Since 5G is seen as the enabler of a truly digital and networked society, privacy issues in the 5G context cover those related to the Internet of Things (IoT), cloud computing, big data, and mobile communications, among others. For their part, geolocation services may utilize the location sensitive sensors of smart mobile devices, and navigation is one area of application. Various infrastructures may be resorted to in the provision of geolocation services. With 5G, a new layer to existing infrastructures and services is envisioned [40].

For instance, the IoT implies data protection issues since also data, which is legally considered as “personal data”, might be collected and analyzed via various devices. Furthermore, location data might reveal sensitive data, especially when combined and analyzed over a long period of time while constant tracking and locating are apt to intrude personal space and autonomy [113].

The focus here is on the so-called “position privacy” as a specific area. However, in a European context, data protection terminology is employed alongside privacy. Indeed, in the EU, privacy and data protection are two separately safeguarded fundamental rights that are closely related [45,96]. The first is safeguarded in article 7 of the EU Charter of Fundamental Rights (CFR) [31] and the latter in article 8 thereof. Due to the developing EU fundamental rights dimension, the General Data Protection Regulation (GDPR) (EU) [35] 2016/679 builds on data protection whereas its predecessor, the Data Protection Directive (DPD) (95/46/EC), refers to privacy as enshrined in article 8 of the European Convention of Human Rights (ECHR). The GDPR has been in force since May 2016, but will be applicable only from May 2018. In the meantime, the Directive applies [35].

In the context of 5G developments and the potential applications accompanying it, one might ask what room, in practice, is left for privacy in a world where everything is supposedly digitized and connected. However, the fact that a number of cases concerning both personal data protection and the right to private life (e.g. Google Spain; Schrems; Uzun v. Germany; von Hannover v. Germany No. 2, etc.) [23] are brought before both the Court of Justice of the European Union (CJEU) and the European Court of Human Rights (ECtHR) does speak for the importance of privacy.

13.11.1 EU Policy and Legal Framework

In September 2016, the EU Commission published its action plan concerning 5G that aims at deploying 5G by 2020 [25]. This is tangential to the Digital Single Market objectives pursued in the EU. Both are also linked to the proposed European Electronic

Communications Code [26]. The Code aims at regulating the leeway for national regulators and providing for legal certainty in the telecommunications sector of the internet era. It covers networks and services and includes, among others, provisions on planning and coordination of spectrum policy.

The Data Protection Directive and the GDPR both apply to processing of personal data by (partly) automated means, as well as by other means where filing systems are formed (3(1) DPD; 2(1) GDPR). On the one hand, the territorial scope of application also overlaps when it comes to the establishment of the controller in the EU. A notable difference is that the GDPR as a legal instrument is a Regulation with direct applicability across the Member States, while the Directive relied on national implementation, and differing national laws came to exist. On the other hand, whereas the Directive refers additionally to making use of “equipment, automated or otherwise, situated on the territory of the said Member State, unless such equipment is used only for purposes of transit” (art 4(1)(c)), the GDPR speaks of offering goods and services to data subjects in the EU as well as monitoring their behavior where this behavior is taking place in the EU (art 3(2)) [35]. The latter approach more clearly applies to various tracking methods. However, already at present, technologies such as cookies fall under EU Regulation [71,117].

The GDPR does not apply to anonymous data whereby an individual is no longer identifiable; however, in assessing identifiability, all means “reasonably likely to be used” should be accounted for, which means considering the resources required and technology available (GDPR preamble, 26 [35]). Pseudonymous data in turn is covered. Pseudonymization means that attribution to a specific individual requires additional information, which is kept separately and secured (art 4(5)).

While the Data Protection Directive does not explicitly tackle “location data”, the GDPR does. Indeed, “personal data” is defined as “any information relating to an identified or identifiable natural person”, while:

...an identifiable natural person is one who can be identified, directly or indirectly [24,40], in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person” (art 4(1) GDPR).

This means that singling out a user of a specific device is sufficient [39]. The Regulation also refers to location in the context of profiling (GDPR preamble, 71, 75; art 4(4)) [35]. Nonetheless, already under the Directive, geo-location data from smart mobile devices as well as, for instance, the calculated location of a Wi-Fi access point combined with a unique identifier, are personal data [38].

In addition, as a specific instrument regulating data protection and privacy in electronic communications sector, the ePrivacy Directive (2002/58/EC [41], as amended by 2009/136 EC), is applicable with regard to base station data processed by operators – but operations on the application level only (i.e. independent from the telecom network) are out of the scope. Nonetheless, providers of other type of infrastructure, such as those relying on WiFi access points, remain within the scope of general data protection law, and the same applies to application providers and developers of operating systems [32,40]. The ePrivacy Directive is currently being reformed in the context of the

objectives noted above, while consistency with the GDPR is pursued. The purpose is to adopt a Regulation that extends the rules to non-traditional services, such as internet calls [124,125].

13.11.2 Legal Aspects Related to the Processing of Location Data

Processing of location data is considered a delicate issue and, thus, both general and specific instruments provide the legal framework for different aspects of processing location data. In the context of geo-location, there may be several controllers and processors, that is those determining “the purposes and means of the processing” and those processing on their behalf (art 2 DPD; art 4 GDPR) of data [35].

Processing of, including any operation performed upon, personal data requires a legal ground. Consent is one such ground (art 6 (1 a) GDPR; art 7 DPD), and it is applicable in the context of personal smart devices and geo-positioning data. Consent must be informed, freely given, specific with regard to the different purposes of processing, unambiguous and withdrawable (arts 4(11), 6, 7(3); preamble, 32 GDPR). A “clear affirmative act” is required, including ticking boxes or choosing settings (GDPR preamble, 32), while consent must also be “clearly distinguishable from the other matters” in the context of some wider declaration (art 7 GDPR). Thus, it is not sufficient to accidentally “consent” by lack of action or by accepting general terms and conditions of particular service [38,40].

Furthermore, processing of sensitive data (art 9 GDPR), as well as children’s consent (art 8 GDPR), are under specific requirements which, in the former case, refer to the explicit nature of the consent and in the latter, to parental authorization. Information on the details of processing must be provided to data subjects in understandable and accessible form, while taking children especially into account (preamble, 39, 58; arts 13-14 GDPR). Users must also remain informed and be reminded of their device being located – this could be best done in cooperation between app providers and developers of operating systems [38,40].

Furthermore, individuals as data subjects have rights (arts 16-22 GDPR; art 12 DPD), which need to be fulfilled. This means, for instance, access to location data in human readable format, as well as the possibility to rectify and erase data (incl. art 17 GDPR on “the right to be forgotten”), preferably online [32,40].

13.11.3 Privacy Protection by Design and Default

The GDPR includes explicit provisions on so-called data protection by design and default (art 25). This means that systems and services are to be designed so as to implement the principles of processing personal data, including data minimization and security (art 5). Technological and organizational measures are to be executed with a view on “the state-of-the-art, the cost of implementation and the nature, scope, context and purposes of processing”. While the risks for individuals’ rights and freedoms must also be taken into account (art 25(1)). These measures include pseudonymization as an example. Similarly, default settings should also support the principles, especially to ensure processing of necessary data only as well as appropriate storage (art 25(2)).

The Article 29 Working Party [36] has recommended that location services be switched off by default, that the scope of consent is limited in time (with a reminder

or renewal of a minimum of once a year even where no changes are planned) and sufficiently granular to enable precision of location data. Processing may only continue so far as necessary for the provision of a service. Service and app providers are thus to ensure that geolocation data or derivative profiles are deleted following justifiable storage periods – unless anonymized. In addition, third-party access should be logged [32,38,40].

13.11.4 Security Protection

Security of processing is regulated in article 32 GDPR [35]. Much like data protection by design, appropriate security must thereby be ensured by implementing technological and organizational measures, taking into account the state-of-the-art, costs and type of processing as well as the risks involved. Security measures include pseudonymization and encryption, abilities safeguarding confidentiality, integrity and restoration, as well as auditing processes.

13.11.5 A Closer Look at the e-Privacy Directive

For its part, the ePrivacy Directive notes that privacy and data protection must be safeguarded in the development of new applications, which rely on devices connected to publicly available networks or utilizing electronic communications services (Dir. 2009/136 preamble, 57). Restrictions on privacy in terms of identification may be imposed in national law to combat nuisance calls and also with regard to location data to enable emergency services (preamble, 36).

Definitions of “traffic data” and “location data” are included in article 2(2)(b)–(c) of the ePrivacy Directive, whereby the former “means any data processed for the purpose of the conveyance of a communication on an electronic communications network or for the billing thereof” and the latter “any data processed in [such a] network or by an electronic communications service, indicating the geographic position of the terminal equipment of a user of a publicly available electronic communications service.” Traffic data includes routing, duration, time, volume and format of a communication, the protocol in question, the location of the device, the network in question, and duration of a connection (ePrivacyD preamble, 15). Location data includes information such as latitude, longitude and altitude of the device, the direction of travel, the identification of the network cell in question, and the time stamp of the location information (ePrivacyD preamble, 14).

Traffic data may be processed by network and service providers for transmission and billing purposes following the principle of necessity, among others – after which it must be erased or anonymized (art 6 ePrivacyD). Users must be informed while prior consent is needed for processing carried out for marketing and value-added services by service providers (art 6(3)-(4)). Article 9 includes provisions on location data other than traffic data, whereby the requirements for processing include anonymization or informed consent. According to the Working Party, consent might concern a specific operation or a more comprehensive type of service, and providing geolocation data to third parties requires consent; the person consenting must also be the one whose data is concerned (e.g. confirmation messages). Consenting must precede sharing of location data by operators when they provide hybrid geolocation services using different types of

location data [38,40]. With the reform of the ePrivacy Directive, communications metadata, such as location data, would be made more available for provision of additional services also by traditional telecoms operators, provided that end-user consent is secured [124,125]. According to the Regulation proposal, “[l]ocation data that is generated other than in the context of providing electronic communications services should not be considered as metadata” (preamble, 17) [124,125].

Users must also be informed of transfers to third parties for value-added service provision, while subcontracting must be in compliance with data protection law (art 9(1); preamble, 32 ePrivacyD). Alongside consent withdrawal, users must have the possibility of “temporarily refusing the processing of such data for each connection to the network or for each transmission of a communication” (art 9(2); preamble, 35 ePrivacyD). Information on the processing and a possibility for revisiting it must be provided to those whose location data is being collected by those who collect it; this might be the provider of a value-added service or the operator [38,40]. In addition, the ePrivacy Directive includes in its article 5(3) requirements for storing information or accessing information stored on a device (so-called “cookie rule”): consenting in the meaning of data protection law is required, unless it is for transmission or imperative for providing a demanded service. In addition, article 4 of the Directive regulates security of processing in order to safeguard confidentiality and integrity of personal data in authorized use. However, as Minch notes [100], “[with] multiple sensors, device location tracking can reveal much more than a simple cookie [...]”.

The abundance of tracking techniques is acknowledged in the reform of the ePrivacy Directive. Moreover, the reform aims at simplifying the “cookie rule”, while also obliging adoption of tailored privacy settings following the principles of data protection by design and default [124,125].

13.11.6 Summary of EU Legal Instruments

A summary of EU legal instruments is shown in Table 13.3.

13.11.7 International Issues

5G development has strong international linkages. International organizations such as the International Telecommunication Union (ITU) are actively pursuing global approaches. Moreover, Public Private Partnerships (PPP) are encouraged in the area [107]. For its part, the EU is actively promoting 5G technology, as noted above. The EU is also cooperating closely with countries such as Japan and Brazil [37].

In addition to EU law already discussed, there are international instruments, such as the Council of Europe Convention on Data Protection and OECD Privacy Guidelines. However, these are out of the scope of this chapter.

As a comparison, the US system differs from the EU approach in that there is no general data protection law comparable to the GDPR and its predecessor, the Directive. Targeted laws, exist in silos and both on state and federal levels, alongside consumer protection patrolled by the Federal Trade Commission [71]. As a constitutional right, privacy is enshrined under the 4th amendment (US v. Katz), which protects people against unreasonable searches, among others. Privacy also operates on the so-called “reasonable expectations” doctrine, whereby public places and voluntary third-party disclosure imply that no such expectations of privacy exist [56,84]. However, the 4th amendment is being

Table 13.3 Summary of EU legal instruments.

| Instrument | What? | Who? | Legitimate processing |
|------------|----------------------------|---------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DPD | personal data | any controller any processor | <ul style="list-style-type: none"> ● (explicit) consent/legal ground ● information ● principles of processing ● data subject's rights ● security |
| GDPR | personal data | any controller any processor | <ul style="list-style-type: none"> ● (explicit) consent/legal ground ● information ● data protection by design ● data subject's rights ● security |
| ePrivacyD | traffic data/location data | telecom operator | <ul style="list-style-type: none"> ● legal ground/(prior) consent ● information ● user/subscriber rights ● security |

adapted to the digital realm and tracking technologies through case law (US v. Jones). This development has also led to geolocation technologies being viewed from the perspective of reasonable expectations [59]. If personal data are transferred from the EU to the US, the receiving organization based in the latter has to join the so-called Privacy Shield program to guarantee an adequate level of data protection (see art 25 DPD; art 45 GDPR [35]) or some other legal basis must exist (Ch. IV DPD; Ch. V GDPR [35]).

13.11.8 Challenges and Future Scenarios in Legal Frameworks and Policy

Whereas location-based services previously tended to be based on individuals' prior requests – whether for one-off use or more long-lasting use of location data (e.g. for navigation) – currently, many applications and services would rather rely on locating and tracking individuals [38]. This development is bound to accelerate with IoT and new technologies. Furthermore, the integrity of data becomes an important issue. With 5G, there are elevated risks involved with regard to unintentional disclosures of location data or data misuse and abuse; also where data is intentionally made available to service providers and app developers. Indeed, the constant evolution of technologies behind location-based application and services inherently poses risks with regard to location data and position privacy [1,40].

The legal framework provided by the general data protection law in the EU has a wide scope of applications, both in terms of substance and territory. Moreover, more specific issues are tackled, for instance in the ePrivacy Directive. On a global scale, or from a US perspective, the view might not be the same. Since 5G has international implications starting from spectrum and standards all the way to the global reach of potential applications, a lack of a common approach to privacy and data protection could be problematic.

Current policies around 5G development largely focus on investments, business, and technological issues, while in research papers privacy has also gained attention, mainly

as a dimension of technological solutions. Importantly, the creation and exploitation of location-based applications and services must comply with the legal requirements safeguarding data protection and privacy of individuals. Relevant regulation must be incorporated into every layer, starting from networks, operating systems, and access.

However, there are limits to what legal instruments, including the requirement of privacy by default, may achieve. Privacy may be enhanced via various control mechanisms, ranging from technical and legal measures to social control (e.g. social and business practices) [100]. Perhaps people should also appreciate their data more and be less willing to disclose their personal data in exchange for relatively small benefits. Then again, the usefulness of consenting and purpose limitation, and even data protection by design (especially data minimization), might be questionable in the context of IoT, cloud computing, and big data – that is, in the light of collecting massive amounts of data and having possible future uses that may even be of public benefit [1,7,34]. Nonetheless, one solution would be to strengthen users' control mechanisms. This could mean personal privacy assistants controlling the use of personal big data, encryption for identity verification and purpose specification, or biometrics for access control. With regard to unique identifiers, randomization could enhance privacy [6,7,34]. The working party has referred to identity management systems and measures to authenticate access requests, as well as centralized procedures for operators mediating between third-party service providers and users. These arrangements could enhance privacy by leaving individuals unidentifiable by third parties [38].

The steps towards the privacy and security of the location data from a legal perspective are summarized in Figure 13.9.

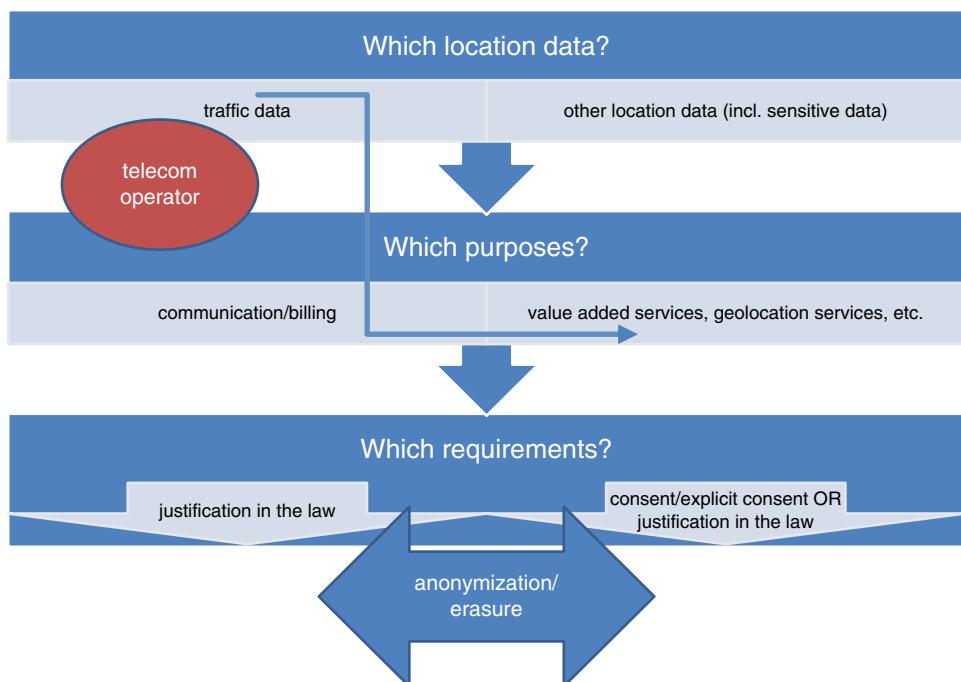


Figure 13.9 Chart of steps toward location privacy protection.

13.12 Landscape of the European and International Projects related to Secure Positioning

This last section focuses on recent projects at EU and international levels about vulnerabilities in wireless positioning, security in 5G, and solutions for secure and privacy-preserving positioning mechanisms. The projects are summarized in Table 13.4. A detailed descriptions about the NSF-funded projects can be found via the NSF search tool [82], by looking after the project number shown in Table 13.4.

Table 13.4 Summary of main EU and international projects somehow related to secure or privacy-preserving positioning.

| Project name and duration | Funding source | Main goals |
|--------------------------------------------------------------------------------------------|---------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Protecting Location Privacy in Location-Aware Computing, 2006–2010 | NSF grant #0627474 | Designing a secure and customizable architecture for privacy-aware location-based services and a suite of privacy protection techniques for reducing the risks of unauthorized disclosure of location information |
| Indoor TDOA Positioning Using a Narrowband RF Transceiver, 2012–2014 | NSF grant #1229899 | Developing a high-accuracy wireless positioning technology for indoor environment, which can be also used for location-based security |
| THE ISSUE, 2011–2014, http://www.theissue.eu/ | EU FP7-REGIONS | Improving the safety and security of citizens and offering innovative solutions for traffic management and urban mobility; real-time positioning, tracking is one of the components serving to reach the improved security and safety goals |
| CSFDA, 2014–2016 | EU FP7-PEOPLE (IIF) | Developing a location-based new modulation for physical-layer security in wireless communications |
| Secure Data Charging Architecture for Mobile Devices, 2014–2017 | NSF grant # 1422835 | Investigating the insecurity aspects of large-scale cellular network infrastructures, identifying their security loopholes, and sketching novel attacks that exploit such loopholes |
| CHARISMA, 2015–2017, http://www.charisma5g.eu/ | EU H2020-ICT-2014-2 | Providing a 5G end-to-end security service chain via virtualized open access physical layer security |
| 5GINSURE, 2015–2017 | EU H2020-ICT-2014-2 | Developing usable security enablers for 5G and initiating a 5G security testbed vision |
| 5G NORMA, 2015–2017, https://5g-ppp.eu/5g-norma/ | EU H2020-ICT-2014-2 | Developing a conceptually novel, adaptive and future-proof 5G mobile network architecture. Security aspects are ones of the addressed dimensions on 5G architecture |
| WiMi, 2015–2018 | NSF grant #1506657 | Creating a software-radio testbed (WiMi) to advance mmWave research and education |

(Continued)

Table 13.4 (Continued)

| Project name and duration | Funding source | Main goals |
|---------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| EAGER, 2017–2017 | NSF grant #1555332 | Providing fundamental knowledge about the 5G radio channels; indirectly this can serve to enhanced security and privacy mechanisms in 5G positioning and communications |
| 5G CHAMPION, 2016–2018, http://cordis.europa.eu/project/rcn/205441_en.html | EU H2020-EUK-2016-1 | Delivering 5G system proof-of-concept in conjunction with the 2018 South Korean Winter Olympics; the architecture will rely on navigation and secure communication aspects |
| SUCCESS, 2016–2018, http://www.success-energy.eu/ | EU H2020-DRS-2015 | Focusing on secure communications solutions using NFV, LTE and 5G technologies complemented by data privacy studies to ensure the acceptability of the results by consumers, also focusing on the vulnerabilities introduced by Smart Meters |
| PICASSO, 2016–2018, www.picasso-project.eu/ | EU H2020-ICT-2015 | Fostering the EU-US ICT policy dialogue by contributions related to, e.g. privacy, security, internet governance, interoperability, ethics |
| INSURE, 2016–2018, www.insure-project.org | Academy of Finland | Studying the security and privacy issues and solutions in GNSS and non-GNSS wireless positioning |
| EARS, 2016–2018 | NSF grant #1642920 | Addressing the key challenges in the development of signal processing algorithms, network protocols, and a prototype hardware design to enable scalable low-latency mm-wave MIMO networks with high degrees of spatial multiplexing |
| Enabling Ultra-Dense Future Cellular Networks (5G), 2016–2019 | NSF grant #1559483 | Fostering UK-US collaboration, by using a one of a kind, outdoor, large scale 5G test-bed for wireless cellular system innovation that has recently been established at the University of Surrey |
| EN 16803-2 & EN 16803-3, ongoing EU standardization work, 2015–2017 | EC and EFTA (European Commission Mandate M/496) | Standardization by CEN/CENELEC: Use of GNSS-based positioning for road Intelligent Transport Systems (ITS). Includes security attacks modeling and definition of performance features and metrics related to security and assessment field tests for security performances of GNSS-based positioning terminals for road Intelligent Transport Systems |

References

- 1 Internet of Things, Privacy & Security in a Connected World. FTC Staff Report, January 2015, pp. ii, 13–14, 35–37. Available at: <https://www.ftc.gov/system/files/documents/reports/federal-trade-commission-staff-report-november-2013-workshop-entitled-internet-things-privacy/150127iotrpt.pdf> (accessed 14 November 2016).
- 2 3GPP, *Third Generation Partnership Project*. Available at: www.3gpp.org/

- 3 5GPP-5G Forum (2016) 5G empowering vertical industries. White paper. Available at: https://5g-ppp.eu/wp-content/uploads/2016/02/BROCHURE_5PPP_BAT2_PL.pdf
- 4 5GPP-5G Forum (2015) 5G white paper: New wave towards future societies in the 2020s. March 2015 [Online]. Available at: http://www.5gforum.org/5GWhitePaper/5G_Forum_White_Paper_Service.pdf
- 5 Abu-Mahfouz, A. and Hancke, G.P. (2012) Distance bounding: a practical security solution for real-time location systems. *Proceedings of the IEEE Transactions on Industrial Informatics*, 9(1), 16–27.
- 6 Cavoukian, A. and Cameron, K. (2011) Wi-Fi positioning systems: beware of unintended consequences – issues involving the unforeseen uses of pre-existing architecture. Information and Privacy Commissioner, Ontario, Canada, June.
- 7 Cavoukian, A. (2015) Evolving FIPPs: proactive approaches to privacy, not privacy paternalism. In: *Reforming European Data Protection Law* (S. Gutwirth, R. Leenes and P. de Hert, eds), Springer pp. 293–309.
- 8 Favenza, A., Rossi, C., Pasin, M. and Dominici, F. (2014) A cloud-based approach to GNSS augmentation for navigation services. *Proceedings of the IEEE/ACM 7th International Conference on Utility and Cloud Computing (UCC)*, London, pp. 489–490.
- 9 Dempster, A.G. and Cetin, E. (2016) Interference localization for satellite navigation systems. *Proceedings of the IEEE*, 99, 1–9.
- 10 Hakkarainen, A., Werner, J., Costa, M., Leppänen, K. and Valkama, M. (2015) High-efficiency device localization in 5G ultra-dense networks: prospects and enabling technologies. *Proceedings of the IEEE 82nd Vehicular Technology Conference (VTC Fall)*, Boston.
- 11 Sulyman, A.I., Alwarafy, A., Seleem, H.E., Humadi, K. and Alsanie, A. (2016) Path loss channel models for 5G cellular communications in Riyadh City at 60GHz. *Proceedings of the 2016 IEEE International Conference on Communications (ICC)*, Kuala Lumpur, pp. 1–6.
- 12 Konstantinidis, A., Chatzimilioudis, G., Zeinalipour-Yazti, D., Mpeis, P., Pelekis, N. and Theodoridis, Y. (2015) Privacy-preserving indoor localization on smartphones. *Proceedings of the IEEE Transactions on Knowledge and Data Engineering*, 27(11), 3042–3055.
- 13 Narayanan, A., Thiagarajan, N., Lakhani, M., Hamburg, M. and Boneh, D. (2011) Location privacy via private proximity testing. *Proceedings of the 18th Annual Network & Distributed System Security Symposium (NDSS 2011)*, The Internet Society.
- 14 Perrig, A., Canetti, R., Tygar, J.D. and Song, D. (2000) Efficient authentication and signing of multicast streams over lossy channels. *Proceedings of the 2000 IEEE Symposium on Security and Privacy*.
- 15 Shahmansoori, A., Garcia, G., Destino, G., Seco-Granados, G. and Wymeersc, H. (2015) 5G position and orientation estimation through millimeter wave MIMO. *IEEE Globecom Workshops*.
- 16 Shaik, A., Borgaonkar, R., Asokan, N., Niemi, V. and Seifert, J.-P. (2016) Practical attacks against privacy and availability in 4G/LTE mobile communication systems. *Network and Distributed System Security Symposium (NDSS)*, February.
- 17 Tahat, A., Kaddoum, G., Yousefi, S., Valaee, S. and Gagnon, F. (2016) A look at the recent wireless positioning techniques with a focus on algorithms for moving receivers. *IEEE Access*, 4(99), 6652–6680.

- 18 Adebomehin, A.A. and Walker, S.D. (2016) Enhanced ultrawideband methods for 5G LOS sufficient positioning and mitigation. *Proceedings of the IEEE 17th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, Coimbra, pp. 1–4.
- 19 Menezes, A.J., van Oorschot, P.C. and Vanstone, S.A. (1996) *Handbook of Applied Cryptography*. CRC Press, Florida, 816 pp.
- 20 Leng, B. and Gao, T. (2014) A passive method of positioning indoor target based on TDOA. *Proceedings of the IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, Guilin, pp. 563–566.
- 21 Drowk, C. (2008) Differential privacy. *Proceedings of the International Conference on Theory and Applications of Models of Computation – TAMC LNCS*, 4978: 1–19.
- 22 Gentry, C. (2009) Fully homomorphic encryption using ideal lattices. *Proceedings of the 41st ACM Symposium on Theory of Computing (STOC)*, ACM. pp. 169–178.
- 23 CJEU Case: Google Spain SL and Google Inc. v. Agencia Española de Protección de Datos and Mario Costeja González (C-131/12; Grand Chamber 13 May 2014).
- 24 CJEU's Cases: Scarlet Extended SA (C-70/10; 24 November 2011); Breyer v. Bundesrepublik Deutschland (C-582/14; 19 October 2016).
- 25 COM(2016)588 Final: Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: 5G for Europe: An Action Plan.
- 26 COM(2016)590 Final: Proposal for a Directive of the European Parliament and of the Council establishing the European Electronic Communications Code (Recast).
- 27 Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (ETS No. 108), Strasbourg 1981. Available at: <https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/108> (accessed 10 November 2016).
- 28 Akopian, D., Melkonyan, A., Yalamanchili, S. and Chen, P. (2011) Integrated monitoring and mobile implementation aspects of WLAN positioning. *Proceedings of the ICSSE*, Macau, China, pp. 688–693.
- 29 Boneh, D., Sahai, A. and Waters, B. (2011) Functional encryption: definitions and challenges. *Theory of Cryptography – TCC. LNCS* 6597, 253–273.
- 30 Lohan, E.S. and Borre, K. (2016) Accuracy limits in multi-GNSS. *IEEE Transactions on Aerospace and Electronic Systems*, 52(5), 2477–2494.
- 31 EU (2016) *Charter of Fundamental Rights of the European Union* (OJ 2016/C 202/2).
- 32 EU (2013) Fifteenth Annual Report of the Article 29 Working Party (adopted 3 December 2013), pp. 8–9. Available at: http://ec.europa.eu/justice/data-protection/article-29/documentation/annual-report/files/2013/15th_annual_report_en.pdf (accessed 8 November 2016).
- 33 EU (2016) *Official Journal of the European Union*, Regulation EU 2016/679 of the European Parliament and of the Council, 27 April 2016. Available at: <http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=en>
- 34 EU Data Protection Supervisor (2016) EDPS opinion on personal information management systems. Towards more user empowerment in managing and processing personal data. Opinion 9/2016. Available at: https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2016/16-10-20_PIMS_opinion_EN.pdf (accessed 21 October 2016).

- 35 EU (2016) *General Data Protection Regulation (GDPR)*. Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0001.01.ENG&toc=OJ:L:2016:119:TOC (accessed 25 November 2016).
- 36 EU – Protection of personal data. Available at: http://ec.europa.eu/newsroom/just/item-detail.cfm?item_id=50083 (accessed 25 November 2016).
- 37 EU – Towards 5G. Available at: <https://ec.europa.eu/digital-single-market/en/towards-5g> (accessed 10 November 2016).
- 38 EU Working Party WP 115: Working Party 29 – Opinion on the use of location data with a view to providing value-added services. November 2005. Available at: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2005/wp115_en.pdf (accessed 11 November 2016).
- 39 EU Working Party WP 136: Opinion 4/2007 on the concept of personal data (adopted 20th June. Available at: http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf (accessed 11 November 2016).
- 40 EU Working Party WP 185: Opinion 3/2011 on Geolocation services on smart mobile devices (adopted 16 May 2011. Available at: http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2011/wp185_en.pdf (accessed 11 November 2016).
- 41 EUR-Lex, Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications). Available at: <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32002L0058>
- 42 van Diggelen, F. (2003) Indoor GPS theory and implementation. *Proceedings of the IEEE Position, Location, and Navigation Symposium*, Palm Springs, CA, April 15–18, pp. 240–247, IEEE Xplore.
- 43 Yu, F., Jiang, M., Liang, J., Qin, X., Hu, M. et al. (2014) Expansion RSS-based Indoor Localization Using 5G WiFi Signal. *Proceedings of the International Conference on Computational Intelligence and Communication Networks (CICN)*, Bhopal, pp. 510–514.
- 44 Caparra, G. et al. (2016) Design drivers and new trends for navigation message authentication schemes for GNSS systems. *Inside GNSS*, 11(5), 64–73.
- 45 Gonzalez Fuster, G. and Gutwirth, S. (2013) Opening up personal data protection: a conceptual controversy. *Computer Law & Security Review*, 29(5), 531–539.
- 46 Wang, G., Chen, H., Li, Y. and Ansari, N. (2024) NLOS Error Mitigation for TOA-Based Localization via Convex Relaxation. *IEEE Transactions on Wireless Communications*, 13(8), 4119–4131.
- 47 Hancke, G.P. and M.G. Kuhn, M.G. (2005) An RFID distance bounding protocol. *Proceedings of the First International Conference on Security and Privacy for Emerging Areas in Communications Networks (SECURECOMM'05)*, pp. 67–73.
- 48 Pesonen, H. (2013) *Bayesian Estimation and Quality Monitoring for Personal Positioning Systems*. PhD thesis, Tampere University of Technology, February.
- 49 de Castro, H.V., van der Maarel, G. and Safipour, E. (2010) The possibility and added-value of authentication in future Galileo open signal. *Proceedings of the 23rd International Technical Meeting of the Satellite Division of the Institute of Navigation*, ION GNSS 2010. Portland, OR, September.

- 50 Sharp, I. and Yu, K. (2016) Improved indoor range measurements at various signal bandwidths. *IEEE Transactions on Instrumentation and Measurement*, 65(6), 1364–1373.
- 51 Balasch, J. et al. (2010) PrETP: Privacy-preserving electronic toll pricing. *Proceedings of 2010 USENIX Security Symposium*, pp. 63–78.
- 52 Nurmi, J., Lohan, E.S., Wymeersch, H., Seco-Grandados, G. and Nykänen, O. (eds) (2017) *Multi-Technology Positioning*, Springer International Publishing, Cham, ZG, Switzerland, 348 pp.
- 53 Syrjärinne, J. (2001) *Studies of Modern Techniques for Personal Positioning*. Doctoral dissertation, Tampere University of Technology.
- 54 Talvitie, J., Renfors, M. and Lohan, E.S. (2016) Novel indoor positioning mechanism via spectral compression. *IEEE Communications Letters*, 20(2), 352–355.
- 55 Werner, J., Costa, M., Hakkarainen, A., Leppanen, K. and Valkama, M. (2015) Joint user node positioning and clock offset estimation in 5G ultra-dense networks. *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, San Diego, CA, pp. 1–7.
- 56 Whitman, J.Q. (2004) The two Western cultures of privacy: dignity versus liberty. *The Yale Law Journal*, 113(6), 1151–1221.
- 57 Nurse, J.R.C., Agrafiotis, I., Creese, S., Goldsmith, M. and Lamberts, K. (2013) Building confidence in information-trustworthiness metrics for decision support. *Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, Melbourne, VIC, pp. 535–543.
- 58 Nohl, K. and Munaut, S. (2010) Wideband GSM sniffing. *Proceedings of the 27th Chaos Communication Congress*. Available at: <http://goo.gl/wT5tz>
- 59 Pomfret, K. (2016) *Implications of Evolving Expectations in the United States. GNSS & the Law: Geolocation Privacy*, pp. 46–49. Available at: <http://www.insidegnss.com/node/5095> (accessed 3 November 2016).
- 60 Radnosrati, K., Gunnarsson, F. and Gustafsson, F. (2015) New trends in radio network positioning. *Proceedings of the 18th International Conference on Information Fusion (Fusion)*, Washington, DC, pp. 492–498.
- 61 Thangarajah, K., Rashizadeh, R., Erfani, S. and Ahmadi, M. (2012) A hybrid algorithm for range estimation in RFID systems. *Proceedings of the 19th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, Seville, pp. 921–924.
- 62 Chen, L., Julien, O., Thevenon, P., Serant, D., Pena, A.G. and Kuusniemi, H. (2015) TOA estimation for positioning with DVB-T signals in outdoor static tests. *IEEE Transactions on Broadcasting*, 61(4), 625–638.
- 63 Schauer, L., Dorfmeister, F. and Wirth, F. (2016) Analyzing passive Wi-Fi fingerprinting for privacy-preserving indoor-positioning. *Proceedings of the International Conference on Localization and GNSS (ICL-GNSS)*, Barcelona, pp. 1–6.
- 64 Wirola, L. (2010) *Studies on Location Technology Standards Evolution in Wireless Networks*. Doctoral dissertation, Tampere University of Technology.
- 65 Godara, L.C. (1997) Application of antenna arrays to mobile communications. Part II: Beam-forming and direction-of-arrival considerations. *Proceedings of the IEEE*, 85(8), 1195–1245.
- 66 Godara, L.C. (2004) *Smart Antennas*. CRC Press, Florida, 472 p.
- 67 Kuhn, M.G. (2006) Positioning security from electronic warfare to cheating RFID and road-tax systems. *Proceedings of the ESCAR Conference*, Florida. Available at: <http://www.cl.cam.ac.uk/~mgk25/escar-2006.pdf>

- 68 Gašparović, M., Nicolau, P., Marques, A., Silva, C. and Marcelino, L. (2016) On privacy in user tracking mobile applications. *Proceedings of the 11th Iberian Conference on Information Systems and Technologies (CISTI)*, Las Palmas, pp. 1–6.
- 69 Gupta, M., Gao, J., Aggarwal, C.C. and Han, J. (2014) Outlier detection for temporal data: a survey. *Proceedings of the IEEE Transactions on Knowledge and Data Engineering*, 26(9), 2250–2267.
- 70 Koivisto, M., Costa, M., Hakkarainen, A., Leppänen, K. and Valkama, M. (2016) Joint 3D positioning and network synchronization in 5G ultra-dense networks using UKF and EKF. Accepted for publication in *IEEE International Workshop on Localization and Tracking: Indoors, Outdoors, and Emerging Networks (GLOBECOM Workshops)*, December 2016
- 71 Paez, M. and La Marca, M. (2016) The Internet of Things – emerging legal issues for businesses. *Northern Kentucky Law Review*, 43(1), 29–71.
- 72 Svensson, M., Paladi, N. and Giustolisi, R. (2015) 5G: Towards secure ubiquitous connectivity beyond 2020. *ICS Report*. Available at: http://soda.swedishict.se/5933/1/T2015_08.pdf
- 73 Bhuiyan, M.Z.H., Kuusniemi, H., Söderholm, S. and Airos, E. (2013) The impact of interference on GNSS receiver observables – a running digital sum based simple jammer detector. *Radio Engineering*, 23(3), 898–906. Available at: http://www.radioeng.cz/fulltexts/2014/14_03_0898_0906.pdf
- 74 Agarwal, N., Basch, J., Beckmann, P., Bharti, P., Bloebaum, S. et al. (2002) Algorithms for GPS operation indoors and downtown. *GPS Solutions*, 6(3), 149–160.
- 75 Bhushan, N., Li, J., Malladi, D., Gilmore, R., Brenner, et al. (2014) Network densification: the dominant theme for wireless evolution into 5G. *IEEE Communications Magazine*, 52(2), 82–89.
- 76 Fei, N., Zhuang, Y., Gu, J., Cao, J. and L. Yang, L. (2015) Privacy-preserving relative location based services for mobile users. *China Communications*, 12(5), 152–161.
- 77 National Institute of Standards and Technology (2008) *The Keyed-Hash Message Authentication Code (HMAC)*. FIPS PUB 198-1. Available at: http://csrc.nist.gov/publications/fips/fips198-1/FIPS-198-1_final.pdf
- 78 National Institute of Standards and Technology, Digital Signature Standard (DSS) (2013) FIPS PUB 186–4. Available at: <http://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.186-4.pdf>
- 79 National Institute of Standards and Technology, Secure Hash Standard (2015) FIPS PUB 180–4 Available at: <http://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf>
- 80 National Institute of Standards and Technology. Advanced Encryption Standard (AES) (2001) FIPS PUB 197. Available at: <http://csrc.nist.gov/publications/fips/fips197/fips-197.pdf>
- 81 NGMN Alliance (2015) 5G White Paper. Available at: <https://www.ngmn.org/5g-white-paper/5g-white-paper.html>
- 82 NSF search tool. Available at: <http://www.nsf.gov/awardsearch/>
- 83 Garcia-Morchon, O., Keoh, S., Kumar, S., Hummen, R. and Struik, R. (2013) Security considerations in the IP-based Internet of Things. *IETF Internet-Draft*.
- 84 Kerr, O.S. (2007) Four models of fourth amendment protection. *Stanford Law Review*, 60(2), 503–551.

- 85 Simeone, O., Maeder, A., Peng, M., Sahin, O. and Yu, W. (2016) Cloud radio access network: virtualizing wireless access for dense heterogeneous systems. *Journal of Communications and Networks*, 18(2), 135–149.
- 86 OECD Privacy Framework, Part I (2013) Available at: http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf (accessed 10 November 2016).
- 87 Clark, P. (2016) Toto, we're not in Satnav anymore: does the Law protect Mobile Users from a misuse of their Location Data?, TaylorWessing TechFocus entry. Available at: https://united-kingdom.taylorwessing.com/download/article_satnav.html
- 88 Closas, P. and Vilà-Valls, J. (2016) NLOS mitigation in TOA-based indoor localization by nonlinear filtering under skew t-distributed measurement noise. *IEEE Statistical Signal Processing Workshop (SSP)*, Palma de Mallorca, pp. 1–5
- 89 Davidson, P. and Piche, R. (2016) A survey of selected indoor positioning methods for smartphones. *IEEE Communications Surveys & Tutorials*, 99, 1–1.
- 90 Druschel, P., Backes, M. and Tirtea, R. (2011) The right to be forgotten – between expectations and practice. *European Network and Information Security Agency (ENISA)* deliverable, 8 October 2011.
- 91 Kela, P., Costa, M., Turkka, J., Koivisto, M., Werner, J. et al. (2016) Location based beamforming in 5G ultra-dense networks. *Proceedings of the IEEE Vehicular Technology Conference (VTC 2016 Fall)*, Montreal, Canada.
- 92 Paillier, P. (1999) Public-key cryptosystems based on composite degree residuosity classes. *Advances in Cryptology – EUROCRYPT LNCS*, 1592, 223–238.
- 93 Walker, P., Rijmen, V., Fernández-Hernández, I., Bogaardt, L., Seco-Granados, et al. (2015) Galileo open service authentication: a complete service design and provision analysis. *Proceedings of the 28th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2015)*, pp. 3383–3396.
- 94 Cramer, R., Damgård, I.B. and Nielsen, J.B. *Secure Multiparty Computation and Secret Sharing*. Cambridge University Press, Cambridge.
- 95 Ferraro, R. and Aktihanoglu, M. (2011) *Location Aware Applications*. Manning Publications, Connecticut, USA.
- 96 Gellert, R. and Gutwirth, S. (2013) The legal construction of privacy and data protection. *Computer Law & Security Review*, 29(5), 522–530.
- 97 Harle, R. (2013) A survey of indoor inertial positioning systems for pedestrians. *IEEE Communications Surveys & Tutorials*, 15(3), 1281–1293.
- 98 Piqueiras Jover, R. (2016) LTE security, protocol exploits and location tracking experimentation with low-cost software radio. ArXiv preprint, arXiv:1607.05171. Available at: <https://arxiv.org/abs/1607.05171>
- 99 Piqueiras Jover, R. (2015) Security and impact of the IoT on LTE mobile networks. In: F. Hu (ed.), *Security and Privacy in the Internet of Things (IoT): Models, Algorithms, and Implementations*. Taylor & Francis LLC, CRC Press, Florida.
- 100 Minch, R.P. (2015) Location privacy in the era of the internet of things and big data analytics. *Proceedings of the 48th Hawaii International Conference on System Sciences*, pp. 1521–1530.
- 101 Brands, S. and Chaum, D. (1994) Distance-bounding protocols. *Advances in Cryptology*. EUROCRYPT, LNCS, 765: 344–359.
- 102 Capkun, S. and Hubaux, J.P. (2006) Secure positioning in wireless networks. *IEEE Journal on Selected Areas in Communications*, 24(2), 221–232.

- 103 Plosz, S., Farshad, A., Tauber, M., Lesjak, C., Ruprechter, T. and Pereira, N. (2014) Security vulnerabilities and risks in industrial usage of wireless communication. *Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA)*, Barcelona, pp. 1–8.
- 104 Sun, S., MacCartney, G.R. and Rappaport, T.S. (2016) Millimeter-wave distance-dependent large-scale propagation measurements and path loss models for outdoor and indoor 5G systems. *Proceedings of the 10th European Conference on Antennas and Propagation (EuCAP)*, Davos, pp. 1–5.
- 105 Pullen, S. and Gao, G.X. (2012) GNSS jamming in the name of privacy. *Inside GNSS*, 7(2). Available at: <http://www.insidegnss.com/node/2976>
- 106 Shaik, A., Borgaonkar, R., Asokan, N., Niemi, V. and Seifert, J-P. (2016) Practical attacks against privacy and availability in 4G/LTE mobile communication systems. *NDSS'16*, February 2016.
- 107 SWD 306 final: Commission Staff Working Document (2016) 5G Global Developments, pp. 3–4.
- 108 Symantec (2016) *Internet Security Threat Report*, vol. 21. April 2016.
- 109 Kivimäki, T., Vuorela, T., Peltola, P. and Vanhala, J. (2014) Review on device-free passive indoor positioning methods. *International Journal of Smart Home*, 8(1), 71–94. Available at: <http://dx.doi.org/10.14257/ijsh.2014.8.1.09>
- 110 Rappaport, T.S., Heath, RW. Jr, Daniels, R.C. and Murdock, J.N. (2015) *Millimeter Wave Wireless Communications*. Pearson/Prentice Hall, New Jersey.
- 111 Hengartner, U. and Steenkiste, P. (2003) Protecting access to people location information. *Proceedings of the First International Conference on Security in Pervasive Computing*, Boppard.
- 112 Lucas-Sabola, V., Seco-Granados, G., López-Salcedo, J.A. García-Molina, J.A. and Crisci, M. (2016) Cloud GNSS receivers: new advanced applications made possible. *Proceedings of the International Conference on Localization and GNSS (ICL-GNSS)*, Barcelona, pp. 1–6.
- 113 Renaudin, V., Dommes, A. and Guilbot, M. (2016) Engineering, human, and legal challenges of navigation systems for personal mobility. *IEEE Transactions on Intelligent Transportation Systems*, 99, 1–15.
- 114 Diffie, W. and Hellman, M.E. (1976) New directions in cryptography. *IEEE Transactions on Information Theory*, 22(6), 644–654.
- 115 Meng, W., Xiao, W., Ni, W. and Xie, L. (2011) Secure and robust Wi-Fi fingerprinting indoor localization. *Proceedings of the 2011 International Conference on Indoor Positioning and Indoor Navigation*, Guimaraes, pp. 1–7.
- 116 Stutzman, W.L. and Thiele, G.A. (2012) *Antenna Theory and Design*. John Wiley & Sons, West Sussex, UK.
- 117 Working Party, WP 223 (2014) Opinion 8/2014 on the Recent Developments on the Internet of Things (adopted on 16 September 2014), p. 10. Available at: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp223_en.pdf
- 118 Kim, Y., Shin, H., Chon, Y. and Cha, H. (2015) Crowdsensing-based Wi-Fi radio map management using a lightweight site survey. *Computer Communications*, 60, 86–96.
- 119 Shu, Y., Bo, C., Shen, G., Zhao, C., Li, L. and Zhao, F. (2015) Magicol: indoor localization using pervasive magnetic field and opportunistic WiFi sensing. *IEEE Journal on Selected Areas in Communications*, 33(7), 1443–1457.

- 120 Wu, Y., Hirakawa, S., Reimers, U.H. and Whitaker, J. (2006) Overview of digital television development worldwide. *Proceedings of the IEEE*, 94(1), 8–21. Available at: <http://www.tijbc.com/pruebas-7419/I0782902.pdf>
- 121 Zhou, Y.B. and Feng, D.G. (2005) Side-channel attacks: ten years after its publication and the impacts on cryptographic module security testing. *IACR ePrint Archive*, Report 2005/388. IACR.
- 122 Jin, Z., Bu, Y., Liu, J., Wang, X. and An, N. (2015) Development of indoor localization system for elderly care based on device-free passive method. *Proceedings of the Sixth International Conference on Intelligent Systems Design and Engineering Applications (ISDEA)*, Guiyang, pp. 328–331.
- 123 Ma, Z., Kargl, F. and Weber, M. (2009) A location privacy metric for V2X communication systems. *Proceedings of the IEEE Sarnoff Symposium, SARNOFF'09*. Princeton, NJ, pp. 1–6.
- 124 COM (2017) 10 Final: Proposal for a Regulation of the European Parliament and of the Council concerning the respect for private life and the protection of personal data in electronic communications and repealing directive 2002/58/EC (Regulation on Privacy and Electronic Communications). art 4, art 6, art 8, art 10, preamble 6, 11, 17, 20–24.
- 125 <https://ec.europa.eu/digital-single-market/en/proposal-eprivacy-regulation>

Part IV

5G Cloud and Virtual Network Security

14

Mobile Virtual Network Operators (MVNO) Security

Mehrnoosh Monshizadeh^{1,2} and Vikramajeet Khatri¹

¹ Nokia Bell Labs, Finland

² Aalto University, Finland

14.1 Introduction

Due to the vital role of mobile operators in providing Internet services and the fast growth of cloud computing technology, mobile operators have considered reforming themselves as one of the cloud providers for networking services. Telecommunications Service Providers (TSPs), especially Mobile Network Operators (MNOs), have invested huge amounts of resources in maintenance and expansion of their infrastructures, while cloud providers such Amazon and Google sell their services at the expense of telecom operators. A physical mobile network can host several network operators, called Mobile Virtual Network Operators (MVNOs). Each MVNO can have its own support systems or they may become customers of a Mobile Virtual Network Enabler (MVNE). In order to meet the growing demands of data and reduce the costs, an efficient development method is to implement MVNE services using cloud computing.

Since November 2012, the European Telecommunications Standards Institute (ETSI) has hosted industry specification group for Network Function Virtualization (NFV). The idea is to apply mainstream IT virtualization technologies to specify standardized network elements, which can be run on a cloud service and can be used as building blocks to create communication services [1]. This would enable cloud-based centralized network elements, which are logically separate per each MVNO, but can share software and still use operator specific data.

Both MNVE and NFV can be considered to run mostly on cloud layers [2]. We can also call this an applied platform for telecommunications, which provides Telecommunication network as a Service (TaaS). TaaS is a platform for sharing physical and virtual resources of cloud infrastructure among multiple MVNOs. TaaS is composed of software, hardware and application functions, which are also called Virtualized Network Functions (vNFs). In TaaS, each mobile operator has interconnection with cloud layers Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS), depending on the type of services it provides to customers. Based on standard security consideration, the proposed platform can be implemented with a combination of cloud deployment model.

Obviously, the mobile operators can benefit from cloudification by sharing physical resources. The new technology can also make it easier for new companies to enter not only the telecommunications service provider market place but also as software vendors for virtualized network functions and perhaps also as cloud service providers. On the other hand, with the help of network data intelligence, the operators can share the critical information (i.e. customer segmentation) with third parties (with privacy preservation) to extend their business in order to gain additional profits.

However, cloudification of mobile operators introduces several advantages, but security is still one of the largest challenges. It is assumed that 5G will follow the Software Defined Networking (SDN) principles of separating the control and data planes as well as using NFV for running network (control) functions on the cloud infrastructure, which brings additional security challenges and increases attack surface.

Due to resource sharing, various internal or external cyber-attacks (data leakage, data corruption, etc.) can target MVNOs. Although traditional prevention mechanisms such as backup-recovery, encryption, Intrusion Detection System (IDS), Internet Protocol Security (IPSec) and secure protocols can be used, we still need to confront new security challenges in a 5G cloudified network and therefore implement TaaS [3].

In this chapter, we introduce the concept of cloudification of mobile operators and present a new platform called TaaS in a cloud environment. In addition, we discuss the security challenges of TaaS, the new threats it introduces, and also the prevention mechanisms to resist these threats. Later, we discuss TaaS deployment and propose a framework to mitigate deployment security, and based on the cloud deployment model, we propose a security framework, Cloud Security Framework for Operators (CSFO), in order to achieve TaaS security. The remainder of this chapter discusses new threats introduced by NFV and mechanisms to prevent them. We also investigate Open Platform for NFV (OPNFV) security group work and see how many security requirements for TaaS have been covered by them. NFV security challenges, data and application layer threats, as well as their mitigation mechanisms, will also be investigated and NFV security requirements will be mapped to the TaaS platform. Furthermore, we discuss OPNFV security group activities and finally the conclusion is presented.

14.2 Related Work

Although some articles investigate security requirements of cloud as general, only a few cover security requirements at each layer of the cloud environment for MVNOs. However, our intent here is not to provide a complete survey of the previous studies, but to select some related works that address the security challenges of a cloudified MVNO.

The Alcatel-Lucent white paper [4] recommends security mechanisms, including hypervisor introspection and centralized security management for NFV deployment. Depending on the deployment model, identity and access management, security zones, FireWalls (FWs), hypervisor introspection and hardening, must be applied to prevent unauthorized access. Regarding Denial of Service (DoS) attacks, virtual load balancers and virtual Domain Name System (DNS) servers should be utilized. A secure key storage should be provided using specialized Hardware Security Models (HSM), so it is not

accessible and visible to third-party vNFs. Various security aspects for NFV are discussed in this chapter, but security requirements for each layer in the cloud are not discussed. Tsai *et al.* [5] studied virtualization security issues and their impact on different cloud layers. Their study includes Virtual Machine (VM) hopping, VM mobility, VM diversity and VM denial of service.

With VM hopping, an attacker can gain access from one VM to other VMs. VM mobility emphasizes spread of vulnerable configuration. However, security management across diverse domains is a challenge, but Service Level Agreements (SLAs) can help. Malicious VMs may cause denial of service, but this can be mitigated by proper policy enforcement on resource usage. Lin *et al.* [6] proposed an extended SDN architecture for NFV with a case study on intrusion prevention. The architecture reduced traffic overhead towards the controller. For this purpose, they redirected traffic to IPS vNF using service chaining. Their research did not mention any specific threat for NFV. Jang *et al.* [7] surveyed common interfaces for NFV-based security services. For access networks, security applications included traffic inspection, traffic manipulation and traffic impersonation. Required functions for these applications are Deep Packet Inspection (DPI), IPS, firewall, Virtual Private Network (VPN) and honeypots. In the mobile network environment, security applications include security configuration, security function negotiation and security request from a user device. For network security function, this chapter concludes with the use of common interfaces, regardless of where they are located and which operator they belong to in a cloud environment. Last but not least, ETSI NFV has a working group focusing on security problems. They have categorized security issues to host security, infrastructure security, vNF/tenant security, trust management and regulatory concerns. Their work covers many if not all aspects of the domain [8].

14.3 Cloudification of the Network Operators

According to the National Institute of Standard and Technology (NIST) [9], cloud computing is a process to enable on-demand access to a shared pool of configurable resources such as storage, applications and services, which can be rapidly provisioned and released with minimum provider interaction. On-demand service, broad network access, rapid provision, resource pooling and measured services are the main characteristics of this process.

The shift to cloud computing technology introduces diverse delivery models to telecom operators. In this transition, mobile operators can act as cloud network providers and based on common characteristics such as geographical zone and availability, offer networking services either to end users or other operators. For this purpose, we introduce a new functionality called TaaS. TaaS is a platform for creating functionalities to be used for commercial MNO business. TaaS is composed of software, hardware and application functions (also known as vNFs), as outlined by NFV industry standards. The combination of these functions is proposed, as TaaS and could be sold as a service product to emerging MNOs or MVNOs. The TaaS platform hosts various mobile operators. Each mobile operator has interconnection with cloud layers (IaaS, PaaS and SaaS), depending upon on the type of services it provides to the customers. Figure 14.1 shows the TaaS stacks in the cloud layers.

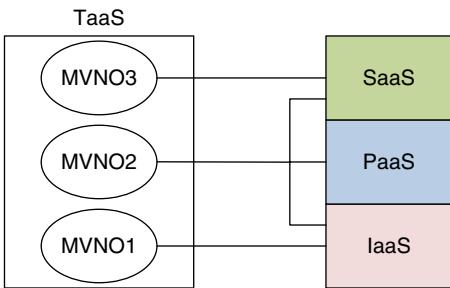


Figure 14.1 TaaS stack.

- *Infrastructure as a Service (IaaS)*: provides virtualized infrastructures. Mobile operators can rent out their network elements, storage resources, computing system and licenses to other operators;
- *Platform as a Service (PaaS)*: an interface between applications in SaaS and VMs in IaaS. PaaS controls VMs. This virtual platform is provided to developers for programming and web management. This programming can be related to network optimization, adding new features and so on. The main added value for PaaS comes from providing an easy-to-use mechanism to deploy customer's software applications to the cloud service and providing scaling for server capacity;
- *Software as a Service (SaaS)*: an application layer that provides different kinds of application software services to mobile operators, when they are relying on cloud base services. The applications can be used for bandwidth control, Quality of Service (QoS) management, network configuration, system backup and so on.

The proposed software stack can be implemented as a combination of the cloud deployment model: public cloud, private cloud, community cloud and hybrid cloud. The combination is based on security consideration and will be discussed in the next section.

14.4 MVNO Security

Security is the main issue of cloud services provided for mobile operators. Due to cloud characteristics such as virtualization and multi-tenancy, application sharing and open source software, the associated security threats such as authentication, information leakage and data corruption are also growing in the TaaS cloud environment. On the other hand, due to the open nature of IP in mobile technologies, these networks are potential targets of cyber-attackers to intrude services and cause problems to the end users and mobile operators. Although in later phases, extensions such as IPSec and Authentication Authorization and Accounting (AAA) have been added into mobile network implementations, security is still a main challenge in cloud computing, because of the inconsideration in initial design of the Internet [10–11].

Due to the nature of 5G networks in extremely fast communication, it is important how and which authentication mechanisms are chosen, since even the lowest latency can have considerable effects on communication. Moreover, in 5G networks, delay in setting up control plane security will impact the delay of sensitive applications such as machine-to-machine (M2M). Common Non-Access Stratum (NAS) signaling

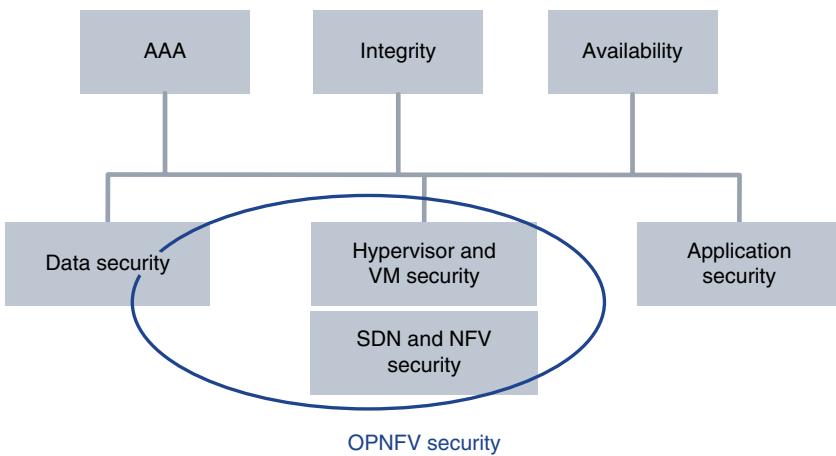


Figure 14.2 TaaS security classification.

(state transition) is an example of control plane signaling, which may introduce delays in 5G security. The user enabled security feature is another aspect to be considered in 5G networks. In current mobile networks, it is MVNO who decide which and when security mechanisms should be applied, while it would be more feasible if the user could enable the security feature voluntarily, depending on the needs of the application used [12].

In general, it is possible to build 5G security based on 4G security mechanisms considering their significant robustness. On the other hand, we cannot yet introduce a certain security model for 5G, since its architecture has not been finalized. Therefore, in this chapter, we concentrate on MVNO security for the cloud environment, regardless of the generation (4G or 5G).

To fulfill security requirements such as availability, integrity and AAA, we need to address TaaS vulnerabilities. We assume that these requirements will be applied in some way also on 5G networks. In Figure 14.2, TaaS security issues are classified into three main aspects:

- 1) Data Security;
- 2) Hypervisor and VM security, which covers SDN and NFV security; and
- 3) Application Security.

14.4.1 Data Security in TaaS

Mobile operators that act as cloud providers are responsible for their customer data protection. The customers are either end users or other mobile operators. Herein we define data vulnerabilities based on the security requirements:

- 1) *Authentication, authorization and accounting*: refers to the access control mechanism that is used against unauthorized access or privilege logging. In this scenario, an attacker tries to modify, corrupt, steal and intercept the data of Control Plane (CP) and User Plane (UP). In addition to account control, telecom operators that act as a cloud provider should also consider other aspects of data protection;

2) *Integrity*: points out data correctness as to whether the data has been modified or corrupted. Malicious codes can be distributed by both insider and co-tenant or via external attackers on data storages [13]. Data encryption, data isolation, secure protocols and intrusion detection can support data integrity and prevent data modification and corruption.

Another aspect of data integrity is data leakage prevention that can be achieved by data sanitization [11,14–15] (encryption and data cleanup). Since multiple tenants may share the same infrastructure or VM; e.g. virtual Home Subscriber Server (vHSS), the cloud service provider is responsible for a complete data cleanup before handing over VM to the next tenant;

3) *Availability*: covers the basic concepts of security, such as data recovery and resource availability. Availability can be achieved via load balancing, redundancy and data backup to prevent data loss. Threats such as DoS should be prevented by an intrusion detection mechanism.

In addition to the security requirements already discussed, legal aspects such as security warranties and compensation agreements among operators belonging to TaaS look necessary. On the other hand, location of the cloud provider [16] (where the parent company is registered) is important, since different countries have diverse laws; regardless of data centers location, in special circumstances, authorities will have access to customer data.

14.4.2 Hypervisor and VM Security in TaaS

The concept of virtualized threats refers to every kind of attack against availability, integrity and confidentiality of the hardware and software in a virtualized mobile network. There are three elements in a virtualized network: hypervisor, VMs (virtual hardware and images), and applications; all these elements should be adequately secured against unauthorized access, change and destruction.

In a virtualized mobile network, the hypervisor itself is not directly connected to any end user, and most threats arise through malicious VMs, therefore having a reliable hypervisor requires secure VMs. While traditional security techniques such as IDS, antivirus and FWs are still applicable for virtualized networks, isolation could be an important approach towards security of VMs. Isolation will ensure that if one VM is attacked, other VMs are not infected [17,18].

There are different methods such as security zones and traffic separation for VM isolation. VMs with similar functionality and security requirements could be grouped in same hardware. Each zone could be controlled by a different access list defined in FW or dedicated IDS and so on. DeMilitarized Zone (DMZ) is an example of a security zone. Traffic separation is another method for VM isolation; similar to traditional networks, traffic with different characteristics, functionality (e.g. CP and charging) and security requirements would be assigned to different Virtual Local Area Networks (VLANs) or VPNs, in this case sensitive traffic would be separated [19].

After discussing virtualized network security concerns, we now go through security requirements such as authentication, availability and integrity:

- 1) *Authentication, authorization and accounting:* refers to a mechanism such as certificate-based authentication that should be utilized to avoid unauthorized access. Keys and signatures must be stored in a secure storage such as HSM [4] to make it invisible to third parties. Hardening should be applied to the infrastructure layer and wherever needed to block any backdoor access. Virtual FWs must be used inside VMs and proxy and traditional FWs must be used where needed to prevent unauthorized traffic. Backup must be maintained for all VMs, so that data can be restored in case of failure. AAA should be maintained by logging actions from each VM and modules; and logs should be stored in a safe storage so that in the case of attack or failure, the logs will not be affected and could help to reveal the root cause. Encryption must be used so that data is not readable to unintended parties, even if it is accessed without authorization. Security policy should be enforced to make sure that all users in the cloud have similar security policy and are in line with SLA [20].
- 2) *Integrity:* refers to protection against threats in the virtualized network. These threats could vary from attacking different virtual machines such as virtual Mobility Management Entity (vMME), vHSS, and their virtual functions, misconfigurations or abuse of resources, corrupting operating systems, switches and management software (in SDN), and inducing any kind of malicious applications.
- 3) *Availability:* can be improved by applying techniques such as load balancing, redundancy and data backup, as discussed earlier.

14.4.2.1 SDN Security in TaaS

In addition to the three main aspects of TaaS security (data security, hypervisor VM security and application security), here we discuss other security and threats of cloud computing, therefore TaaS.

It should be considered that introducing new technologies normally also brings new security challenges. Except for the security issues of SDN that are new technologies for supporting TaaS, other security threats and their detection-prevention mechanisms are almost similar to traditional networks. However, traditional network implementations rely on dedicated hardware and private connections between network elements. Their control plane connections are not exposed to the public, unless there is a configuration error somewhere. The traditional network has survived quite well until today, and will continue to survive with very little security awareness.

SDN is a new approach to separate UP and CP in mobile networks. As shown in Figure 14.3, based on its functionality, SDN can be considered in IaaS (SDN switch) or in PaaS (SDN controller). The SDN controller in mobile networks only carries CP (MME, or Serving/Public Data Network Gateway (S/P GW) VMs), and is located in PaaS. In this case, SDN is an interface between the infrastructure and application layer. SDN in its switching functionality (S/P GW VM) only carries UP and considers part of infrastructure layer. UP and CP use OpenFlow protocol for communications with each other in SDN. The vNFs are running on virtual machines and they provide a certain subsection of the whole functionality of a telecommunications network.

However, from the security point of view, SDN brings several advantages, such as [21]:

- *Centralized management:* simpler maintenance and debugging;
- *Programmability:* faster security solution implementation and easy feature deployment;
- *Cost saving:* with sharing security techniques and service chaining; and
- *Centralized and virtualized function:* centralized monitoring and detection techniques.

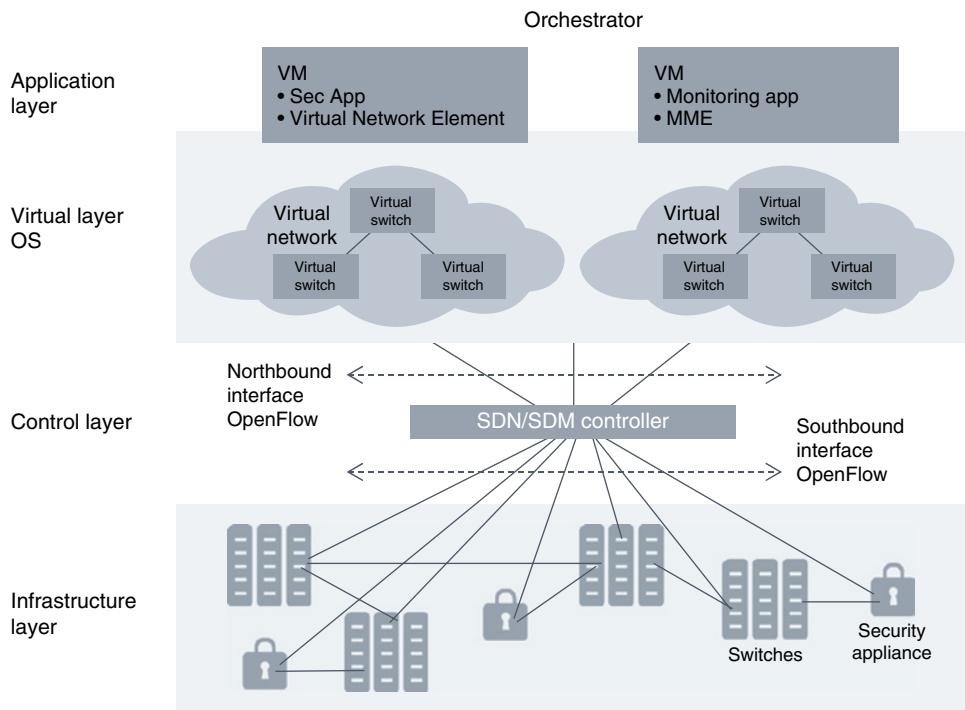


Figure 14.3 Layer-based SDN architecture.

They also have disadvantages for mobile networks and therefore for TaaS:

- *Centralized controller*: potential for single attack;
- *Vulnerable southbound interface (OpenFlow) between controller and data-forwarding*: degrade network, availability, performance and integrity via DoS attack;
- *Vulnerable northbound interface*: between controller and applications;
- *Programmability*: applications have access to controller to program the network;
- Reduced isolation of network functions; and
- Expensive and vulnerable cryptographic keys.

SDN carries most of the three-layer threats such as configuration, authorization and access control, as well as software and images vulnerabilities. The Open Networking Foundation (ONF) has identified the southbound interface between controllers and data forwarding devices (SDN switches) as vulnerable. This interface uses OpenFlow protocol, which could be vulnerable against spoofing if the authentication between controllers and switches are not implemented correctly or is compromised [21]. Therefore, communication between controller and switch must be over the Transport Layer Security (TLS) or IPSec, to avoid eavesdropping, tampering and DoS attacks at the controller. Considering SDN, secure Application Programming Interface (API) techniques must be utilized at the northbound interface.

Below are listed some of the monitoring, detection and prevention techniques that could be used for SDN, NFV and OpenFlow security [21]:

- Deep Packet Inspection (DPI);
- Deep Flow Inspection (DFI);
- Shallow Packet Inspection (SPI);
- Virus scanners;
- Intrusion Detection Systems (IDS);
- Firewall (FW);
- Security zones;
- Policy enforcement (PCRF) to define access rules and flow rules for access control and authorization;
- Secure protocols, i.e. FlowTagging (flow tracking);
- Simple Network Monitoring Protocol (SNMP);
- Remote Monitoring (RMON);
- NetFlow or sFlow; and
- SDN Monitoring (SDNM).

14.4.2.2 NFV Security in TaaS

Virtualization is the main component in cloud services provided by mobile operators. Multi-tenancy, application sharing and open source software lead to security threats such authentication, information leakage and data corruption in cloud environments, including TaaS.

Usually, open source software may contain vulnerabilities, bugs and other security holes and therefore not in line with enterprise security requirements. In a poorly secured open-source environment, attackers can easily have access to the system [22]. Some common vulnerabilities in open source include Heartbleed and ShellShock. Heartbleed is a bug in the OpenSSL software library that allows theft of protected information. This bug has infected many web and email services [23–24]. ShellShock is a vulnerability in bash that allows the non-authorized user (hacker) to remotely execute commands and take over the system [25–26]. These vulnerabilities were discovered and then correction patches were applied, but if vulnerabilities are not detected early enough, open-source software brings security challenges. Considering similar cases, open-source software adds more security concerns to the cloudified environment, and it should be carefully evaluated and tested before utilizing it.

NFV in TaaS refers to any network function that runs on mobile network equipment over a hypervisor. There are three attack profiles in NFV [27]:

- 1) *Intra-MVNO attacks*: include attacks on an MVNO by its own employee to occupy and degrade network services;
- 2) *Inter-MVNO attack*: refers to any type of attack from one MVNO towards another MVNO(s), in order to extract the competitor's information, corrupt or misuse their services;
- 3) *Attacks by end user*: this category covers the attacks that are caused by mobile network end users within the same MVNO or other MVNOs.

In a cloud environment with NFV, network functions will be deployed as vNFs that bring security challenges. Different solutions, such as security zone and grouping, isolating applications by VMs and licensing are recommended for NFV security. NFV acts in the hypervisor and other parties can see the encryption keys, therefore providing a signature beside the keys [4]. FW and orchestration both are recommendations for

NFV and platform security. Some of the major threats on vNFs and their mitigation mechanisms are explained here [28–30]:

- 1) *Malicious loops that are caused by routing loops, unavailability of management network due to network failure*: to prevent these threats, the network should be logically validated to make sure that management interfaces are accessible; even if vNFs are down;
- 2) *Improper data removal due to VM crash, execution of malicious vNF and therefore unauthorized changes to Basic Input/Output System (BIOS) or Unified Extensible Firmware Interface (UEFI), hypervisor and Operating System (OS)*: for mitigation secure boot, i.e. Trusted Platform Module (TPM) and crash protection can be used;
- 3) *Abuse of hypervisor resources by malicious VM (impacting other VMs) and QoS degradation*: performance isolation by segregating resources to each VM is recommended as a prevention mechanism [31];
- 4) *Insufficient vertical and horizontal VM AAA mechanism*: to prevent this threat, AAA mechanisms among vNFs, between vNFs and the application layer and between vNFs and management stations, should be revised;
- 5) Software unreliability such as:
 - *Coding flaws*: that affect all MVNOs using the same software: Correction and security patches should be applied on all VMs using the same software;
 - *Configuration changes or correction patches*: that need reboot and cause service outage on MVNOs; backup and load balancing are the mitigation mechanisms for such a threat;
 - *Test and monitoring backdoors*: closing test and monitoring and debug interfaces are recommended;
 - *Stored password and private keys in VM images*: using unique private key for each image could prevent these threats.

14.4.2.3 OPNFV Security

OPNFV is an open-source platform used for vNF deployment and Proof of Concept (PoC). While the OPNFV security group improves security of OPNFV via code review, vulnerability management and documentation, in general, their focus is on network virtualization, SDN controller framework, OpenStack and virtual storage [32–34]. At a research level, there is not much on OPNFV security, but they are following the ETSI security group activities and will eventually develop the upstream, and improve the audits and security guide, which covers how to secure an OPNFV-based deployment.

For AAA, most of the threats are already mitigated in various OpenStack projects. In the Virtual Infrastructure Manager (VIM) or SDN controller, AAA is inherent in key-stone Open DayLight (ODL). There are also some blueprints being worked on for the Federation and use of Security Assertion Markup Language (SAML), OpenID connect, etc. However, these considerations are not applied if we intend to go a layer above the VIM.

In data security, OPNFV only covers transport of data; for example, various service APIs. The data at rest or in motion of the application is not covered.

For hypervisor, SDN and NFV security, OPNFV only concentrates on OpenStack, while some security issues such as execution of malicious and non-verified vNFs have

Table 14.1 OPNFV security focus.

| | Research domains | | |
|----------------|------------------------------------------------|----------------------------|-------------|
| | Data | Hypervisor, SDN and NFV | Application |
| | | | |
| OPNFV Security | Virtualization (KVM, QEMU, XEN) | | ✓ |
| | Network Virtualization (DPDK, ODP, OVS) | ✓ | ✓ |
| | SDN controller framework (ONOS, ODL, OpenFlow) | ✓ | ✓ |
| | OpenStack | | ✓ |
| | Virtual storage | | ✓ |

not yet been mitigated. In documentation, the plan is to cover these and a guide is being collaboratively worked on, although most of it is covered in the OpenStack security guide, even Kernel-based Virtual Machine (KVM), etc. Secure boot, trusted compute, etc., are covered in the OPNFV security guide, although it is still very much a work in progress. Outside of that, it comes down to vendor implementation as to how they configure the TPM to be harnessed by the host OS.

The application security domain has not been considered, since OPNFV concentrates only on NFV Infrastructure (NFVI), VIM, and MANagement and Orchestration (MANO).

Since OPNFV is about upstream code contributions, several parts are covered by upstream projects. Therefore, it is not easy to make a direct comparison at a functional level. In Table 14.1, five aspects of OPNFV security (virtualization, network virtualization, SDN controller framework, OpenStack and virtual storage) are compared with TaaS security domains (data, hypervisor SDN/NFV and application).

In Table 14.1, none of the five aspects of OPNFV security covers application security, only two aspects (network virtualization and SDN controller frame work) cover data security and all five aspects cover hypervisor security. The comparison shows OPNFV security does not cover application security; however, it partially covers data security and considerably covers hypervisor, SDN and NFV security. Therefore, OPNFV security needs to be revised in areas of data and the application domain, especially to meet the security requirements for TaaS [27].

14.4.3 Application Security in TaaS

Another aspect of virtualized network security refers to protection against threats that are related to an application server or a web server connected to the Internet.

Based on the concept of SaaS, software applications should be accessible over the Internet that makes security a very critical challenge for mobile operators. Beside the mechanisms such as data encryption, access control and authentication, back up and redundancy, the mobile operator could implement sensitive applications that do not require end user intervention (i.e. billing application), by using PaaS that is accessible only to limited professional users among mobile operators [35].

14.4.4 Summary

In Figures 14.4 to 14.7, both traditional and cloud specified threats and their mitigation mechanisms for three domains of data, hypervisor (VM, SDN and NFV) and application are classified in different categories (AAA, integrity and availability).

Although some attacks are common to traditional networks, therefore similar mitigation mechanisms are used [36]. Still attacks that are targeting hypervisor, SDN and NFV domains are cloud specific and therefore their mechanisms are different from traditional networks [37]. In addition to hypervisor specified threats, a new mitigation mechanism is introduced in the application domain for cloudified network. For this purpose, system-related applications are implemented in PaaS rather than SaaS, since these applications should be accessible by a limited group of developers and not by the end users.

| █ Threats | █ Mitigation | |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------|
| Data | Hypervisor, SDN and NFV | Application |
| <ul style="list-style-type: none"> • Probing • Man-in-the-middle attack • User to remote • Remote to local • IP spoofing • Phishing | <ul style="list-style-type: none"> • Side channel attacks • Stored password and private keys in VM image • Back doors, test and monitoring interfaces | <ul style="list-style-type: none"> • Spyware • Cookie poisoning • Service injection attack |
| <ul style="list-style-type: none"> • AAA • Firewall • Rule-based policy control • Secure protocols <ul style="list-style-type: none"> • Encryption and Hardening • Data cleanup before switching tenant | <ul style="list-style-type: none"> • Hypervisor monitoring • VM isolation • Unique private keys for each VM • Patching and closing test and monitoring interfaces | <ul style="list-style-type: none"> • Encrypting cookie data • Software updates and security patches |

Figure 14.4 AAA requirements for cloudified network.

| Data | Hypervisor, SDN and NFV | Application |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"> • DoS • Unexpected system failure • Data removal | <ul style="list-style-type: none"> • Image loss • Configuration loss • Misconfiguration • Availability of management station | |
| <ul style="list-style-type: none"> • Redundancy • Backup • IDS • Firewall <ul style="list-style-type: none"> • Load balancing and resource isolation (NW, CPU, memory) | <ul style="list-style-type: none"> • Configuration test • Disabling not used test and debug interfaces • Logical network validation for management stations | <ul style="list-style-type: none"> • Implementing system-related applications in PaaS |

Figure 14.5 Availability requirements for cloudified network.

| Data | Hypervisor, SDN and NFV | Application |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"> • Corruption and tampering (botnet, malware, ransomware...) | | |
| <ul style="list-style-type: none"> • IDS • Vulnerability and virus scanning • Firewall • Encryption • Validation and error checking mechanisms | <ul style="list-style-type: none"> • Software updates and security patches • Secure coding | <ul style="list-style-type: none"> • VM isolation (security zone and traffic separation) • Secure boot – Trusted platform module (TPM) • Secure crash |
| | | <ul style="list-style-type: none"> • Secure browser |

Figure 14.6 Integrity requirements for cloudified network.

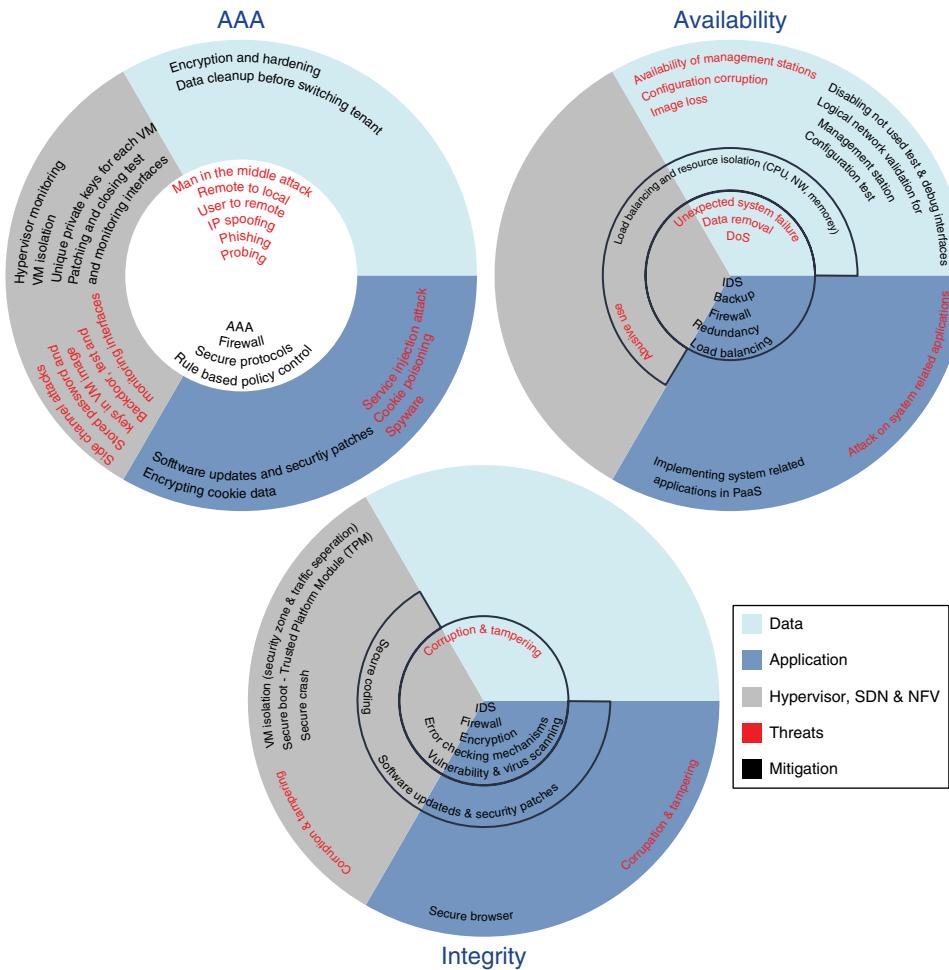


Figure 14.7 Security requirements for cloudified network.

Table 14.2 TaaS security benchmark.

| Domain | Requirements | | | | | | Affected Layer | |
|-------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------|--|
| | AAA | | Availability | | Integrity | | | |
| | Threats | Prevention | Threats | Prevention | Threats | Prevention | | |
| Data | Unauthorized access and privileged access <ul style="list-style-type: none"> ● Probing ● Remote to local ● User to remote ● Man-in-the middle ● IP Spoofing ● Phishing | <ul style="list-style-type: none"> ● AAA ● FW ● Rule-based policy control ● Encryption ● Hardening ● IPSec | Data loss and resources unavailability <ul style="list-style-type: none"> ● Data removal ● Unexpected system failure ● Abusive use ● DoS | <ul style="list-style-type: none"> ● Backup ● Redundancy ● Load balancing ● IDS ● FW ● AAA ● IPSec | Data corruption, tampering and leakage | <ul style="list-style-type: none"> ● FW ● AAA ● IPSec ● IDS ● Vulnerability scanning ● Encryption SSL/TLS ● Data cleanup before switching tenant | SaaS PaaS IaaS | |
| Hypervisor and VM | <ul style="list-style-type: none"> ● Probing ● Remote to local ● User to remote ● Man-in-the middle ● IP-Spoofing ● Phishing | <ul style="list-style-type: none"> ● AAA ● FW ● Rule-based policy control ● IPSec ● Hypervisor monitoring ● VM isolation | <ul style="list-style-type: none"> ● Image loss ● Configuration loss ● Misconfiguration ● DoS ● Abusive use | <ul style="list-style-type: none"> ● Backup ● Redundancy ● Load balancing ● IDS ● FW ● AAA ● IPSec ● Configuration test | Data corruption, tampering and leakage <ul style="list-style-type: none"> ● Botnet ● Malware | <ul style="list-style-type: none"> ● VM isolation ● Security zone ● Traffic separation VLAN and VPN ● SSL/TLS ● DPI ● IDS ● FW ● AAA ● IPSec | PaaS SaaS | |
| Application | <ul style="list-style-type: none"> ● Probing ● Remote to local ● User to remote ● Man-in-the middle ● IP-Spoofing ● Phishing ● Spyware ● Cookie poisoning ● Service injection | <ul style="list-style-type: none"> ● AAA ● FW ● Rule-based policy control ● IPSec ● Encrypting cookie data | <ul style="list-style-type: none"> ● Unexpected system failure ● DoS | <ul style="list-style-type: none"> ● Redundancy ● Implementing system related applications in PaaS | <ul style="list-style-type: none"> ● Application corruption ● Botnet ● Malware ● Adware ● Ransomware | <ul style="list-style-type: none"> ● IDS ● Secure coding ● Secure browser | PaaS SaaS | |

14.4.5 MVNO Security Benchmark

In order to propose a security framework for MVNOs in TaaS, an evolved benchmark on threats and their mitigation mechanisms against security requirements is needed. Therefore, in Table 14.2 and Figure 14.7, based on security requirements and for each domain of a cloudified environment, the TaaS threats and their prevention mechanisms are listed [36,38–40].

Some of the threats listed in Table 14.2 are described here:

- 1) *Probing*: is an attempt to monitor a computer or network and steal important information such as open ports and IP addresses for devices connected to a network. Examples include port scanning and IP sweep attacks;
- 2) *Remote to Local (R2L)*: this threat tries to access target machines without having an account and permissions on that machine. Access is made possible by exploiting a vulnerability and other related means. An example is the File Transfer Protocol (FTP) write attack, which exploits a common anonymous FTP misconfiguration. Other examples include dictionary attacks, Hyper-Text Transfer Protocol (HTTP) tunnel attacks and Xsnoop attacks;
- 3) *User to Root (U2R)*: is an attempt to get administrator or super privilege access while the user has only local access to a victim machine. A vulnerability in the victim machine is exploited in order to gain root access. An example is the yaga attack, which adds the attacker to the domain admins group by hacking the registry and can crash a service on the victim's machine. Other examples include ps-attack and Xterm attack. The vulnerability CVE-2016-0728 has been found in Linux kernels 3.6 and later versions, which is a reference leak in the keyrings facility and occurs when an error message is generated if a process tries to replace its current session keyring with the same one [41];
- 4) *Man-in-the middle attack*: this attack occurs when an attacker gains access to the communication channel established between two legitimate users. The attacker is capable of performing unauthorized activities such as intercepting and modifying communications including send and receive data that is meant for someone else. It is a type of eavesdropping attack that occurs when a malicious actor inserts himself as a relay/proxy into a communication session between people or systems. Examples include man-in-the middle attack on HTTP and a poorly-implemented Secure Sockets Layer (SSL);
- 5) *IP Spoofing*: in this attack, a user/device creates IP packets with a false source IP address, in order to hide the identity of sender or introduce itself as another device and steal the data;
- 6) *Phishing*: in this attack, an attacker tries to learn account information or login credentials from a user by presenting himself/herself as a reputable and genuine entity or person. The communication channels include email, phone call, instant message and others. Typically, a web link is sent to the user, which looks safe and asks for user credentials. Upon giving credentials, the credentials are forwarded to attacker;
- 7) *Spyware*: is all kinds of software that monitors computers and networks to collect unauthorized information or steal credentials, such as passwords and credit card numbers;

- 8) *Cookie poisoning attack*: when a user visits a webpage in a web browser, personal information of the user along with session information is stored in a cookie. In this attack, a cookie is modified by an attacker to gain unauthorized information about the user for identity theft purposes;
- 9) *Service injection attack*: injection attacks targeting a service belong to this category. Examples include Structured Query Language (SQL) injection attack, eXtensible Markup Language (XML) injection attack and cross-site scripting attack;
- 10) *Botnet*: a group of connected vulnerable computers in a network, which are remotely controlled by a master computer (hacker). Similar to robots, they automatically perform some functions that are predefined by the botmaster and forward information like viruses to target computers, which can cause denial of service. Botnets often use basic applications like Internet Relay Chat (IRC) and HTTP, and communications among them are encrypted that makes it difficult to detect them;
- 11) *Malware*: refers to all kinds of software codes, i.e. viruses, worms, Trojans and drive-by download. These attacks are programmed to perform malicious operations on a networked device;
- 12) *Adware*: all kinds of software that use some form of advertising delivery system to replace banner ads on web pages with those of other content providers;
- 13) *Ransomware*: any kind of software that locks a computer and demands some form of payment to make the computer unlock.

14.5 TaaS Deployment Security

What matters in cloud computing is the combination of layers and deployment to propose a new security model. Our proposed platform Cloud Security Framework for Operators (CSFO) not only recommends for each layer a proper deployment but also emphasizes on specific detection-prevention mechanism for different layers [42]. Figure 14.8 shows our proposed security framework for TaaS:

14.5.1 IaaS

- 1) *Layer point of view*: Since infrastructures are fully managed by a mobile cloud provider, the security mechanism is also responsibility of the provider. The tenants usually have minimum control and interaction on the network elements. They do not have access to the control plane VMs, even though they still could reach some of the network elements, such as Home Location Register (HLR) or the Policy Control and Charging Function (PCRF) server, to pull their subscribers' information (i.e. subscriber profile, billing information). However, IaaS is less accessible by customers (end users or tenants); still insider attackers need to be highly considered.

For this layer, techniques such as data isolation through VMs, ciphering to protect data against unauthorized access, backup and recovery for data reliability and IDS for preventing malicious attacks should be considered by the cloud provider.

- 2) *Deployment point of view*: Considering high security requirements for infrastructures, limited accessibility, geographic location and high cost of network elements, mobile operators are recommended to use a private cloud for this layer.

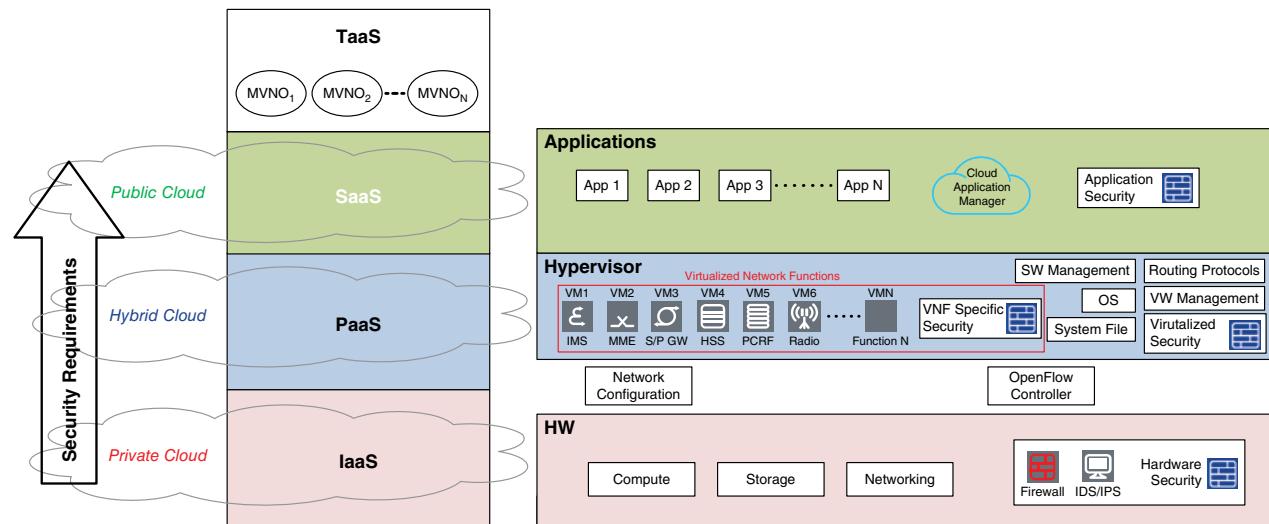


Figure 14.8 Cloud security for operators.

14.5.2 PaaS

- 1) *Layer point of view:* This layer is normally used by developers to program and run their applications. Generating software bug or file system corruption, unauthorized access or privilege upgrade and denial of service are the security threats that should be considered at this layer. Strong authentication and access right control is required for this layer to limit the user base that can make critical modification to configuration. Logging of all management actions are important to trace misbehaving users and to learn from mistakes. Therefore, a policy control mechanism could evaluate the requested access and decide whether or not to grant the access to developer.
- 2) *Deployment point of view:* This layer will be used by a limited group of professionals and does not need to be accessed by all end users, therefore community or hybrid deployment for this layer is recommended. Mobile operators could take advantage of the public cloud, while for sensitive parts of the system software, they could just provide a private cloud.

14.5.3 SaaS

- 1) *Layer point of view:* Since the application layer is the closest layer to the end users, they could easily install different kinds of malware or spyware and steal the information or cause data corruption at this layer. Secure protocols and malware detection methods are some of the prevention mechanisms that should be considered in this layer.
- 2) *Deployment point of view:* In order to gain the initial cloud computing benefits, such as elasticity and economies of scale, SaaS should be available to all customers (end users and other tenants), therefore public cloud is recommended for SaaS.

In Figure 14.8, a larger area is dedicated to the IaaS layer, to highlight the higher security requirement for this layer.

14.6 Future Directions

A cloud system is distributed over many geographically separate computing sites, and if one site breaks down unexpectedly (e.g. by earthquake or severe cyber-attack), it is challenging to leverage such a cloud system. Although there are many researches on the security concerns of cloud computing, there are still open issues requesting further investigation in future studies:

- 1) Various new business opportunities opened by cloudification disruption in the operator domain should be investigated and analyzed as to whether these business opportunities trigger further technological breakthroughs or not. On the other hand, further research is necessary to understand the requirements to TaaS and their mutual priorities. It should be studied whether there will be lightweight MVNOs, which operate almost without their own staff. To achieve this, the expectations of various stakeholders, i.e. MNOs, MVNOs, equipment vendors and end users in the security domain, should be listed.

- 2) Fighting against governmental cyber war attacks is one of the common concerns among potential target governments, privacy aware end users, and consequently also service providers. Some governments are not financially restricted while preparing cyber-attacks and espionage and they may sponsor cracker and hacktivists to leverage their technological curiosity for ideological purposes. For this purpose, a novel means is by at least detecting these attacks as early as possible, or even prevention appears necessary.
- 3) Cloud computing enables fast update cycles for software components such as VM images. It should be investigated as to whether lowering quality assurance (testing) effort of finding typical software bugs like buffer overruns is feasible or not. Lowering quality assurance poses a risk of enabling vulnerabilities that can break into a system. It is important that such breaches can be detected quickly and also their fixes are distributed before any major damage occurs. On the other hand, it is worth considering if the possible development cost savings and profits from faster time to market, exceed potential expenses caused by damages and bug fixing.
- 4) From the end user perspective, one main benefit for MNO cloudification is utilization of a higher bandwidth for lower costs. It can also provide flexibility benefits such as on-demand services and dynamic charging patterns for bandwidth fluctuations, etc.
- 5) From the legacy perspective, legislation protecting end users against MNO/MVNOs with malicious intentions may be needed in the future. Small capital lightweight MVNO may be bought by the wrong people with malicious intent, and legislation should be introduced to disconnect such malicious MVNOs.

We need to learn more about specifications and results of ongoing work at ETSI NFV. Proposed TaaS will implement at least partial security requirements outlined in the ETSI NFV specification. The TaaS concept needs to be analyzed further and compared with NFV to find commonalities and also to understand how it can be positioned in NFV context. ETSI NFV compliant open source software implementation OPNFV release Colorado became available on September 26, 2016 [31,32], and it should be analyzed as to how well it addresses security concerns outlined for TaaS. In addition, OPNFV security needs to be revised to accommodate the cloudified environment. While multiple open source projects are being released for cloud and NFV, its security aspects from the mobile operator point of view should be investigated and tuned to meet their demands in a cloudified environment.

14.7 Conclusion

Emerging traditional mobile operators who follow similar interests introduce potential demand for a new service model in cloud computing called Telecommunication network as a Service (TaaS). According to a location-based, customer-based or service-based agreement, mobile operators could be grouped to considerably improve their cost structure, time and quality efficiency and therefore their speed to market.

TaaS provides the possibility to understand mobile operator's threats in a wide range and based on their provided cloud layers. While the majority of earlier studies have concentrated only on a few threats for a specific layer, this chapter has discussed the mobile cloud threats for all layers of cloud and from an MVNO point of view towards

a secure and flexible 5G architecture. Also, it is important to define the security requirements for 5G networks at the initial stage. On the other hand, the proposed cloud security model helps mobile operators to understand how to tradeoff and merge their services based on the deployment and importance of the provided services. In Figure 14.8, private deployment is assigned to Infrastructure as a Service (IaaS) that requires highest security consideration.

For Software as a Service (SaaS), which is the closest layer to the end users, public cloud is recommended. Some of the reasons for this recommendation come from application availability to a wide range of end users, scattered end users (geographic location) and roaming condition. Finally, hybrid cloud is the proposed deployment for Platform as a Service (PaaS) layer; that means private and public deployment could be considered for provided platforms and based on their sensitivity and security concerns. The discussed CFSO model is a combined security model that considers different threats and vulnerabilities for each layer, their modules, services and protocols, and helps TaaS to find the best combination of deployment solution.

In addition, we reviewed three categories of Mobile Virtual Network Operator (MVNO) attack profiles: intra-MVNOs attacks, inter-MVNOs attacks and end-user attacks and some of the Network Function Virtualization (NFV) specific threats and their mitigation mechanisms for a cloudified MVNO were introduced. However, the majority of cloudified MVNO threats and their mitigations are similar to traditional networks, and still there are new threats introduced by virtual Network Functions (vNF). Abuse of unremoved Virtual Machines (VM) data by new tenants, vNF malicious loops, malicious VMs, insufficient AAA mechanisms or non-unique keys for VMs, are some of these new threats. Furthermore, we reviewed TaaS security domains (data, hypervisor and application) and applied three main security requirements (AAA, availability and integrity) for each domain and addressed virtualized network threats and their prevention mechanisms.

In addition, Open Platform for NFV (OPNFV) security activities has been discussed, since OPNFV has been popular in the open source community for development and Proof of Concepts (PoCs). Based on security key criterions, we showed that the OPNFV security group does not cover data security partially and application security. OPNFV uses OpenStack as a hypervisor platform and relies on its security to cover hypervisor, SDN and NFV security. Considering the security requirements outlined for TaaS, OPNFV security needs to be revised to accommodate the cloudified environment. While multiple open source projects are being released for cloud and NFV, its security aspects from the mobile operator point of view should be investigated and tuned to meet their demands in a cloudified environment.

References

- 1 Chiosi, M. and Wright, S. (2014) Network functions virtualisation – White paper # 3, ETSI, Darmstadt, Germany. Available at: https://portal.etsi.org/Portals/0/TBpages/NFV/Docs/NFV_White_Paper3.pdf
- 2 Monshizadeh, M., Yan, Z., Hippeläinen, L. and Khatri, V. (2015) Cloudification and security implications of TaaS. *Proceedings of the World Symposium on, Computer Networks and Information Security (WSCNIS)*, pp. 1–8.

- 3 Katica, N. and Tahirovic, A. (2012) Opportunities for telecom operators in cloud computing business. *Proceedings of the 35th International Convention, MIPRO*, 2012, pp. 495–500.
- 4 Why service providers need an NFV platform (white paper). Alcatel-Lucent, Tech. Rep. C401-01087-WP-201409-1-EN, 2013.
- 5 Tsai, H.Y., Siebenhaar, M., Miede, A., Huang, Y., and Steinmetz, R. (2012) Threat as a service? Virtualization's impact on cloud security. *IT Professional*, 14(1), 32–37.
- 6 Lin, Y.D., Lin, P.C., Yeh, C.H., Wang, Y.C. and Lai, Y.C. (2015) An extended SDN architecture for network function virtualization with a case study on intrusion prevention. *IEEE Network*, 29(3), 48–53.
- 7 Jang, H., Jeong, J., Kim, H. and Park, J.S. (2015) A survey on interfaces to network security functions in network virtualization. *Proceedings of the IEEE 29th International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, Gwangju, pp. 160–163.
- 8 Chiosi, M. (2012) Network functions virtualisation – introductory white paper. ETSI, Darmstadt, Germany. Available at: https://portal.etsi.org/nfv/nfv_white_paper.pdf
- 9 Mell, P. and Grance, T. (2009) The NIST definition of cloud computing. *National Institute of Standards and Technology*, 53, 50.
- 10 Oredope, A., McConnell, A., Peoples, C., Singh, R., Gonsalves, T.A. et al. (2013) Cloud services in mobile environments: the IU-ATC UK-India mobile cloud proxy function. *Wireless Conference (EW), Proceedings of the 2013 19th European*, pp. 1–7.
- 11 Fernandes, D.A., Soares, L.F., Gomes, J.V., Freire, M.M. and Inácio, P.R. (2014) Security issues in cloud environments: a survey. *International Journal of Information Security*, 13, 113–170.
- 12 Schneider, P. and Horn, G. (2015) Towards 5G security. *2015 IEEE Trustcom/BigDataSE/ISPA*, Helsinki, pp. 1165–1170.
- 13 Subashini, S. and Kavitha, V. (2011) A survey on security issues in service delivery models of cloud computing. *Journal of Network and Computer Applications*, 34, 1–11, 1.
- 14 Kronabeter, A. and Fenz, S. (2013) Cloud security and privacy in the light of the 2012 EU data protection regulation. In: *Cloud Computing, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering* (M. Yousif and L. Schubert, eds), Springer, Berlin, pp. 114–123.
- 15 Soares, L.F.B., Fernandes, D.A.B., Gomes, J.V., Freire, M.M. and Inácio, P.R.M. (2013) Cloud security: state of the art. In: *Security, Privacy and Trust in Cloud Systems* (S. Nepal and M. Pathan, eds), Springer, Berlin, pp. 3–44.
- 16 *Cloud Service Level Agreement Standardisation Guidelines* (2014) European Commission, Brussels. Available at: http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?action=display&doc_id=6138
- 17 Ali, M., Khan, S.U. and Vasilakos, A.V. (2015) Security in cloud computing: opportunities and challenges. *Information Science*, 305, 357–383.
- 18 Doelitzscher, F., Reich, C., Knahl, M. and Clarke, N. (2012) Understanding cloud audits. In: *Privacy and Security for Cloud Computing* (S. Pearson and G. Yee, eds), Springer, Berlin, pp. 125–163.
- 19 Building secure telco clouds, White paper. Nokia Networks, Tech. Rep. C401-01087-WP-201409-1-EN, 2014.

- 20 Petcu, D. (2014) SLA-based cloud security monitoring: Challenges, barriers, models and methods. In: *Euro-Par 2014: Parallel Processing Workshop* (L. Lopes, J. Žilinskas, A. Costan, R. Casella, G. Kecskemeti, *et al.*, eds), Springer, Berlin, pp. 359–370.
- 21 d. Oca, E.M. and Mallouli, W. (2015) Security aspects of SDMN. In: *Software Defined Mobile Networks: Beyond LTE Network Architecture* (M. Liyanage, A. Gurto and M. Ylianttila, eds), Wiley & Sons, Ltd, West Sussex, UK, pp. 331–356.
- 22 Vadalasetty, S.R. (2014) Security concerns in using open source software for enterprise requirements. SANS Institute Reading Room, October. Available at: <https://www.sans.org/reading-room/whitepapers/awareness/security-concerns-open-source-software-enterprise-requirements-1305>
- 23 *The HeartBleed Bug*. Available at: <http://heartbleed.com/>
- 24 CVE-2014-0160, Available at: <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2014-0160>
- 25 *Bash Code Injection Vulnerability via Specially Crafted Environment Variables (CVE-2014-6271, CVE-2014-7169)*. Available at: <https://access.redhat.com/articles/1200223>
- 26 CVE-2014-6271. Available at: <https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2014-6271>
- 27 Monshizadeh, M., Khatri, V. and Gurto, A. (2016) NFV Security considerations for cloud-based mobile virtual network operators. *Proceedings of the International Conference on Software, Telecommunications and Computer Networks (SoftCom)*.
- 28 ETSI GS NFV-SEC 003. *NFV Security; Security and Trust Guidance*. Available at: http://www.etsi.org/deliver/etsi_gs/NFV-SEC/001_099/003/01.01.01_60/gs_NFV-SEC003v010101p.pdf
- 29 ETSI GS NFV-SEC 001 v1.1.1. *Network Functions Virtualisation (NFV); NFV Security; Problem Statement*. Available at: http://www.etsi.org/deliver/etsi_gs/NFV-SEC/001_099/001/01.01.01_60/gs_NFV-SEC001v010101p.pdf
- 30 ETSI Group Specification: ETSI GS NFV 002. *Network Functions Virtualisation (NFV); Architectural Framework*. Available at: http://www.etsi.org/deliver/etsi_gs/NFV/001_099/002/01.02.01_60/gs_NFV002v010201p.pdf
- 31 Lukyanenko, A., Nikolaevskiy, I., Kuptsov, D., Gurto, A., Ghodsi, A. and Shenker, S. (2014) *STEM+: Allocating Bandwidth Fairly To Tasks. Technical Report TR-14-001*, ICSI, April. Available at: https://www.icsi.berkeley.edu/icsi/publication_details?n=3651
- 32 *OPNFV Delivers Open Source Software to Enable Deployment of Network Functions Virtualization Solutions*. Available at: <https://www.opnfv.org/news-faq/press-release/2016/09/open-source-nfv-project-delivers-third-platform-release-introduces-0>
- 33 *Technical Overview | Open Platform for NFV (OPNFV)*. Available at: <https://www.opnfv.org/software/technical-overview>
- 34 *Security Home – Security – OPNFV Wiki*. Available at: <https://wiki.opnfv.org/display/security/Security+Home>
- 35 Yrjo, R. and Rushil, D. (2011) Cloud computing in mobile networks – case MVNOM. *Proceedings of the 15th International Conference on Intelligence in Next Generation Networks (ICIN)*, pp. 253–258.
- 36 Monshizadeh, M. and Yan, Z. (2014) Security-related data mining. *Proceedings of the IEEE International Conference on Computer and Information Technology (CIT)*, pp. 775–782.

- 37 Liyanage, M., Abro, A.B., Ylianttila, Y. and Gurkov, A. (2016) Opportunities and challenges of software-defined mobile networks in network security. *IEEE Security & Privacy*, 14(4), 34–44.
- 38 Binu, S. and Misbahuddin, M. (2013) A survey of traditional and cloud specific security issues. In: *Security in Computing and Communications*. Springer, Berlin, pp. 110–129.
- 39 Zhang, N., Liu, D. and Zhang, Y. (2013) A research on cloud computing security. *Proceedings of the International Conference on Information Technology and Applications (ITA)*, pp. 370–373.
- 40 Chhabra, B. and Taneja, B. (2011) Cloud computing: towards risk assessment. In: *High Performance Architecture and Grid Computing* (A. Mantri, S. Nandi, G. Kumar and S. Kumar, eds). Springer, Berlin, pp. 84–91.
- 41 CVE-2016-0728. Available at: <https://cve.mitre.org/cgi-bin/cvename.cgi?name=cve-2016-0728>
- 42 Brunette, G. and Mogull, R. (2009) Security guidance for critical areas of focus in cloud computing, v2. 1, *Cloud Security Alliance*, pp. 1–76. Available at: <https://cloudsecurityalliance.org/guidance/csaguide.v2.1.pdf>

15

NFV and NFV-based Security Services

Wenjing Chu

Futurewei Technologies, Inc.

15.1 Introduction

In this chapter, we discuss 5G and security in the context of Network Functions Virtualization (NFV), a new transformative technology that combines the technology developments in distributed systems, software virtualization, and cloud computing to modernize telecommunication infrastructure and services. We will cover the content in three parts. The first part will discuss what NFV is and how NFV relates to 5G and security. In the second part, we survey the new security challenges that NFV brings and the new opportunities to solve security problems using NFV technologies. In the third part, we look at several exciting advancements in NFV-based security solutions that can have a huge impact on the success of 5G and networking in general.

15.2 5G, NFV and Security

5G is both a quantitative and qualitative great leap forward for the wireless industry. While the standardization and planning of early commercial deployments are just getting under way at the time of writing, the goals that the industry set are very ambitious and far-reaching. For example, the NGMN Alliance white paper [1] broadly defined the characteristics of 5G in three aspects:

- 1) Use cases, faster and more common broadband access, IoT, vehicular networks, verticals like healthcare and robotics, ultra-real-time tactile networks, AR/VR;
- 2) Business models, XaaS, value-add services and OTT applications; and
- 3) Value creation, encompassing diverse sets of value propositions to different customers ranging from consumers and enterprises to many vertical markets and partners.

The white paper systematically examined the requirements of 5G from user, system and business perspectives:

- *User experiences*: high data throughput, low latency;
- *System performance*: high user density, spectrum efficiency;

- *Device requirements:* e.g. battery life;
- *Enhanced services:* e.g. security, location services;
- *New business models:* XaaS; and
- *Network deployment, operation and management:* innovation agility, flexibility, operational efficiency.

By this broad definition of 5G, NFV and related technologies (e.g. SDN, telco cloud computing) are an integral part of what it means to be 5G. How we experience a 5G network as a consumer, an operator, or as a business or partner will be defined or made possible by the ideas and technologies of NFV. While some may still stick to a narrower view of defining 5G in terms of radio technology advances, it is clear to us that achieving the diverse use case goals in much broader industries outside of traditional mobile internet access and the success of 5G commercial ecosystems require NFV.

Regardless of how we prescribe the importance of NFV to 5G by definition or by necessity, we will look at the intersections of 5G and NFV in this chapter, and explore the problems and opportunities in the security domain. We will assume that readers are familiar with wireless networks and 5G. We start with a short introduction to NFV, and a section discussing the overlaps and differences of NFV, SDN and other cloud computing technologies. For the remainder of this chapter, we will use the term NFV in a broader sense to mean the wide body of technologies known collectively as “NFV and related technologies”. In other literature, people may also use terms such as Software Defined Infrastructure or Telco Cloud to mean the same concept of a cloud and software-centric infrastructure system.

15.3 A Brief Introduction to NFV

In the fall of 2012, a group of technologists from major telecom operators authored a white paper [2] that popularized the ideas around network function virtualization and served as a call to action. A few months later, in January 2013, the European Telecommunications Standards Institute (ETSI) launched the first of a series of NFV focused meetings in Sophia-Antipolis, France. This was the beginning of the NFV ISG (Industry Specification Group) [3] with a different working model that distinguishes itself from more traditional telecom industry standard bodies by refraining from heavy and rigid standard specifications and by embracing widely successful IT industry technologies and approaches such as commodity hardware, virtualization, software defined networking, cloud computing and open-source.

From the very beginning, the moniker NFV spanned a wide range of meanings. As a start, NFV means implementing network systems by using commodity servers, storage and network switches and realization of the network functions in software. However, the mandate of NFV goes far beyond this narrow scope. It encompasses the use of general-purpose operating systems, hypervisors and containers, cloud management software that enables XaaS, and the transformation of the network's OSS (operations support systems) and BSS (business support systems).

Let us first look at the simpler, narrow case of NFV. Figure 15.1 is a graphic depiction of this transformation of moving from vertically-integrated physical appliances to virtual appliances implemented on a virtualized infrastructure, based on a pool of commodity

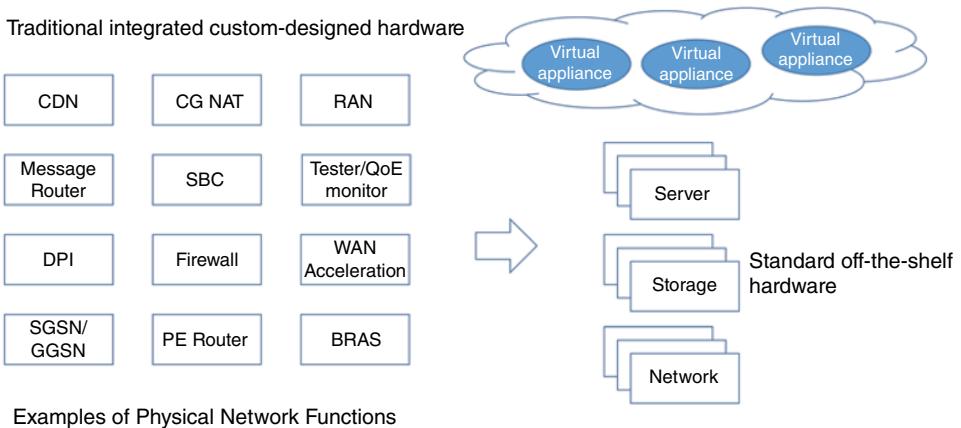


Figure 15.1 A simple view of NFV.

physical resources that can be shared and reallocated on demand. We will refer to the appliances shown on the left side of Figure 15.1 as Physical Network Functions (PNF), and the equivalent software implementation as Virtual Network Functions (VNF). From a security perspective, the VNF's role within a solution architecture stays the same as the corresponding PNF, but its implementation environment has changed. For example, the VNF's developers can no longer use customized physical design features for security (e.g. by limiting certain device management ports), nor can they do so to the common operating system or hypervisor or container environment in the virtualization layer. These common systems represent both a potentially larger attack surface and a place where fundamentally new security solutions can be delivered. It is critical that the virtualization layer provides equivalent or stronger security features for the VNFs. The virtualization layer is an intermediary that can potentially provide more unified, uniformly enforced, and stronger security than an individual PNF developer can achieve on his/her own. In addition, virtualization means that a VNF for security (e.g. a firewall or a DPI appliance) can be deployed to anywhere in the virtual infrastructure almost instantaneously with little overhead or cost. This also opens up new opportunities for stronger security features that were not feasible before NFV.

Beyond virtualization of infrastructure and VNFs, NFV also integrates the management, orchestration and OSS/BSS to the architecture picture. In many aspects, the management and orchestration layer may be more fundamental than virtualization itself. The ETSI NFV ISG's work in the NFV reference framework is the most widely known example of the current thinking in this area. In Figure 15.2, three components are introduced in the management and orchestration (MANO) area. The Virtual Infrastructure Manager (VIM) constructs an abstract consumption model for the virtualized infrastructure. The VNF Manager, whether it is a common instance for many VNFs, or a specific instance (e.g. per vendor) for a given subset of VNFs, helps to manage the lifecycle of the corresponding VNFs in a virtual infrastructure. The Orchestrator (at least in ETSI's naming convention) looks at the service-level abstractions, such as the service catalog, and presents those abstractions to the broader users of the overall system, such as OSS and BSS. While there are clearly alternative ways of

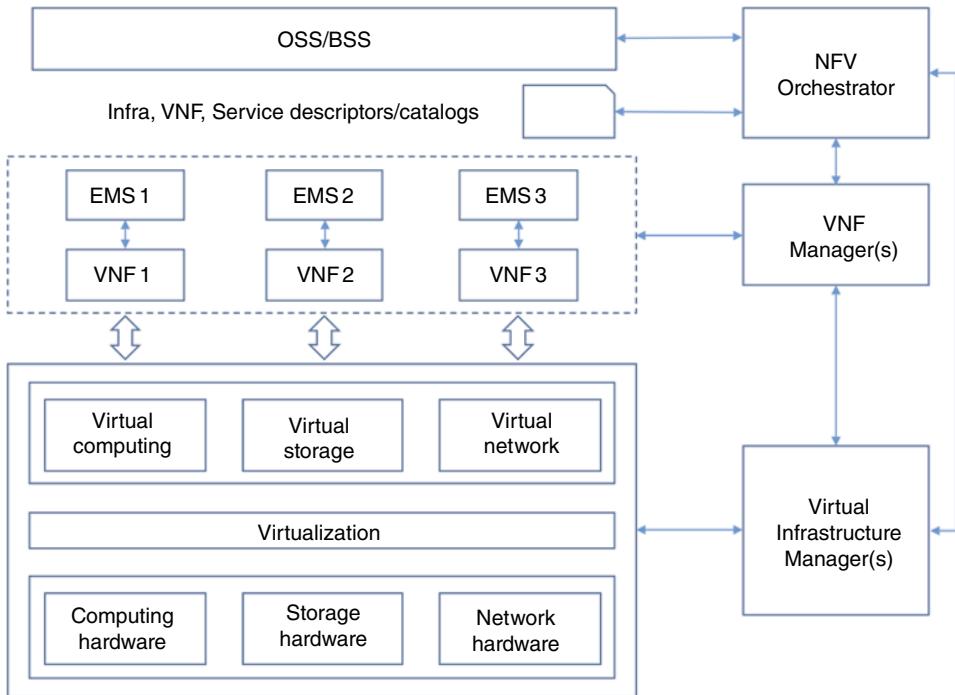


Figure 15.2 ETSI NFV reference framework.

constructing such an architecture, and some aspects of ETSI's formulation are still works in progress, let us use the ETSI framework at a high level to help us familiarize ourselves with the functional components and vocabulary and thus gain some appreciation of NFV's security challenges and opportunities.

A deep dive into the ETSI NFV Reference Framework is beyond the scope of this chapter; we refer interested readers to [4]. NFV is abstracting and automating many of the operational and business processes by software, and as such it requires operators to redefine trust relationships between many existing components and roles. For example, in a physical infrastructure, creating a new private network may involve a request-and-approval process – and potentially, employees with physical access permissions, and additional testing procedures to finally deliver the requested private network. NFV enables an on-demand service model where the equivalent private network can be delivered immediately without any human intervention or physical rewiring. We therefore must have an automated mechanism for approval, authentication and accounting, and for validating network topology and security policy configurations. Automation allows uniformed policy enforcement, nimbler ways of defending against attacks, and deploying enforcement points to wherever they best fit the need. Like any other digitization of a human process, it also opens up attack surfaces and methods that may not exist in the physical world. These issues are similar to what we found in the IT cloud computing world, as are the solutions. We will therefore not spend much time on these general security questions. In Section 15.6, we will discuss this topic in more detail and also look at some open-source examples of how new generations of software mechanisms found in NFV can help us minimize security threats and enhance delivered services.

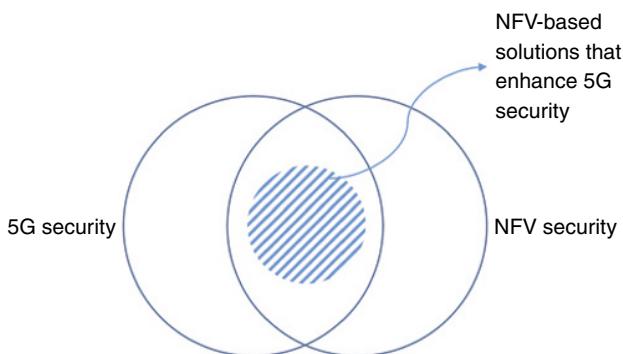


Figure 15.3 5G Security and NFV.

As we will stress in this chapter, NFV should also be viewed as a step towards a service provider cloud, with all the benefits of cloud computing. This includes business model innovations, for example, Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), adopted in the telecom world. These models of service deployment (collectively abbreviated as XaaS) bring up multi-tenancy-related security questions, as well as issues related to trust relationships and roles between different levels of service providers.

All of these NFV-based technology enablers promise a new generation of security products and services that will go far beyond what the industry has been accustomed to. Many researchers in academic and industry, as well as innovative entrepreneurs, have been actively working in this area and are bringing them to market in some cases. In Section 15.7, we will survey some of the work to illustrate the huge potential that NFV and the telco cloud bring to 5G and future networks.

Figure 15.3 summarizes the relationship between 5G security and NFV. The two significantly overlap; NFV brings new challenges related to security as well as new opportunities that can solve many security requirements in 5G and beyond.

Next, let us devote a short section to better clarify the overlap and differences between three terms that some use interchangeably but that others separate for various (and good) reasons; while many simply use the combination “NFV/SDN” to avoid this question. Our short discussion here is not motivated to sway opinions in any direction, but rather only to clarify and to avoid any confusion for the rest of the chapter.

15.4 NFV, SDN, and a Telco Cloud

Software Defined Networking (SDN) started as a proposal to separate a network switch's control plane from its data plane, and centralize the control plane in a software entity called a Controller, where new behaviors can be programmed in a relatively easy fashion without modifying the switch itself. It is also commonly associated with the protocol that the proponents proposed to realize such a design: OpenFlow [5]. Open Network Foundation (ONF) [6] was later formed to develop this vision further.

As with many good things in life, the basic concept of SDN has been stretched in all directions since its origin, and it can be difficult to tell when someone is referring to

OpenFlow style SDN, or even ONF's interpretation of SDN. The phrase "software defined" has also been applied to other components in the data center or other complex systems, such as "Software Defined Storage", or "Software Defined Data Center", or "Software Defined Infrastructure". We will not attempt to distinguish between these.

We will use the term SDN to mean a concept different from the OpenFlow protocol, and similarly to apply it to any network device other than just an Ethernet switch. We think the essential benefits of SDN are the central control of decision-making at a high level, not centralizing the entire control plane in OpenFlow. For the purpose of this chapter, the data plane device can be any Network Function (NF) – whether it is PNF or VNF, or a software entity in NFVI virtualization layer, such as a virtual switch or a virtual router. For instance, software-defined radio is a good example of SDN. The common entity is a logically centralized controller designed with modern interfaces and distributed, loosely-coupled architecture.

Under such a definition of SDN, the role of SDN falls into two common cases. In the first case, SDN is deployed to the networking subsystem in a data center. In this case, the SDN Controller can be thought of as a part of the Virtual Infrastructure Manager (VIM) (see Figure 15.2), and the corresponding data plane can be seen as the NFVI's hardware networking or virtual networking subsystems. In the second case, the SDN controller may reside within a data center while controlling network hardware devices outside of the data center, for instance, the aforementioned Software Defined Radio. Another example is Software Defined WAN (as in physical routers or optical switches of the wide area network). Opinions vary, but this second case is often seen as a separate domain, because the VIM usually does not provide an umbrella abstraction and common interface. This is partially because the VIM for a data center is often developed separately by the IT, not the CT industry, for enterprise uses. Service provider networks usually face fundamentally more complex requirements and constraints than the typical IT users. There is an active area of development on how to design or integrate these cases together. The following diagram in Figure 15.4 illustrates both cases in an end-to-end example of NFV and SDN, and this is the view that we will use throughout this chapter.

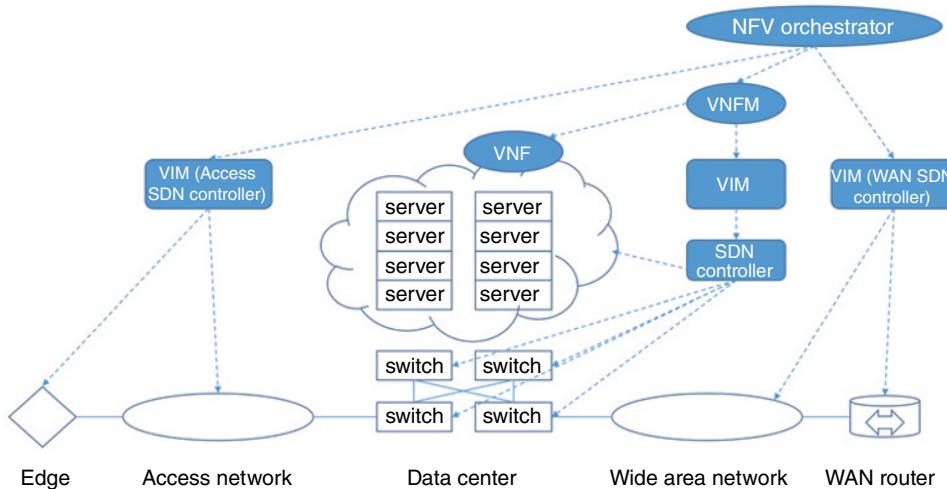


Figure 15.4 NFV and SDN end-to-end reference.

We believe that the efforts around SDN and NFV are part of an overall move towards a cloud model of developing, deploying, operating and monetizing operators' infrastructure, services – and increasingly, content. It is not coincidental that 5G and NFV happen at the same time. Going back to the NGMN Alliance's white paper on 5G goals, we can see that operators in the era of 5G do much more than provide a faster wireless pipe; 5G is much more than a speed dial-up from 4G. NFV, SDN and cloud technologies are integral to 5G. A telco cloud, in this sense, is the adoption of a cloud model in all aspects of an operator's business, and NFV and SDN are components of the strategy to move telco infrastructure, operational processes, and business models to that vision.

To summarize, we will primarily use the simple term NFV in this chapter in a broad and forward-looking context.

15.5 Common NFV Drivers

In this section, we briefly look at some common NFV drivers and use cases and how they relate to 5G and 5G security.

15.5.1 Technology Curve

Developing custom products for the telecom industry is a costly business. This cost shows up in the extraordinary long technology cycles. It is common for this process to take a decade or more from R&D to commonly-deployed services for customers. This cost also shows up in the proprietary systems such as gateways, switches, routers, and operating systems, middle-ware, and applications. The hardware is slower to upgrade because of this cost, which deprives the industry of the full benefits of Moore Law in silicon improvements. Its customized operating systems, tool chains, and middle-layer and application software do not fully take advantage of rapid software and methodology advancements, such as distributed system algorithms, distributed or NoSQL data bases, programmable APIs, Big Data, and machine learning, just to name a few. Other segments of the IT industry have taken leaps and bounds with these software technologies, which the telecom industry cannot readily adopt because of the proprietary approach it has taken. There are, of course, unique problems in the telecom industry, such as real-time constraints and high reliability needs. But the question is, can we achieve these based on standard COTS (common off-the-shelf) hardware and modern software technologies? The answer is not only yes, but that we must. NFV can help the telecom industry not only ride the technology curve but lead it.

15.5.2 Opportunity Cost and Competitive Landscape

Proprietary systems are not only costly to develop but also slow to adapt to new customer needs and too inflexible to compete with nimble new competitors. 5G's success depends on satisfying new customer needs that we may not fully appreciate yet, and tapping into innovative ideas of many different types of partners. IoT, connected and driverless cars, mobile healthcare – these are entirely new markets different from the traditional mobile phone services. VR/AR, immersive video, intelligent agents – these are brand-new user experience and engagement models. To fully seize these new

opportunities, agility and flexibility are a competitive advantage we must attain. New competitors, often referred to as OTTs (over the top), do not have the burden of legacy systems and can often outsmart telcos. Operators cannot justify the huge investment in 5G without a strategy to capture a large share of its promised bounty. NFV is part of that strategic toolset.

15.5.3 Horizontal Network Slicing

Networks are very expensive to build, maintain and operate. 5G envisions networks not only for mobile consumers on smart phones, but also for television, healthcare, automobiles, and vast swathes of other segments of the economy, as in IoT. Each of these networks has its own unique requirements. But we cannot build, maintain and operate entirely separate networks for each of them. It will be cost-prohibitive and extremely inefficient. What we need is a shared 5G mobile infrastructure that can be sliced end-to-end, or horizontally, into separate virtual networks that fit best for each market segment. This slicing applies to not only network access and topology but also to functions, such as security, provided by the slice. SDN and NFV together promise rapid provisioning and bringing up these slices on a shared and virtualized infrastructure.

15.5.4 Multi-Tenancy

Having different slices of networks is good, but we also need the virtual infrastructure to be able to support management and operations by different teams, by different functions in an organization, and by different partners and players in the industry in order to deliver sophisticated applications on shared infrastructure. NFV architecture is designed with multi-tenancy in mind as one of the major differences from traditional enterprise and telco systems.

15.5.5 Rapid Service Delivery

As we discussed earlier, the new competitive landscape demands business agility. But that is not all. Many types of requirements in 5G need the ability of rapid service delivery. For example, an “around the globe” service platform may require provisioning and tearing down of resources in a “follow the sun” pattern. Another example is the Continuous Integration/Continuous Deployment (CI/CD) model that requires applications to rapidly update live. Energy conservation is yet another example – being able to shut down part of the infrastructure when the work load is low, and then restart again when demands pick up. That also depends on the ability of rapid provisioning, and security, detecting attacks and rapidly provisioning resources to respond is a strategic arsenal in cyber defense.

15.5.6 XaaS Models

Service providers are no strangers to providing functions as a service, of course. The problem is that the legacy service definitions are rigid, hard to change (regulated by governments), and narrowly targeted. Cloud computing providers such as Amazon Web Services (AWS) pioneered much more flexible service models with a programmable API.

Whether the consumable functions are at the infrastructure level, Infrastructure-as-a-Service (IaaS), platform level (PaaS) or the software level (SaaS), or anything in between or in combinations, the flexibility and on-demand nature of these services open up a vibrant ecosystem of innovative partners who come up with a vast variety of ways to consume the shared infrastructure and create value. This type of ecosystem is a must for 5G network builders to fully leverage and monetize their expensive investment. Operators can also leverage each other's investments and provide better services to their customers. Monetizing fully the operators' (and vendors', and partners') collective investment is one of the most strategically important factors in the success of 5G.

15.5.7 One Cloud

We have talked about the telco cloud as powered by NFV and related technologies so far. But is there any fundamental reason that the telco cloud needs to be distinct from any IT cloud? If we look at the long horizon (or even if one looks back in history), the answer is no. Communication, as distinctive as it is, is closely intertwined with computing; future applications, as envisioned in 5G and in IT industries, predominantly require both computing and communication. Therefore, it is a fair question to ask: Is there a One Cloud into which we will eventually converge, regardless of IT or CT? It means that we are all competing towards the same technology foundation that will one day meet many, if not all, of our IT and CT needs. That day is already here in many market segments, and we may see the contour of what it looks like and make decisions to better prepare for the future. NFV is a major pillar in the foundation that will enable the networking industry to turn its expertise into advantages in the converged one cloud.

15.6 NFV Security: Challenges and Opportunities

Because of the dynamic and loosely-coupled nature of NFV, security – from solutions to processes – must be embedded into it from day one as a basic fabric. This security infrastructure, in identity services and role-based access control (RBAC) for example, has been developed and matured in the cloud computing space and is being adopted in NFV. In this section, we will focus on new challenges and new opportunities brought forth by NFV.

15.6.1 VNF Security Lifecycle and Trust

When a physical network device is introduced into an operational network, there is established trust of this device due to the fact that this device was installed and configured by a trusted employee, delivered to a secured location by a trusted courier, and developed and manufactured by a trusted vendor with a contract or a certification, and so on. For VNFs, this chain of trust relationships needs to be created and maintained in a NFV environment throughout its lifecycle.

A VNF's lifecycle is illustrated in Figure 15.5.

A VNF has several important trust relationships, as shown in Figure 15.6.

In a private NFV environment, as in a private cloud – where NFV infrastructure, MANO and OSS/BSS are in the hands of one administrator and all VNFs are managed

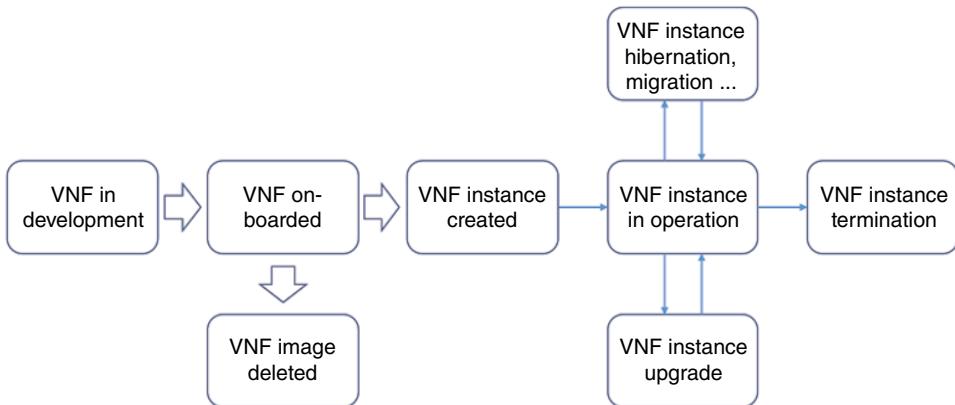


Figure 15.5 A VNFs lifecycle.

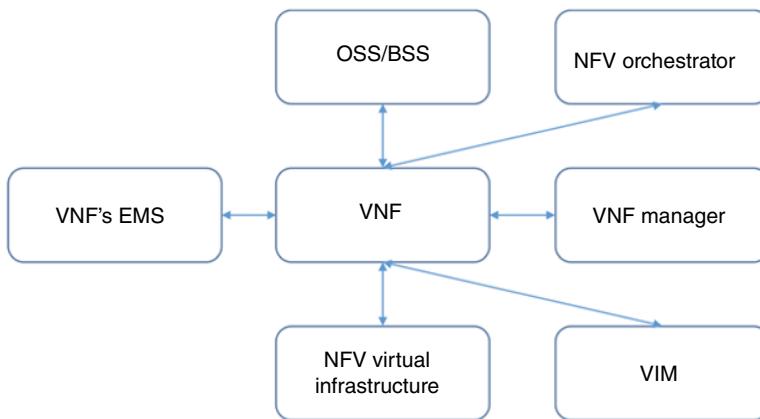


Figure 15.6 A VNF's trust relationship.

by one operator – trust relationships can be simplified. In this section, we focus on this simpler case. We will look at multi-tenancy and XaaS in a later section.

In the development phase, a VNF vendor must design the VNF for operation in an NFV environment. For physical devices, a lot of trust is often assumed. Even when a simple password mechanism is implemented, it is common that the password has a publically-known or easy-to-guess default value. The default password, or private cryptographic key, is often stored in the image itself. All these practices must be thoroughly eliminated in order for the VNF to operate in a much more open and less-trusty environment. Comprehensive public key authentication with an identity management infrastructure must be designed into software products, and backdoors eliminated. Software integrity must be strictly maintained at every step of the lifecycle with cryptographic signing using a traceable certificate from a Certificate Authority (CA) established or recognized by the operator.

In the on-boarding phase, the operator must be able to assure the VNF's authenticity and integrity by validating the vendor signature with an authorized CA. The operator

will still need to perform the normal acceptance testing procedure. However, with VNFs, this testing procedure can be fully or largely automated. As security-related risks or bugs are identified, new releases can be quickly on-boarded and tested again through full automation in a Continuous Integration/Continuous Deployment (CI/CD) chain. Security verification is an important step in that tool chain. It also ensures that VNFs from different vendors follow the same rigorous process to achieve same level of security level and same security architecture standard. An early example of an automation security testing component in CI/CD is Amazon Web Service's Inspector [7].

Once on-boarded, a VNF is ready for deployment. Deployment involves creation of an instance of the VNF, for example, one or more virtual machines or containers. It is given an identity and a private management IP address for communication, first with its VNF Manager (VNFM). The identity is usually provided as a service by the Virtual Infrastructure Manager (VIM). With the identity and management IP address, and the assistance of the Identity and Authentication/Authorization service, the VNF instance can now establish a trust relationship with VNFM, VIM and other MANO entities. While a VNF could be simple and contained within a single instance of VM or container, more often it is much more complex, with many VMs or containers collaborating to accomplish its functionalities. Each of these could have its own unique trusted identity, and the identity manager with the VIM facilitates trust establishment among them. It is also possible that a complex VNF or a legacy implementation may conduct the VNF internal trust management on its own. For example, a container manager or a vendor-specific PaaS layer may substitute for or complement the identity function.

The VNF Manager (VNFM) can be a common entity for many or all VNFs in a NFV environment. We can also see VNFM proprietary to a vendor and its VNFs. In either case, the VNF needs to locate the VNFM and then establish a trusted communication channel with the VNFM. Finding VNFM can be facilitated by the VIM's or MANO's instance creation phases via metadata to the VM (e.g. a model configuration), or by a common service discovery protocol. In a model-based approach, this VNFM identity can be one of the meta data in the descriptor that defines this VNF. There are efforts underway to standardize how a VNF should be formatted and described or modeled. Among other benefits, a model-based approach can ensure that security policies are uniformly expressed and enforced. Either way, the VNF establishes a trusted and secured management channel with the VNFM, and now it can be managed.

Let us first look at the simple case where the VNF is the only entity required for the example service. In this case, once the VNF can be securely managed and configured, it is now operational. There are a few things the VNF typically needs from the VIM to be fully operational, and they do have security implications. Common examples include publically addressable floating IP address, DNS service, and NTP time service. These services are usually provided by the VIM and NFVI, and therefore it is the responsibility of the operator to secure these services. DDoS attacks against DNS service have been widely reported in recent years, for example in [8]. The operator can also outsource some of these services, for example DNS, to professional DNS service providers. We will further discuss these XaaS scenarios in a later section.

The next step in a VNF's lifecycle is updating its software image. Software upgrade is a major undertaking in physical network devices, including extensive testing and preparation. In the NFV methodology, the Continuous Integration and Continuous Deployment (CI/CD) [9] paradigm is a large part of changing the whole feature delivery

pipeline to enable much more agile and flexible operation and business models. For the security domain, this is a great opportunity to create a threat intelligence pipeline, through which new threats are detected and analyzed, then new security patches or defenses are developed, tested and delivered through CI/CD routinely. The agility of this pipeline must beat the new threats in this game to maintain security in the future. NFV-based CI/CD is an enabler for this fundamental capability that future 5G systems must have.

The last stage of a VNF's lifecycle is termination. Some of the tasks in a termination phase are the removal of cryptographic material from the image, orderly archival and removal of kept data records, etc.

Further discussions regarding the VNF's lifecycle and security in each stage can be found in ETSI NFV ISG's TST working group publications such as [10].

To summarize, in this simpler, private NFV environment, a VNF's lifecycle is standardized by a model VNF Descriptor (VNFD) and automated through VIM, VNFM and service layer. This standardization and automation, coupled with rapid updating capability by CI/CD, will make security in NFV stronger than the manual system of today. We often hear that the most common security vulnerabilities are due to human error or human weakness exploitation. NFV-based automation will change the dynamics and allow operators to put their resources directly into defeating threats.

15.6.2 VNF Security in Operation

Managing a VNF's lifecycle is only one aspect of VNF security. A VNF spends most of its time in the operational state. Several important factors determine a VNF's security property.

First, let us remove one of the simplifications we made earlier: the assumption that a VNF is a single entity, for example, a virtual machine. More commonly, a VNF consists of a group of virtual machines with the same (a cluster) or different software images (VNF Components or VNFCs). In either case, each virtual machine has a unique identity and they are distributed to various computing units, not necessarily on the same physical servers.

These virtual machines will need a network over which they can: (i) coordinate their work by passing control messages among themselves, or (ii) pass around network packets (traffic) between them for distributed or pipeline processing. This is similar to a backplane in physical systems where control modules and line modules are interconnected, or there is an Ethernet-based network fabric in a data center. Inside a VNF, this interconnection is provided by a virtual network created dynamically. This virtual private network is provided on demand by virtual overlay such as VXLAN [11], a software abstraction supported by the hypervisor or by a virtual switch (vSwitch in layer 2) or a virtual router (vRouter in layer 3). Virtual networking or network slicing is a fundamental building block of the NFV infrastructure, and must support two basic requirements:

- 1) Topology isolation; and
- 2) Security rules.

Topology isolation means the virtual network is private, with its accessibility to the outside world fully controlled. Isolation alone is usually insufficient; security rules provide firewall-like security to the virtual machines that reside inside the virtual network.

These software abstractions can be several magnitudes more scalable than physical firewalls such that security rules can be applied with very fine granularity. This is sometimes referred to as micro-segmentation [12].

Virtual networks and security rules only protect a VNF from other virtual machines that are sharing the same infrastructure. VNFs must also connect with each other to collaboratively construct a network service. This is referred to as Service Function Chaining or SFC [13]. SFC can be realized by the same mechanism using virtual networks and security rules. It is not very efficient, however, because data packets must be in and out of processors repeatedly stressing the data path bottlenecks, and the common packet parsing and classification tasks have to be repeated in each stop of the pipeline where that information of a packet is needed. An active task in IETF [13] is defining a new encapsulation scheme for SFC to address some of these issues. SFC does not provide security protection between different VNFs, however. If security is needed, a virtual firewall VNF is inserted in the chain to work the same way as a physical firewall or security gateway.

Some of the VNFs that comprise the service must eventually connect with the outside world via physical network devices to backbone networks, access networks or devices that are geographically remote. The VNFs are gateways that must contain access to physical ports directly or indirectly. These edge VNFs must be protected from threats just like any other PNF in this case. They do have one advantage compared to the physical ones: they can scale up and down on demand, and that gives them better ability to deal with DDoS attacks.

Naturally, networking is the most important aspect of security for VNFs, but not the only one. Virtualization of computing and storage also exposes new attack surfaces to VNFs and adds new protection capabilities at the same time. Because VNFs share a server's processors, memory and disk with other virtual machines, the isolation capability provided by the hypervisor is usually not as strong as physical isolation. These types of issues have been addressed, however, and more capabilities are being added to a processor's architecture to strengthen protections. Other issues include the *noisy neighbor problem*, where a less-restrained virtual machine sharing the same physical processor resources could intentionally or inadvertently abuse them and cause performance degradation, or more severe issues. To properly address these types of issues, new generations of processors will add stronger isolation mechanisms, for example, Intel Xeon's cache reservation [14]. VIM's resource management software will develop schemes to allocate resources and place virtual machines more intelligently.

The virtualization layer, for example the hypervisor, provides a strong level of protection for the VNFs running on the top. Modern hypervisors all have built-in security features in addition to the protection provided by the operating systems. VNFs can also migrate, scaling up and down and in and out, and new security rules or security VNFs can be added on demand. All these functions substantially enhance the overall ability of the network services to withstand threats both old and new.

15.6.3 Multi-Tenancy and XaaS

So far we have assumed that the entire NFV system is private, that is administered by a single operator. NFV facilitates XaaS models in many different layers that can enable innovative business and operational approaches. For a given service, the provider must

support multi-tenancy in the infrastructure, so that the same infrastructure can be provided to more than one consumer in a shared fashion. For the consumers, this is functionality provided to them as a service (XaaS). They can build higher-layer applications on top of the XaaS interface.

Let us take an example of IaaS (infrastructure as a service) where the provider operator offers its data center infrastructure to partners. The infrastructure is NFVI and VIM in NFV terminology and it supports multi-tenancy. The provider who may own the infrastructure and the operation of it offers the use of this infrastructure in a cloud fashion. NFVI consists of the virtualized assets being shared, and VIM provides the interface or API so that the partners can consume this service. Multi-tenancy means the provider allows the partner/consumer to slice a virtual pool of computing resources under the partner's administrative control; the partner can virtually construct its own network services, composed of VNFs, within its slice of resources. Consumers of Amazon Web Services will be familiar with this consumption model; NFV extends this to telco services.

Multi-tenancy requires not only an efficient way of slicing the resources, but also security and privacy protections offered by the provider to its partners/consumers. Virtual private networks and their associated security rules, as described in the last section, are the basic foundation. But additional capabilities are required, including virtualization of supporting functions such as DNS, DHCP, identity, monitoring and reporting. Moreover, additional services are required for virtual gateway routing so that the virtual private network can securely connect with the Internet. We will often see that software originally developed for private use, even private cloud use, cannot adequately support these multi-tenancy requirements.

The provider may optionally support more features for its partners or consumers. These can be common tools that are needed to develop or deploy VNFs, such as performance monitoring and reporting, logging, data analytics, enhanced security services, databases, high availability, scaling and billing. These are often referred to as either Platform as a Service (PaaS) or Software as a Services (SaaS), depending on how we classify them into middle-ware or application categories. All these services will also need to support multi-tenancy.

With 5G networks and applications, this concept of resource-sharing and multi-tenancy are being pushed to new levels to include network-sharing. For instance, a new class of end-user devices (say, IoT or healthcare or automobiles) can be supported with a virtual slice of wireless access. Associated with each class is a slice of access network infrastructure (virtual networks), and a slice of virtual computing infrastructure and service applications (NFV data centers and VNFs). Each such class of services can have its own capacity, its own security policy, its own SLA, billing and support models, and so on.

15.6.4 OPNFV and Openstack: Open Source Projects for NFV

Why open source? Open source may seem like an off-topic subject for 5G, security or NFV, but it is not – for two very important reasons.

First, as we alluded to earlier, open source has been increasingly seen as an indispensable component in the development of NFV. The traditional standardization processes made tremendous accomplishments in getting mobile networks to where we are today.

That being said, we see shortcomings too, such as long and slow processes from ideas to products and services, and difficulty to adopt those kinds of processes to the software domain. Open-source projects such as Linux and Openstack have emerged as de facto standards within certain domains and can be just as effective in achieving many of the goals, like interoperability and rich ecosystems, which standards try to achieve. In the security realm, the prominence of open-source components like OpenSSL is a good example. Premium standard organizations such as ETSI, 3GPP and IETF recognize this trend and see the important role open-source is likely to play for NFV, IoT, and many other areas of 5G. OPNFV (Open Platform for NFV) [15] is one such project launched in September 2014 to accelerate the adoption of NFV by integrating an open-source reference platform for NFV.

Second, as more and more people come to realize, open-source cloud computing software based on commodity off-the-shelf hardware have made incredible progress in delivering unparalleled economic computing power in the last decade. One may venture to say that this is the main reason why we started the NFV initiative in the first place. Taking advantage of the open-source technology base to accelerate NFV seems a natural and obvious thing to do. Why re-invent the wheel when we can stand on the success that open-source already created for IT, web and cloud computing industries? That is why, in about two years after ETSI's launch of the NFV program, the Open Platform for NFV (OPNFV) project was created – with leading operators, network vendors, IT vendors, and open-source software developers joining together for the first time in industry history. It became a Linux Foundation project in September 2014 with a mission to “create a reference NFV platform to accelerate the transformation of enterprise and service provider networks”.

If we take a high-level framework like that of Figure 15.2, how do we then use open-source components to build a platform together to satisfy operators' needs? That is essentially what OPNFV set out to do (Figure 15.7). The OPNFV reference platform is the one example of NFV design that we can closely examine – and, for those of us who are so inclined, run it in a lab to experience it. The following diagram is a rendition of the OPNFV reference platform by the author. Many will readily admit that there is not one single reference platform but several different ways we can compose the platform together. Not all technical questions have been settled yet, as is often the case in the open-source world, but we can still benefit from studying their most recent release still in development at the time of writing, code-named Danube.

As shown in Figure 15.7, the centerpiece of the OPNFV reference platform is Openstack [16] as the VIM (virtual infrastructure manager). Openstack provides abstracted services as APIs.

Nova is Openstack's virtual computing API. For the NFV data plane, we first need a hypervisor, such as KVM [17] or LXD [18]. There are other choices for either hypervisor or container. Well-recognized examples include ESX® from VMWare and Docker®. For simplicity, we only show the components that have been actively integrated within the OPNFV at the time of writing (during the Danube release cycle). The same applies to all the discussions in this section.

The hypervisor is critical in providing security to VNFs in this environment. Hardening KVM is mandatory to prevent common vulnerabilities such as guest execution escape, guest-triggered DOS, and information leaks to guest. Modern hypervisors are safe for most common-use cases.

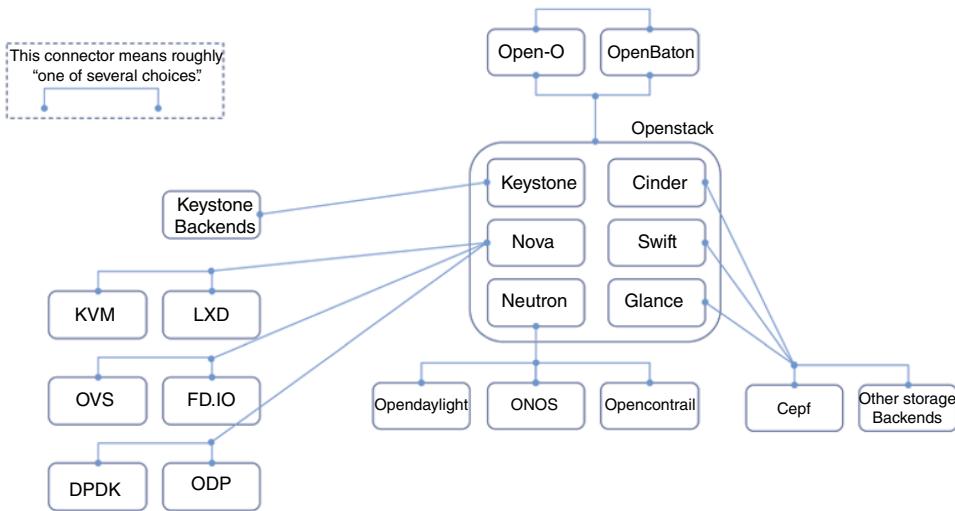


Figure 15.7 OPNFV reference platform for NFV.

In addition to the hypervisor/container, NFV requires special functions to support the network data plane. Open vSwitch (OVS) [19] is an open-source implementation of software-based switching capabilities based on SDN and is the most commonly deployed virtual switching implementation, according to Openstack user surveys. FD.io [20] is an alternative way of supporting a high-performance network data plane.

The networking data plane puts an extreme demand on packet movement (I/O) and processing to the underlying servers. The kernel TCP/IP stacks (and drivers) known to many software developers are often inadequate due to performance and other limitations. DPDK [21] is an open-source SDK that can improve, dramatically in some use cases, the data-packet movement and processing performance by taking advantage of special features in the processors. DPDK is developed by Intel® for Xeon processors. Similarly, ODP [22] (Open Data Plane) is an open-source project aiming to provide a cross-platform data plane SDK for SoCs and servers. The performance improvement levels of DPDK and ODP are particularly pronounced when we look at smaller packet sizes or packet-per-second measurements [23]. These smaller packets can be common in telecommunications networks for voice, media and others, as compared to IT and database applications in enterprises. In addition, packet-process latency or latency variation (jitter) can also be important considerations. More importantly, we also start to see that the narrow focus on SDK for performance may not be sufficient. A system's approach, where we develop a general framework that takes into account cluster management, reliability, programmability, and other system level issues, may hold more promise [24].

The data plane is the fundamental component of the fabric, for security or any networking functionality. One prominent feature of NFV is virtual network on demand. This network level isolation, in topology by layer 2 or layer 3 means, in virtual addressing, NAT, and security rules applied to the virtual network domains lay the bedrock for NFV security.

The counterpart of network data plane is its control plane. In Openstack, Neutron is the default virtual networking API, and we use an SDN controller to achieve many challenging goals such as scaling, flexibility and reliability. OPNFV integrated several well-developed options for the SDN controller. OpenDaylight (ODL) [25] is a sister Linux Foundation project focused on developing an open-source SDN controller. ODL is backed by many networking industry major corporations. The Open Network Operating Systems (ONOS) [26] is another open-source project spearheaded by ON.lab and other members, which has also become a Linux Foundation project. ONOS closely collaborates with the AT&T CORD (Central Office Re-architected as Data center) [27] project. As we have discussed, the SDN controller is the back end to deliver a certain networking area function abstraction. In Openstack, Neutron is the abstraction that comes natively by default. Let us take Open vSwitch (OVS) as an example of the data plane, and ODL as an example of the controller. Users make a Neutron API call to create an on-demand virtual network. This request passes to and is implemented by ODL, which in turn interacts with OVS via OpenFlow. There is much more complexity in the steps taken to realize the simple virtual network abstraction, of course, but Neutron abstraction itself is quite simple.

For NFV, Neutron's simple abstraction is often not enough, because it has to have the ability to represent many different networking technologies and use cases. For instances, virtualization of broadband access networks (vCPE: virtual Customer Premises Equipment) and VPN based on BGP/MPLS [28] for core networks will not fit into the simple Neutron model neatly. If we consider 5G use cases such as IoT, automobile, healthcare and other industries outside of traditional telco, this situation is even more complex. We have several ways to solve this problem. We could extend Neutron, but this option can make Neutron very complex and be a burden on enterprise applications that may not need or want that complexity. We could develop a separate service that specifically addresses telco networking needs (e.g. the Openstack Gluon project) [29]. Or, we could implement those networking abstractions outside of Openstack altogether by deploying SDN controller as a separate instance of VIM (Figure 15.4).

There are several abstractions in Openstack to capture different storage needs. Glance is an image service for all application software images; Swift is an API for object store; and Cinder is an API for block storage. OPNFV currently integrates a Ceph open-source project [30] to implement virtual storage from various storage hardware, for example, economical local hard drives in the servers. Other storage back ends exist and can be more common in deployments.

In Openstack, Keystone implements Openstack's identity API, supporting API client authentication, service discovery, and distributed multi-tenant authorization services for the entire Openstack deployment. A client asks Keystone for an authentication token by providing its valid credentials. This token is then used in the REST API to access an Openstack service, such as the X-Auth-Token request header. Keystone also supports integration with existing LDAP directories for authentication and authorization.

Finally, in the Danube release cycle, OPNFV is integrating two MANO (Management and Orchestration) open-source projects. OPEN-O [31] is a Linux Foundation-backed project "that enables telecommunications and cable operators to effectively deliver end-to-end services across Network Functions Virtualization (NFV) Infrastructure." The Open Baton project [32] is supported by TU Berlin, Fraunhofer FOKUS, and the

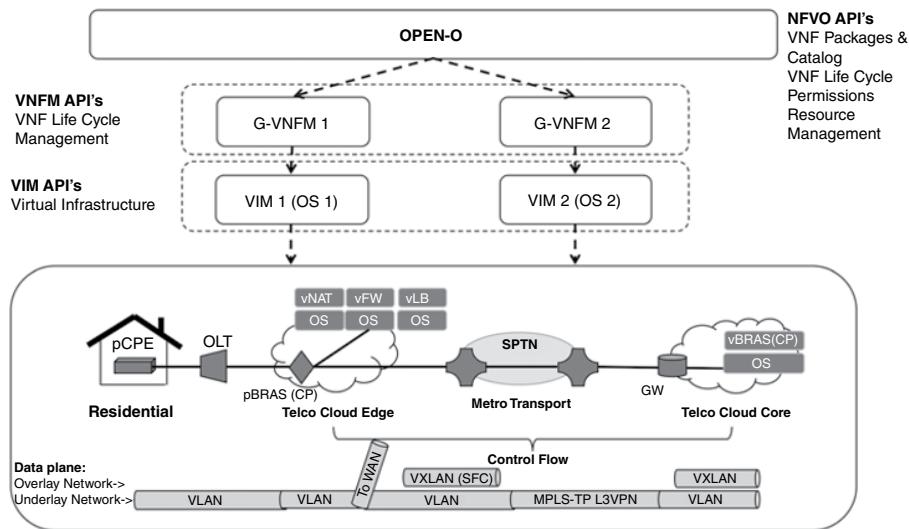


Figure 15.8 vCPE example use case by OPEN-O integrated with OPNFV.

5G-Berlin program in Germany, and it “is a ETSI NFV compliant MANO framework. It enables virtual Network Services deployments on top of heterogeneous NFV Infrastructures.” While these MANO projects are still new at the time of writing, they complete the OPNFV stack so that an end-to-end NFV deployment reference platform is now possible in open-source through OPNFV.¹

Figure 15.8² illustrates an end-to-end use case for an operator to deploy residential vCPE service with OPNFV integrated with OPEN-O VNF Manager and Orchestrator. This end-to-end network service slicing capability allows a common 5G network infrastructure to efficiently support multiple partners, multiple market segments, and use cases.

15.7 NFV-based Security Services

In the last part of this chapter, we will highlight several exciting advances in recent years regarding NFV-based security service. We call a new security service NFV-based if the technology has been made fundamentally more potent and impactful because of the development of NFV. These technologies may not be entirely new, but their full potential was not realized until NFV started to transform telecom industry infrastructure, services, and business models.

1 Other open-source MANO projects such as AT&T’s ECOMP [41] and Open Source MANO (OSM) [42] have not been integrated with OPNFV at the time of writing. For that reason alone, we choose not to discuss them in this section. Right before finalizing this chapter, we have just learnt that the Linux Foundation had announced a merger between Open-O and ECOMP [44].

2 Derived from image by OPNFV Opera project [15] and Open-O project [31] under CC-BY 3.0.

15.7.1 NFV-based Network Security

Network security services are typically provided by a gateway device. They are ubiquitous today in our computing and networking infrastructures in the forms of firewalls, DPI (deep packet inspection), IDS/IPS (intrusion detection systems or intrusion prevention systems), and many more. They can be stand-alone appliances or can be embedded into other network devices. These gateways are inline devices, meaning that they must be inserted into the network data path and handle the full load of traffic in order to be most effective. They are therefore prime candidates for virtualization. We look at different ways in an NFV system in which this can be achieved, and their benefits.

15.7.1.1 Virtual Security Appliances

Virtual security appliances are easy to adopt with NFV. The virtual appliance is a VNF that has equivalent security function as its PNF counterpart. The virtual appliance also has several fundamental advantages:

- The VNF can be created on-demand within seconds, anywhere in the NFV infrastructure resource pool;
- The VNF can easily scale up or down on demand;
- The VNF can be virtually *dropped in* to an existing network function chain;
- If the VNF is designed in a modern way, or *cloud native*, then it can also easily scale out horizontally in many cases;
- The VNF can be managed remotely and automatically by applying uniform security policies.

How can these properties of the VNF improve security? We can look at one recent security incident: the DDoS attack against DNS service provider Dyn.com. By Dyn's statement [33], the attack was the work of "tens of millions" IoT devices infected by the Mirai botnet. With the deployment of 5G and future IoT use explosion, we should expect similar incidents, with several orders of higher-scale magnitude. How can a NFV system better defend against such an attack?

First, NFV's dynamic scaling feature will be able to absorb the first wave of attack much more effectively. In a physical appliance deployment, the system's capability is static. Let us say the Dyn engineers foresaw an attack that may spike the load of the system by up to 10× of the normal load. That is the system capacity that they would have installed, and the malicious party only needs to generate 11× the load to cause a problem. In a NFV system, the entire data center's resources are a shared pool among many work loads. When the DNS system is under attack, the DNS system can scale up dynamically (including by shifting resource allocation from less critical system services) and give the system much more flexibility and time to absorb the attack wave, detect DDoS activities, analyze behaviors and even figure out remedies.

Second, even if the attacker can muster a large enough load to drain the extra capacity of the entire data center, Dyn's NAC operators would have much better MANO tools to deploy a solution quickly. The Dyn's official statement said that it took them a heroic "two hours" to recover from the first wave of the attack. A better MANO may reduce the deployment of the remedy in minutes or even seconds, and therefore practically render the attack a non-event.

Sometimes new attacks may be deployed that no one has seen before, and existing system software may not have a ready configuration we can use to fix the problem. NFV enables developers to rapidly prototype a new network gateway using programmable SDN, and to deploy it inline in the front end, via SFC, without interrupting the existing DNS system. This third method can be an important tool to deal with ever-more sophisticated zero-day attacks that we cannot foresee.

15.7.1.2 Distributed Network Security Services

Network security can also be distributed, either implemented within the hypervisor or as a hardened distributed service of the NFVI. Since the enforcement of security rules is dispersed to all servers, their load is distributed and scaled naturally, because a percentage of each server's computing resources are allocated to security. The network security "device" only exists virtually in a management system's abstraction. This allows users to use security gateways pervasively, since they basically have no material cost.

This last point is very important. Once the cost of adding a network security device is eliminated, we can abstract many security solutions to general policies, and the management system can automatically provision the necessary rules without human intervention – and the associated inevitable human errors.

This fine-grained abstraction of security also allows automated gathering logs for compliance audit and monitoring. It paves the way for further uniformed policy enforcement and audit.

15.7.1.3 Network Security as a Service

Many network security functions can be abstracted and delivered as a service. One of the common network security functions is VPN access. A corporation may deploy VPN gateways in its HQ or offices around the country or the world. Its employees from home or remote locations can safely access information and computing resources located in the HQ and data centers. This pattern is very common and ubiquitous. The same pattern exists for many systems, for example healthcare, or geographically dispersed IoT sensors.

This usage pattern can be abstracted and delivered as a service by one global operator to all its customers around the world, with much lower cost and higher ease of use.

In the healthcare setting, for example, hospitals and other care providers do not wish to spend the time and money to deal with IT systems, communication systems, and security and patient privacy law compliance. An operator can supply much better solutions, including certified compliance to regulations through an API-based service. Not only security and related regulations, but mobile healthcare services can be integrated seamlessly by software on top of a shared operator infrastructure based on NFV.

This XaaS pattern can be applied to many industries and market segments that are expected to become major use cases of 5G.

15.7.2 Policy-based Security Services

The second area we will look into is that of policy-based security services. Policy refers to a higher-level notion of what we want from our systems, as compared to a lower-level procedure of how. Policy-based management shows up in all areas of IT and non-IT human endeavors, from IT security, human resources and financials, to government and law.

Policy-based systems make declarations of what our goals are, without specifying how to achieve those goals. Policy is therefore also called *declarative*, and it must then rely on a system that can translate the declared goals and put them into practice or enforcements. This sounds a lot like how some governments are supposed to function. We can intuitively sense that this is easier *said* than *done*.

One problem with declarations is that they are often ambiguous. We therefore must have a clear language that can express all kinds of policies we may wish to impose. Another problem is that a group of declarations often contradict each other; we will need a way to resolve internal conflicts and still maintain integrity and consistency. But even with a clear and consistent policy specification, the question remains: how do you put the policy into force? Even with a correct procedure implemented in a system that can enforce the said policy, we still do not know if it is truly being enforced. You may say this is two sides of the same coin. If you are mathematically inclined or trained in computer science theories, you know this is indeed an unsolvable problem.

Because of these difficulties, policy-based management of security, networks, and other areas have been studied but not widely implemented on a large scale. With NFV, we see a new possibility that we may come to achieve real deployment, and dramatically simplify and enhance security at the same time. This optimism comes from the properties of NFV systems that we have been discussing throughout this chapter. To illustrate, let us look at some of most recent open-source solutions in this space.

15.7.2.1 Group-based Policy

Network engineers have been configuring access control lists (ACL) or firewall rules for years. These rules may be clear but they are extremely brittle and fragile, hard to maintain because they are specific to the details of each system, and difficult to prove correct. They are also primitive and not very effective.

Group-Based Policy (GBP) is a high-level abstraction to solve some of these challenges. A group is an abstraction of a collection of network endpoints and description of their properties. In NFV, this can be a group of VNFs. A VNF can instantly inherit security and other policies by becoming a member of a group. In GBP, the actual rules are separated out of the devices with an abstract rule set describing the desired behavior. These abstract rules can therefore be “reused” by attaching to many groups. The reusable nature of the rule set allows operators to develop mature policies that comply with known regulations, such as PCI and those developed to meet IoT, healthcare or autonomous cars in the future. These rules include not only common ones that we know in firewalls or ACLs, but also NFV-powered service chaining rules and rerouting rules. For example, you can write rules to reroute “dirty” traffic through a firewall virus-scan or IDS. This redirection capability allows GBP to describe a NFV Forwarding Graph. GBP also allows rules to come from different sources, for example, VNF’s developers specifying the rules for the application, and data center operators specifying rules for operations and administration.

The example in Figure 15.9³ gives a good illustration of the concept of groups and NFV service chaining to enforce dynamic security policies in an NFV system.

³ Derived from image by Openstack Group Based Policy project, CC-BY-3.0 [43].

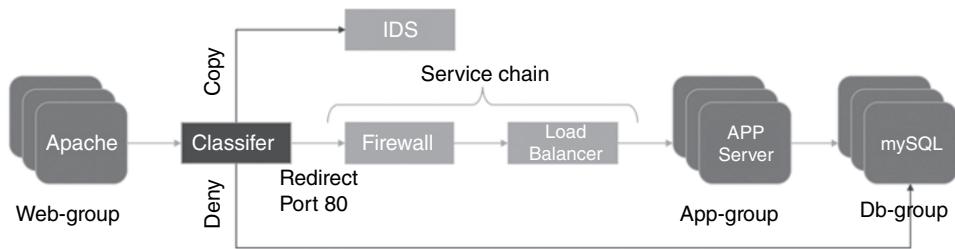


Figure 15.9 Example of GBP with NFV service chaining.

GBP is a policy framework specifically focused on networking and network-based security services. It is otherwise general and has been utilized in Openstack on top of Neutron [34], and in OpenDaylight [35], in addition to commercial products from several networking vendors.

Next, let us look at an example of a policy framework that is more general-purpose by design, beyond just networking.

15.7.2.2 Openstack Congress

The analogy of policy management in a data center to that of a government was not lost on the team developing a policy system for Openstack. They aptly named their system Congress.

Congress is a general-purpose policy framework that is implemented within Openstack as a service, although it could live outside of Openstack. The main components of Congress are:

- *A policy language*: Datalog [36];
- *Data source*: all main services in Openstack such as Nova, Neutron, Cinder, Glance, Swift, Keystone, other data collection services like Ceilometer/aodh, and non-Openstack systems, can all be its data sources; and
- *An implementation of policy services*: for monitoring, enforcement and audit of the policy.

Although the policy language, Datalog, is syntax-wise a subset of Prolog and is evaluated as first-order logic (therefore always consistent), it is closer to a query language like SQL. Congress can support many modes of policy services. For example, in *reactive mode*, Congress reports a policy violation after the fact, and can also trigger a corrective action. In *proactive mode*, Congress can check if a violation would occur if an action were taken and therefore prevent the violation from occurring in the first place. A consumer of Congress's service could consult Congress for pre-clearance. Congress can also be used for monitoring and reporting, and provide audit trails for compliance.

Because Datalog is a general language used for deductive reasoning, Congress can be used to express any policy, for example, environmental policy, power usage – or for our purposes, security and data center operations. Here is a simple example taken from OPNFV Copper project [37] that should be self-explanatory.

- Policy

“Certain software images are not allowed to run in the DMZ.”

- Enforcement

“If such an image is found running in the DMZ, pause the VM to stop such violation.”

- Congress policy specification in Datalog

```
dmz_server(x) :-  
    nova:servers(id = x, status = 'ACTIVE'),  
    neutronv2:ports(id, device_id, status = 'ACTIVE'),  
    neutronv2:security_group_port_bindings(id, sg),  
    neutronv2:security_groups(sg, name = 'dmz')  
  
dmz_placement_error(id) :-  
    nova:servers(id, name, hostId, status, tenant_id, user_id,  
image, flavor, az, hh),  
    not glancev2:tags(image, 'dmz'),  
    dmz_server(id)  
  
execute[nova:servers.pause(id)] :-  
    dmz_placement_error(id),  
    nova:servers(id, status = 'ACTIVE')
```

15.7.3 Machine Learning for NFV-based Security Services

Policy-based approaches tend to favor deductive reasoning by building models of reality and analyzing the models based on logical rules. As we know, this approach has its limits and does not work well in complex, dynamic and large systems by itself. An alternative way is to find truth from observation data and inductive reasoning.

Machine learning (ML) and Big Data have made rapid advancements and have had a huge impact on many problems that IT and CT industries face. It is not new that ML has played a large part in security – both on the positive side in protecting business systems and user privacy, and on the negative side as in hacking or unwanted surveillance. It is out of the scope of this short discussion to go into the general picture of ML applications in security and many of the envisioned 5G use cases. We will briefly touch upon several promising areas where ML can help NFV-based systems to better perform and protect [38,39].

- Autonomous Operations

Within an abstracted and service API-oriented system, ML can use collected real-time data to fine-tune optimization parameters of the overall resource management scheme. It is also promising that the ML system will be able to adapt to load spikes (e.g. during a DDoS attack) with autonomous responses by learning from long-term operational data collected from human expert operators. As we have reviewed throughout this chapter, NFV has made the system highly automatable and also made consistent real-time data readily available. In [38], we call such a highly autonomous system a Sentient Network.

- Self-Defense

There exists a large body of literature now about ML algorithm for anomaly detection and learning latent structures or patterns from network activities. Discovery of invariants in functional, operational, causal and other relationships are crucial in many complex cyber-physical systems such as a smart grid [40]. The telco physical infrastructure itself is such a system. ML can go beyond what traditional data analytics can deliver by discovering deep non-linear, distributed, and time-shifted relationships. Detection of these kinds of complex relationships, and anomalies within them, is a huge step towards creating a large self-defending NFV infrastructure and future applications that will run on it.

- Artificially-Intelligent Customer Service

Beyond the wonders of Photo Captioning, Natural Language Processing (NLP) systems and chat bots, AI and ML can also make the modeling of user needs much more intelligent. NFV allows these systems to have access to data far beyond just the Call Records and mobile device tracking, for example, into end-to-end customer services. Examples include seamless mobility between heterogeneous networks and voice or video emotion detection of quality problem detections. NFV's flexibility allows the system to address the problem before a customer's experience deteriorates. Security is a crucial part of this customer experience. Static systems often put security vs. privacy, or security vs. convenience into a zero-sum game. ML can help us differentiate and make choices to optimize the overall experience – without the hassle of making explicit trade-offs every time we face such a choice.

15.8 Conclusions

NFV and 5G security is a complex subject for a short chapter. We approached this topic with an introduction to the essence of what NFV is all about, and showed that an NFV-based telco cloud is fundamental to what 5G can be – much more than a faster wireless network. NFV brings to us new challenges in security, but also much more so in new opportunities – not only because of the capabilities NFV enables, but also because security is in the basic DNA of NFV from day one. These new capabilities will enable us to be ready for 5G and the envisioned 5G use cases in industries as diverse as health care, transportation, energy and media. Lastly, we introduced NFV-based security services, and looked into two areas where NFV can help transform how we meet security challenges in the future: policy-based and machine-learning-based security services. High-level abstraction and data-driven intelligent automation are absolutely essential to solve the future security problems that will be several orders of higher-scale magnitude.

References

- 1 NGMN Alliance (2015) 5G White Paper, February.
- 2 ETSI (2012) Network Functions Virtualization – An Introduction, Benefits, Enablers, Challenges and Call for Action [Online]. Available at: https://portal.etsi.org/nfv/nfv_white_paper.pdf

- 3 ETSI NFV ISG [Online]. Available at: <http://www.etsi.org/technologies-clusters/technologies/nfv>
- 4 ETSI GS NFV 002 (2013) NFV Architectural Framework [Online]. Available at: http://www.etsi.org/deliver/etsi_gs/nfv/001_099/002/01.01.01_60/gs_nfv002v010101p.pdf
- 5 Anderson, T., Balakrishnan, H., Parulkar, G., Peterson, Jennifer Rexford, J. et al. (2008) OpenFlow: enabling innovation in campus networks, *ACM SIGCOMM Computer Communication Review*, 38(2), 69–74.
- 6 Open Networking Foundation [Online]. Available at: <https://www.opennetworking.org/index.php>
- 7 AWS Inspector [Online]. Available at: <https://aws.amazon.com/inspector/>
- 8 *The Guardian*. DDoS attack that disrupted internet was largest of its kind in history, experts say [Online]. Available at: <https://www.theguardian.com/technology/2016/oct/26/ddos-attack-dyn-mirai-botnet>
- 9 Wikipedia. Continuous integration [Online]. Available at: https://en.wikipedia.org/wiki/Continuous_integration
- 10 ETSI NFV ISG (2014) *Security and Trust Guidance*.
- 11 IETF RFC 7348 (2014) Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks [Online]. Available at: <https://tools.ietf.org/html/rfc7348>
- 12 VMWare, Inc. (2014) Data Center Micro-Segmentation [Online]. Available at: <https://blogs.vmware.com/networkvirtualization/files/2014/06/VMware-SDDC-Micro-Segmentation-White-Paper.pdf>
- 13 IETF RFC 7665 (2015) Service Function Chaining (SFC) Architecture [Online]. Available at: <https://tools.ietf.org/html/rfc7665>
- 14 Intel Corp. (2015) Improving Real-Time Performance by Utilizing Cache Allocation Technology [Online]. Available at: <http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/cache-allocation-technology-white-paper.pdf>
- 15 Open Platform for NFV (OPNFV) [Online]. Available at: <https://www.opnfv.org/>
- 16 The Openstack Project [Online]. Available at: <https://www.openstack.org/>
- 17 Kernel-based Virtual Machine (KVM) [Online]. Available at: http://www.linux-kvm.org/page/Main_Page
- 18 Linux Container (LXD) [Online]. Available at: <https://www.ubuntu.com/cloud/lxd>
- 19 Open vSwitch [Online]. Available at: <http://openvswitch.org/>
- 20 The Fast Data Project (FD.io) [Online]. Available at: <https://fd.io/>
- 21 Data Plane Development Kit [Online]. Available at: <http://dpdk.org/>
- 22 The Open Data Plane Project [Online]. Available at: <http://opendataplane.org/>
- 23 Dell, Inc. (2015) Telecom TV [Online]. Available at: http://static.telecomtv.com/campaigns/Dell/Dell_Custom_High_Velocity_Cloud_FINAL.pdf
- 24 Lan, C. (2016) *A Framework for Network Function Virtualization*. Electrical Engineering and Computer Sciences, University of California at Berkeley, Technical Report No. UCB/EECS-2016-128.
- 25 OpenDaylight: Open Source SDN Platform [Online]. Available at: <https://www.opendaylight.org/>
- 26 The Open Network Operating System (ONOS) [Online]. Available at: <http://onosproject.org/>
- 27 Central Office Re-architected as Data Center [Online]. Available at: <http://opencord.org/>

- 28 IETF RFC 4364. BGP/MPLS IP Virtual Private Networks (VPNs) [Online]. Available at: <https://tools.ietf.org/html/rfc4364>
- 29 The Openstack Gluon Project [Online]. Available at: <https://wiki.openstack.org/wiki/Gluon>
- 30 The Ceph Project [Online]. Available at: <https://ceph.com/>
- 31 The Open Orchestrator Project [Online]. Available at: <https://www.open-o.org/>
- 32 The Open Baton Project [Online]. Available at: <https://openbaton.github.io/>
- 33 Kyle York (2016) Dyn Statement on 21 October 2016 DDoS Attack [Online]. Available at: <http://dyn.com/blog/dyn-statement-on-10212016-ddos-attack/>
- 34 Group Based Policy [Online]. Available at: <https://wiki.openstack.org/wiki/GroupBasedPolicy>
- 35 Group Based Policy in OpenDaylight [Online]. Available at: [https://wiki.opendaylight.org/view/Group_Based_Policy_\(GBP\)](https://wiki.opendaylight.org/view/Group_Based_Policy_(GBP))
- 36 McCarthy, J. (2016) Datalog: Deductive Database Programming [Online]. Available at: <https://docs.racket-lang.org/datalog/>
- 37 OPNFV Copper Project [Online]. Available at: <https://wiki.opnfv.org/display/copper/home>
- 38 Chu, W. (2016) A Sentient Network – How High-velocity Data and Machine Learning will Shape the Future of Communication Services [Online]. Available at: <http://www.slideshare.net/wenjingchu/a-sentient-network-how-highvelocity-data-and-machine-learning-will-shape-the-future-of-communication-services>
- 39 Nagra, J., Ahammad, P. and Jiang, H. (2016) SoK: Applying Machine Learning in Security – A Survey [Online]. Available at: <https://arxiv.org/pdf/1611.03186v1.pdf>
- 40 Zhang, J., Rahman, S., Sharma, R., Ramakrishnan, N. and Momtazpour, M. (2015) Analyzing Invariants in cyber-physical systems using Latent Factor Regression. in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- 41 AT&T ECOMP [Online]. Available at: <http://about.att.com/innovationblog/031716ecomp>
- 42 Open Source MANO (OSM) [Online]. Available at: <https://osm.etsi.org/>
- 43 Openstack GroupBasedPolocy project. Group Based Policy White Paper.[Online]. Available at: https://wiki.openstack.org/w/images/a/aa/Group-BasedPolicyWhitePaper_v3.pdf
- 44 The Linux Foundation. (2017) [Online]. Available at: <https://www.linuxfoundation.org/announcements/linux-foundation-announces-merger-of-open-source-ecomp-and-openo%C2%A0to-form-new-open>

16

Cloud and MEC Security

Jude Okwuibe, Madhusanka Liyanage, Ijaz Ahmad, and Mika Ylianttila

Centre for Wireless Communications, University of Oulu, Oulu, Finland

16.1 Introduction

The extreme capacity and performance demands of 5G networks will necessitate major changes in the way we produce and utilize digital services. 5G will usher in a massive array of new devices, services and use cases driven by technology developments and socio-economic transformations into the mobile network arena. Records will be set on every metric; ubiquitous ultra-broadband, virtual zero latency and gigabit experience will be the key defining features of the 5G networks. Certain key technologies and methods become pertinent towards the evolution to 5G; among these are *Cloud Computing* and *Multi-Access Edge Computing (MEC)*.

The relevance of cloud computing to mobile network has been on upward spiral over the last decade. Cloud computing provides on-demand computing resources and services on a scalable platform to both large and small organizations. At present, popular social media sites such as Facebook, Twitter, YouTube and Netflix are basically running on clouds. Besides, users are increasingly accustomed to carrying multiple mobile devices to meet their different lifestyle and work demands. This buttresses the need for cloud services in two ways; first, to ensure that user data is synchronized on all devices and second, to enable low storage devices to perform storage intensive operations leveraging on the cloud for virtual storage. This not only reduces cost of devices, but also extends battery life and improves overall user experience.

Another initiative for advancing the efficiency and dynamism in mobile networks is MEC. MEC is one of the most recent approaches towards the extension of cloud computing capabilities to the edge of the network. The main idea of the MEC initiative is bringing the clouds closer to the edge of the network as well as to the users.

Other similar approaches are Cloudlets, Fog, Edge computing¹, as well as their hybrids. MEC combines elements of Information Technology (IT) and telecommunication networking, hence extending cloud computing capabilities and IT services to the edge of cellular networks. This minimizes network congestion and also improves the overall performance of the network.

Bootstrapping MEC and cloud computing with their various deployment models to 5G networks come with many benefits to both the cloud service providers (CSPs) and to businesses. Such benefits include providing internet-based services to both small and large organizations at a highly reduced cost, hence creating a level playing field for different levels of investments. Other benefits include cutting down operational costs through reduced backhaul capacity requirement, eliminating resource redundancy through the provision of pay-as-you-use services, providing access to applications based on need; that is, businesses can now easily increase or decrease their capacities without incurring any unwarranted costs, flexible and rapid application deployment, secure radio access provisioning to application developers and content providers, as well as faster response time for cloud applications through reduced end-to-end (E2E) network latency and thus better Quality of Experience (QoE) for fast moving user equipment.

However, these remarkable benefits are not offered without cost. Transitioning to MEC and cloud computing poses a number of security risks to the industry, most of which are covered in this chapter. On the side of cloud computing, the idea of relieving clients of direct control, and transferring both infrastructure and resource management to the CSP could lead to trust issues on the network. Moreover, the centralization of resources requires CSPs to frequently update their cloud security and privacy protection mechanisms, so as to keep pace with the rapid evolution of cloud computing technology [1]. On the side of MEC, the security and privacy concerns are even more threatening, given that MEC is still in its infancy. Security concerns are mainly in the context of the *cloud-enabled IoT* (Internet of Things) environment. Security technologies are geared towards the MEC nodes, for example, MEC servers and other IoT nodes. Threats such as *man-in-the-middle (MitM)* and *malicious mode problems* have been identified [2]. Here, we describe in detail, and in the context of 5G networks, the security vulnerabilities of both cloud computing and MEC. We define different use cases of each technology and outline different associated privacy and security threats, and then we propose adequate solutions to address these security concerns.

16.2 Cloud Computing in 5G Networks

The fifth generation of mobile network standards (5G) will support a massive number of connected devices and provide wireless connectivity for a wide range of new applications and use cases. Unlike the previous generations of mobile networks, the

¹ There is a common misappropriation between edge computing and mobile edge computing. In contrast, edge computing comprises all technologies that leverage on distributed IT architecture to provide a means of collecting and processing data at local computing devices rather than in the cloud or remote data center, while mobile edge computing is one instance of edge computing at cellular base stations where cloud computing capabilities are moved to the edge of the cellular network. Other instances of edge computing are peer-to-peer ad hoc networking, fog/cloud/cloudlets, autonomic self-healing networks and dew computing.

5G network design is inspired by the need for more user-centric services across all network paraphernalia. Both existing and evolving systems, such as LTE-Advanced and WiFi, will be harnessed together with other revolutionary technologies in order to meet the anticipated performance demands of 5G. The Radio Access Technologies (RATs) of 5G will consist of existing RATs, licensed and unlicensed, supported by some novel RATs to support specific deployment scenarios and use cases, especially for ultra-dense deployments [3,4].

In all its ramifications, experts have shown that 5G mobile networks will be heavily supported by cloud services [3–5], whether for self-backhauling or for direct device-to-device (D2D) connectivity. Already cloud-based applications and storage are becoming common in modern networks; this is evident in the recent growth of uplink data in mobile networks. By integrating large-scale cloud architectures, 5G mobile networks need to be able to deliver services flexibly at unprecedented speeds to match the predicted growth in mobile data traffic that will be generated by mobile cloud services. In addition, the radio access infrastructure of 5G will largely depend on cloud architecture technologies for on-demand resource processing, storage and network capacity provisioning [5].

16.2.1 Overview and History of Cloud Computing

The actual origin of the term “cloud” in the context of computing has remained unclear in the literature; however, most technology historians agree that the idea of the cloud came originally from basic *virtualization* of IT infrastructures [7–10]. Prior to adopting the conventional term “cloud”, IT specialist had long practiced the art of replacing actual IT infrastructures with their virtual equivalents. This was initially adopted as a cost-saving strategy, but soon it also became a feasible means of improving flexibility and enhancing the speed of communication networks [6].

The initial idea of what we commonly refer to as “cloud computing” today can be traced back in the 1950s, when firms and organizations began to adopt and optimize the use of large-scale mainframe computers by allowing multiple users simultaneous access to both hardware resources and shared central processing units (CPUs). Companies like IBM and DEC were known for such solutions at the time. Going through the 1960s to the mid-1990s, the idea of cloud computing had evolved through various significant phases and milestones. This includes the *ARPANET* in the late 1960s and 1970s, the *CSNET* in the early 1980s and the *Telescript* in the mid-1990s. However, the most remarkable evolution was seen in the late 1990s, when the *Web 2.0* was introduced; for the first time in computing history, enterprise applications were able to be delivered over the internet [8]. Salesforce.com was a key player at this time, being one of the first companies to experiment with the idea of delivering contents over the web. About the same time, the *Virtual Private Network (VPN)* was born, allowing interconnection between multiple private networks over a public shared network such as the Internet.

The modern version of cloud computing came in the early 2000s after the dot-com bubble burst, with Amazon taking the lead with the implementation of a fully web-based retail service in 2002. Another milestone to cloud computing was seen in 2006 when Google launched its *Google Docs* services, which brought cloud computing services directly to end users [10,11]. In late 2000s, the advent of low-cost, high-capacity network and storage devices and infrastructures led to a wider adoption of the cloud

concept, with big players like NASA, Microsoft, IBM and Oracle all getting on board. At present, Oracle is in a bid to further the course of the cloud into the next generation of computing. With the goal of integrating all key IT service layers to the cloud, that is, the Applications (SaaS), the Platform (PaaS), and the Infrastructure (IaaS), this will become a key component of the modern cloud computing architecture.

16.2.2 Cloud Computing Architecture

The architecture of the cloud presented in Figure 16.1 consists of the essential components and characteristics, as well as the deployment and service models of modern-day cloud computing infrastructure. Together, these components are able to provide clients with the essential features for which the cloud is designed; such features as on-demand self-service, resource pooling, rapid elasticity, disaster recovery and broad network access [12].

The overall cloud architecture is grouped into two sections, the *front-end* and *back-end* platforms, both connected via a virtual network interface or the Internet, as shown in Figure 16.2.

- The *front-end* platform, also referred to as the client platform, consists mainly of applications and interfaces required to access the cloud system. Applications may vary, depending on the nature of cloud services. Email services, for instance, rely on the use of traditional web browsing apps like Chrome, Firefox, and Microsoft Edge, while file management cloud services may rely on the Windows Explorer application.

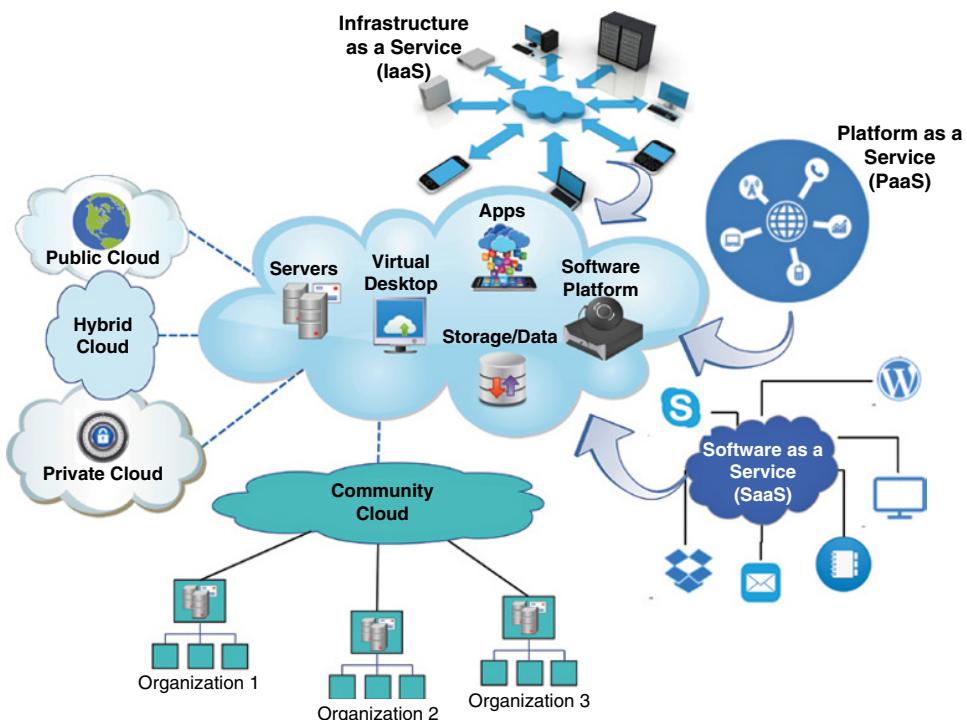


Figure 16.1 A view of the cloud architecture.

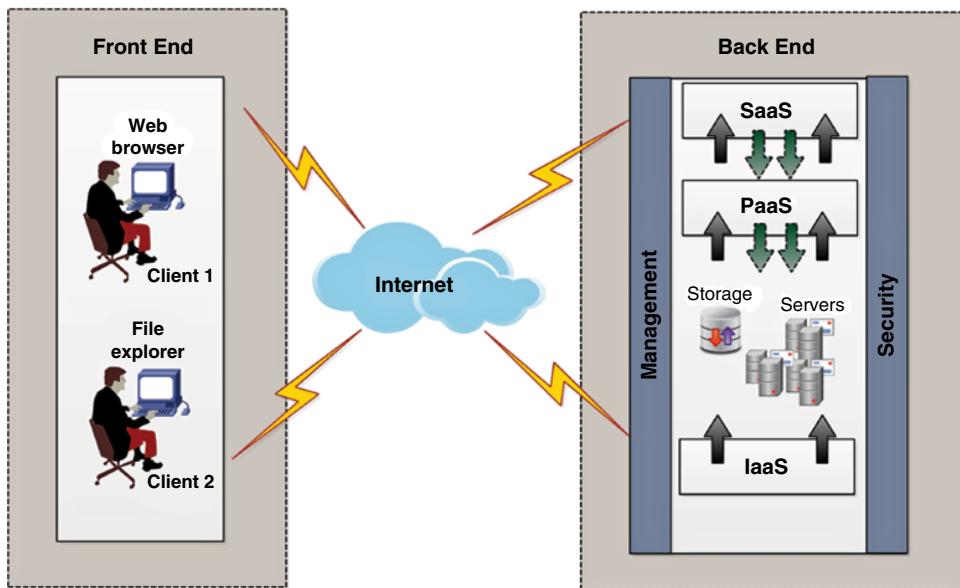


Figure 16.2 Front-end and back-end view of the cloud architecture.

- The *back-end* platform is the core part of the cloud computing architecture, also referred to as the “cloud”. The back-end platform is managed by the cloud service providers; it includes the servers, the cloud storage systems, and the virtual machines. It is in the back-end platform that the protocols designed to run the cloud computing platform run. It is also on this platform that the security mechanisms and traffic control of the cloud platform are provided.

16.2.3 Cloud Deployment Models

Due to several foreseen security and privacy concerns in the cloud environment, different deployment models have been designed to meet different levels of security and privacy for both the cloud infrastructure and stored user data. Four main deployment models are identified:

- 1) *Private Cloud*: The cloud infrastructure in this model is designed to serve the exclusive needs of a given enterprise or organization. These infrastructures may be hosted within or outside the enterprise and can be managed by either the enterprise or a third party. This deployment model demands more capital investment (CAPEX), especially for the hardware requirements. It is also most vulnerable to security and privacy threats, given that access must be exclusive to only the enterprise and with third-party service providers, trust issues may also arise.
- 2) *Public Cloud*: Provides cloud infrastructure for use by the general public. Public clouds share similar or the same architectural features as other deployment models. The key difference comes in their security and privacy policies. Public clouds are the most recognizable model of cloud computing. CSPs, such as Google and Microsoft,

and Amazon Web Service (AWS), offer several free cloud services to the public and these services are easily accessible over the Internet.

- 3) *Hybrid Cloud*: Integrates two or more distinct cloud servers to provide cloud infrastructure for different use cases. The idea is to combine the benefits of multiple deployment models. Cloud servers in hybrid clouds could be either a combination of private, public or community clouds. One of the popular use cases of hybrid clouds is in enhancing security and privacy on the cloud without incurring the overhead costs (CAPEX) of building a private cloud. In this case, non-critical resources like test workloads can be hosted in the public cloud, while critical resources like user data and workloads are hosted internally.
- 4) *Community Cloud*: A multi-tenant model that provides cloud computing infrastructures to multiple organizations within a specific community, who share common interests or concerns. Similar to the hybrid model, the infrastructures in this model could be placed within or outside these communities, and could be managed either internally or by a third party. Usually, the cost of running community clouds are distributed among selected members of the community and not each individual user [12,13].

16.2.4 Cloud Service Models

Different levels of abstraction constitute the back-end platform of the cloud architecture. These abstractions are grouped into the different service levels, depending on what resources are offered as a service for a given abstraction level, and this is depicted in Figure 16.2. According to National Institute of Standards and Technology (NIST), the three standard cloud service models are *Platform as a Service (PaaS)*, *Software as a Service (SaaS)*, and *Infrastructure as a Service (IaaS)* [12]:

- *Platform as a Service (PaaS)*: The transmission of platforms directly as services on the cloud is one of the integrated set of IT solutions that cloud computing provides. Typically, computer programs and mobile applications rely on some sort of middleware platform; this could be a set of hardware, combined with an Operating System (OS) and some libraries. PaaS provides this platform from the cloud, hence allowing users to develop and run applications without the overhead cost (CAPEX) of building and maintaining separate platforms [12,14].
- *Software as a Service (SaaS)*: Presently, SaaS is the most widely-used cloud service model in the IT industry. Using SaaS, a software application is hosted on the cloud and licensed to multiple users who are normally isolated from each other. Such services could be offered for free or on a pay-per-use basis over the web. Access is usually through some lightweight client interface such as a web browser [14]. Companies such as Adobe, Google, Microsoft, Facebook and Salesforce have been offering such services, some for decades now. Software applications such as MS Word, Excel, WhatsApp, Skype, and several others are now offered as SaaS through traditional web browsers.
- *Infrastructure as a Service (IaaS)*: This is the earliest and most fundamental cloud service model in computing. Long before the term IaaS was used, organizations were already optimizing their large-scale mainframe computers by allowing multiple users simultaneous access to both hardware resources and shared CPUs. IaaS may include virtual machines (VMs), servers, storage, network connectivity, firewalls and VLANs, all offered to users as provisioned services.

16.2.5 5G Cloud Computing Architecture

Cloud computing is one of the key technologies that will drive the evolution towards 5G mobile networks. According to the 5G Infrastructure Public-Private Partnership (5G PPP) Architecture Working Group, the overall 5G network architecture will be driven by an extreme demand for flexibility and programmability across all non-radio network segments, ranging from the fronthaul to the backhaul networks. Flexibility and programmability will also extend to mobile networks, access networks, core networks, and aggregation networks, as well as other evolving network segments such as the mobile edge networks, IoT networks, satellite networks, and the software defined cloud networks [15,16].

The 5G evolution will advance the convergence of multiple heterogeneous network environments, hence integrating a wide variety of network technologies for radio access and also introducing some novel access technologies to support specific deployment scenarios and use cases. Cloud computing capabilities will be leveraged to provision services at different network segments of the 5G network, from the Core Network (CN) segment to the Radio Access Network (RAN) segments, as shown in Figure 16.3.

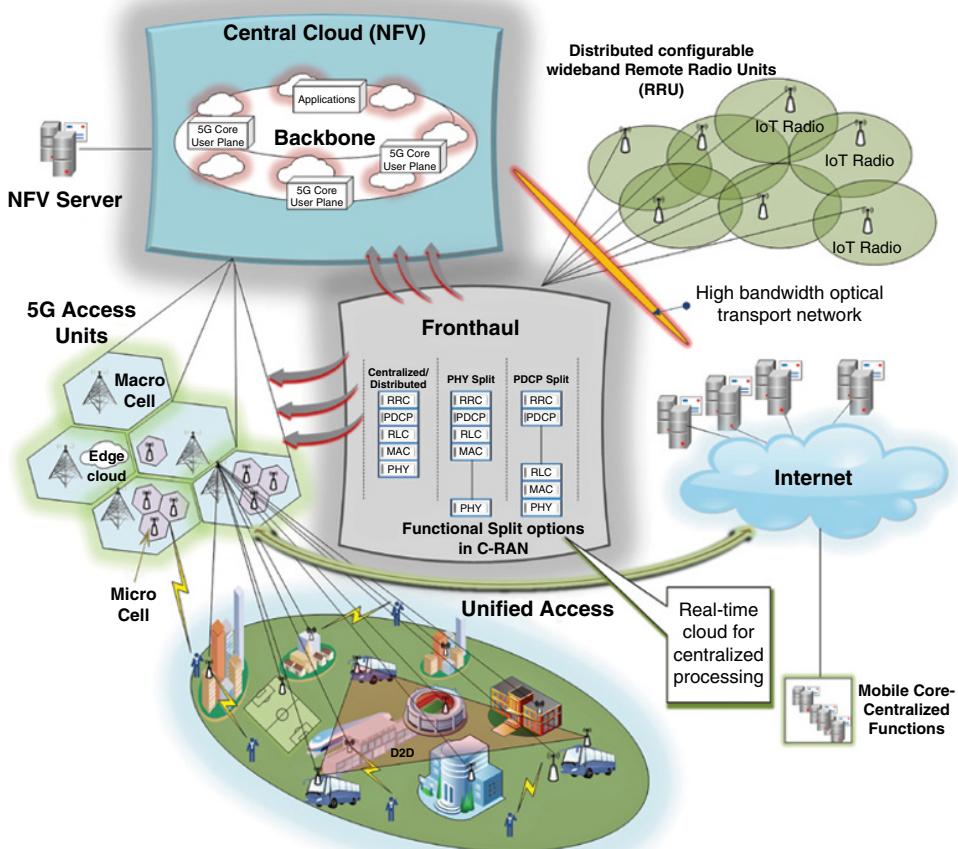


Figure 16.3 5G cloud computing architectures.

- *Cloud Computing in 5G Core Network:* The majority of 5G service plane and core network functions are anticipated to be deployed on cloud computing infrastructures. Cloud computing will be leveraged to handle various core network and service plane functions that are typical on cellular networks. These include multi-domain orchestration, service invocation, end-to-end (E2E) network slicing, and other service-tailored networks functions [16]. Traditional core network functions such as Authentication, Authorizations, and Accounting (AAA), security, traffic control and mobility management are also expected to be provisioned by cloud services.
- *Cloud Computing in 5G RAN:* The 5G framework proposes a novel radio access infrastructure based on cloud architecture technologies, using a scheme called Cloud RAN (C-RAN). Apparently, 5G mobile network will integrate existing RANs with novel access technologies in order to meet its performance and capacity demands; this combination will form the overall 5G RAT family. C-RAN is a novel access technology designed to extend Network Functions Virtualization (NFV) capabilities to the radio interface of future cellular networks. The overall idea of C-RAN is to promote spectral efficiency and multilayer interworking to support different use cases and technologies in 5G networks. C-RAN combines real-time virtualization and centralized baseband unit (BBU) pools to achieve large-scale centralized base station deployment on cellular networks [17].

16.2.6 Use Cases/Scenarios of Cloud Computing in 5G

Experts have defined and are already experimenting with numerous use cases of 5G networks. These use cases come from all major industries across the world, including manufacturing, healthcare, telecommunications, energy, TV and media, transportation, as well as other infrastructures. Although different use cases may tend to have different characteristics; however, for ease of understanding and clarity, METIS² and other groups like the 5G-PPP Architecture Working Group recognize a first-level grouping of 5G use cases into three main categories, namely *Extreme Mobile Broadband (xMBB)*, *Massive Machine-type Communications (mMTC)*, and *Ultra-reliable Machine-type Communications* [16,18]. Although this grouping provides a more or less generic view of all anticipated use cases in 5G networks, it is typical for several use cases to require similar network characteristics; hence, virtually all use cases will fall under one of these categories. Here we discuss each of these categories in the context of cloud computing:

- *Extreme Mobile Broadband (xMBB):* This group of use cases requires reliable provisioning of gigabytes on-demand bandwidth to support extreme business agility and guaranteed moderate rates to support less demanding applications on 5G networks. Typical use cases under this group are Virtual Interactive Presence and Augmented Reality (VIPAR) used in telemedicine, augmented reality, and tele-presence. The key enablers to this group of use cases are C-RAN and Mobile Cloud Computing (MCC). By ensuring unified dynamic operations on the radio

² Mobile and wireless communications enablers for Twenty-twenty (2020) Information Society.

access networks, cloud computing sets to enable broader network access to such applications through rapid elasticity and resource pooling.

- *Massive Machine-type Communications (mMTC)*: The main characteristic of this use case category is the massive number of connected devices. These use cases depend on the interworking of billions of sensors and actuators to support a huge number of low-cost and energy-constrained devices. Typical use cases under this group are mission-critical machine type communications such as remote surgery, industrial process automations, smart grids and intelligent transport systems. With the emphasis on latency, reliability and availability, the role of mobile edge clouds becomes essential in these uses cases.
- *Ultra-reliable Machine-type Communications (uMTC)*: These use cases are mainly characterized by high reliability and availability. Time-critical services and applications such as industrial control applications, V2X³, tactile network applications, autonomous driving, remote control over robots, IoT and other critical machine-type communications fall under this category. The major requirements for these use cases are fast discovery, immediate communication establishment, sporadic data handling and reliable feedbacks. The key enablers for the uMTC use cases are cloud storage and unified radio interface which is found in C-RAN [18].

16.3 MEC in 5G Networks

Virtualization and programmability introduces a major paradigm shift in the evolution towards the next generation of mobile networks. With virtualization, certain network functions, which are usually provisioned by proprietary network elements, are replaced with some form of virtual infrastructure like the cloud, with the aim of providing on-demand; cost-efficient and service oriented network services on-the-fly. Mobile Edge Computing (MEC) is an archetype of such virtualization at the edge of cellular networks, with an aim to reduce congestion at the core of the network; MEC is envisioned to be one of the driving technologies towards the 5G evolution.

16.3.1 Overview of MEC Computing

MEC is a new technology that is currently being standardized by the ETSI⁴ MEC Industry Specification Group (ISG), a European standard institute for Information and Communications Technologies (ICT). ETSI envisions that by providing IT and cloud computing capabilities within the RAN at the edge of mobile networks, developers and content providers can have direct access to real-time radio access information, this will also promote ultra-low latency, provide higher bandwidth and ensure improved overall user experience [19].

The MEC architecture is mainly a complementary unification of information technology and telecommunication domains in a virtualized platform served through a MEC hosting infrastructure, as shown in Figure 16.4. MEC mainly incorporates

³ Vehicle-to-Vehicle/Infrastructure.

⁴ European Telecommunications Standards Institute.

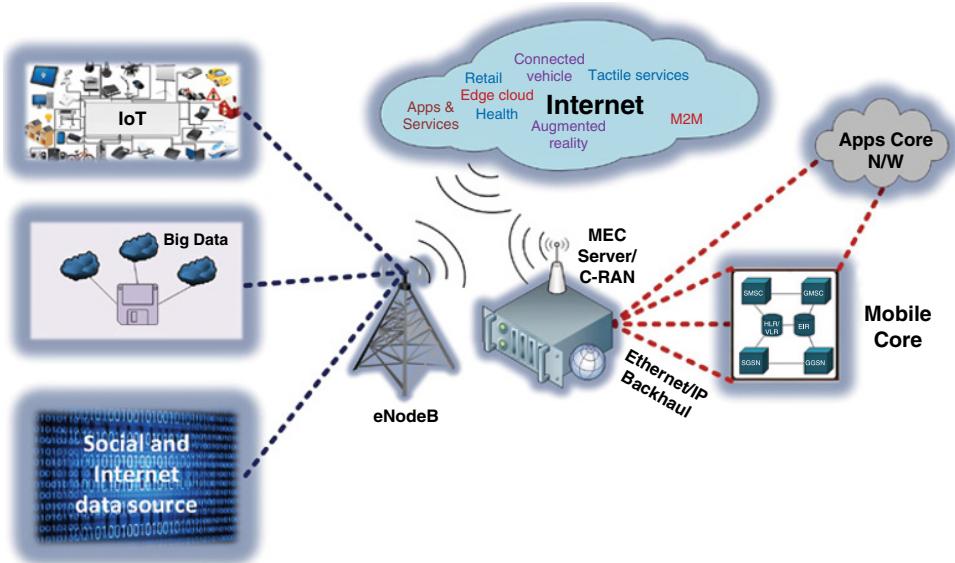


Figure 16.4 MEC system reference model.

application server platform into mobile base stations, its key characterizing features being proximity to data source, application services, RAN technologies and users [20]. MEC offers the following benefits:

- **Reduced Latency:** Extending computing resources to the edge of the network means faster arrival for network packets, and hence faster starts for streamed and real-time contents. MEC eliminates lengthy round trip delays to central servers and provides local support for application services, making the network more suitable for time-sensitive applications such as augmented reality and tactile networks thus allowing mobile devices to run computation-intensive and latency-critical applications more effectively. This also translates into improved QoE for the consumers [20].
- **Higher Efficiency Gains:** By storing contents locally at the edge of mobile networks closer to user applications, the need for frequent traveling of data through the backhaul channel is largely reduced, hence increasing efficiency gains. Contents targeted for the backhaul channel are more efficiently transmitted through opportunistic networking techniques, thereby reducing the signal load of the core network.
- **Backhaul Capacity Gains:** This comes from the reduced volume of signaling on the core network. Content caching has shown the potential for up to 35% reduction in backhaul capacity requirement [22]. Caching of local Domain Name System (DNS) could result to as much as 20% reduction in content download time [21].
- **Cost Savings:** This comes in multiple folds; first we have major cost savings on data delivery and in processing power and energy requirements; this constitutes a major part of the CAPEX for service providers. Then we also have some critical backhaul expenditure savings, which translates into reduced OPEX.
- **Real-time RAN Information:** This benefit is mainly experienced by application designers and content providers. With MEC, operators can open their RAN edge to authorized third parties, hence providing real-time network information at the edge

of the network, where they can easily be accessed and used by developers to optimize user applications. In turn, the RAN provider is able to utilize information from such third parties to make more efficient resource allocations decisions on the network.

16.3.2 MEC in 5G

Network softwarization will be a major driving force for the evolution towards 5G. Emerging technologies such as Software Defined Networking (SDN), Network Functions Virtualization (NFV), Fog Computing and MEC, will play a critical role in this process. The overall target of these technologies is to advance agility, flexibility and scalability at various points on the network. MEC extends IT services and cloud-computing capabilities to the edge of the network away from centralized nodes. It also extends similar functionalities to the radio access networks and to the mobile subscribers. For operators, MEC platform enables them to provide new services through the open RAN edge; this allows application designers to offer over-the-top (OTT) services using the MEC servers.

It suffices to say that MEC as a technology is still relatively in its infancy. However, according to 5G-PPP, MEC is one of the key technologies and architectural concepts that will drive the evolution to 5G networks. Other related technologies are SDN and NFV. MEC will contribute by no small measure in the realization of the cardinal objectives of 5G in terms of throughput, latency, scalability and automation [23]. MEC will serve as a key enabler to edge applications on 5G, while NFV⁵ will be focused on the network functions. With the emphasis on infrastructure re-use, these twin concepts could quite readily be combined in a complementary fashion; hence the 5G design makes provision for a dual hosting of both Virtual Network Functions (VNF) and MEC using the same server. The benefit to this modality comes from optimal infrastructure re-use, which constitutes a major savings on investment [25]. The MEC server is designed for optimal softwarization of functions and efficient infrastructure utilization [26]. As shown in Figure 16.5, all MEC applications and application platform services are software applications running on hardware components. With this design, it is possible to lower the cost of hardware components by combining off-the-shelf components with function virtualizations. For instance, the MEC virtualization manager layer depicted in Figure 16.5 provides IaaS facilities, which will provision for flexible and efficient multi-tenancy, run-time and hosting environment for MEC application platform services [27].

The need for MEC comes natural with the convergence of IT and telecommunications networking. MEC proposes a paradigm shift in existing communication ecosystems that will enable a new vertical business segment and services for both consumers and enterprise customers. By allowing for content caching at the network edge, the core network is relieved of congestion, hence providing a more efficient platform for resource-demanding applications and use cases at the edge, such as augmented reality, video analytics, location services, IoT and other such use cases [24]. The next subsection provides more information on some of these use cases.

⁵ NFV and MEC tend to co-exist in many technical contexts, because they are similar in various ways and also have similar origin; however, they are unique in the network segments they target, while NFV is targeted towards a variety of network functions and applications, i.e. routing, security, firewall, and VPNs. MEC is mainly targeted towards functions associated with wireless services at the edge of the RAN.

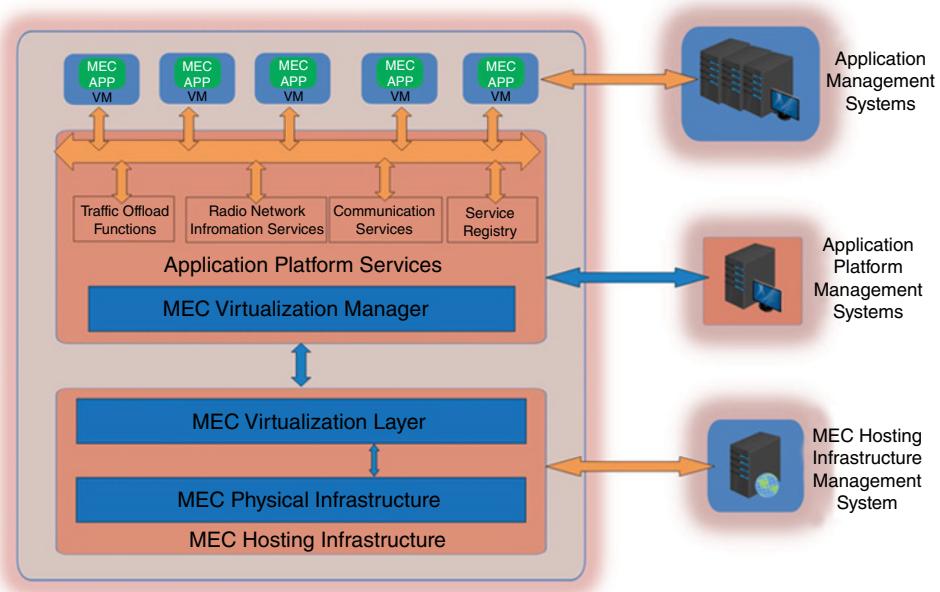


Figure 16.5 MEC server platform.

16.3.3 Use Cases of MEC Computing in 5G

The open architecture of MEC makes it a suitable option for a number of novel applications and use cases. However, given that MEC is still in its infancy, most of its potential use cases and scenarios are not typical in the current networking environment. With existing network infrastructure, a few use cases are already being tested and verified; among these are active device location tracking and distributed content and Domain Name System (DNS) caching:

- *Active Device Location Tracking:* This use case combines a third-party geo-location algorithm with an MEC-based application to perform real-time network measurement. This enables the tracking of devices without reliance on conventional Global Positioning System (GPS) devices.
- *Distributed Content and DNS Caching:* Just like the caching services in conventional web browsers speeding up access to frequently visited sites, this use case is made effective by minimizing the load at the server through caching, hence providing faster data delivery to customers. According to a report by BT TSO – Research & Technology, content caching has the potential to reduce backhaul capacity requirements by up to 35%. Local Domain Name System (DNS) caching can reduce web page download time by 20% [29].

Other experimental use cases of MEC include:

- *Video Analytics:* which uses a video management application to process and store video data, and using some predefined algorithms; it occasionally collects new video streams and compares with pre-recorded streams to detect changes in the environment. Typical applications under this use case are the smart city and public security [28].

- *Augmented Reality:* Here the MEC server uses real-time tracking and content caching to support augmented reality contents on mobile devices. The key driver for this use case is reduced round-trip-time (RTT) and higher throughput.
- *RAN-Aware Content Optimization:* By providing accurate real-time subscriber RAN information to the content optimizer, MEC helps providers achieve dynamic content optimization, and improved QoE and overall network efficiency, which stirs novel services and revenue opportunities on the network [28].

16.4 Security Challenges in 5G Cloud

With the advent of 5G networks, myriads of new businesses, trust models, new mobile technologies and new service delivery models will gain momentum on a global scale. 5G will play active roles in almost all aspects of our day-to-day life. This evolution will attract a corresponding evolution in the threat landscape and increase privacy in concerns at different application areas, hence security and privacy will play a central role in the evolution towards 5G.

The security and privacy concerns in 5G will span beyond technologies involving regulation and legal frameworks. Certain use cases pose special security concerns due to the nature of the applications they support. For instance, in online business, security concerns are indispensable to all parties, as a breach in security can lead to major flaws in trust among transacting parties and this could stagnate the adoption of certain novel technologies in 5G. Moreover, one of the cardinal goals of 5G is to realize a communication network that is more user-centric than previous generations. An indicator to this goal is that users become more aware of the nature of services they receive and the corresponding security vulnerabilities associated with these services. Hence, a highly optimized service with a porous security or privacy model could easily be rejected by users. With big data analytics, such security and privacy concerns become even more threatening. With unification of multiple communication domains, a security breach in one network segment could result to a ripple-effect across other segments. In this section, we discuss the security vulnerabilities that come with the cloudification of services on different segments of the 5G networks.

16.4.1 Virtualization Security

As discussed in previous sections, virtualization is the main driving force to cloud computing. Obviously, cloud computing will leverage on multiple virtualized systems in order to optimize available resources and deliver on its proposed benefits. The term “cloud” was initially adopted in science to represent a high level of abstraction in describing a collection of resources whose complexities are left undefined in a given context. This interpretation still remains relevant in defining cloud computing. Users still lack a clear understanding of the modalities to the processing and storage of their data, the location of storage, and the security of the entire process.

Security concerns with virtualization may range from potentials for data misplacement to Denial of Service (DoS) attacks. In [30], Lindstrom outlines five key security concerns of virtualization, which he called the *five immutable laws of virtualization security*:

- 1) An attack on a virtualized system is tantamount to attacking the actual hardware components that is virtualizes;
- 2) Security vulnerability of a virtualized system is the combination of actual system vulnerabilities and the vulnerabilities of the hypervisor⁶;
- 3) Security on virtualized systems can be improved by separating functionality and content, e.g. separating the user data from the network functions;
- 4) Aggregating multiple virtualization platforms on the same physical systems will increase risk, except if the hypervisor is configured to avert this risk; and
- 5) A trusted virtualized system on a mistrusted infrastructure is at higher risk than a mistrusted virtualized system on a trusted infrastructure.

16.4.2 Cyber-Physical System (CPS) Security

CPS integrates multiple networking, computing and physical resources on different spatial scales controlled by computer-based algorithms. CSPs consist of several communication technologies, sensors and actuators all linked together over a software system. Typical applications CPSs include monitoring and control of physical and organizational or business processes, for example, SCADA⁷, integration of different technical disciplines and application domains, for example, smart grids, distributed or interconnected systems of systems, as well as other sensor based smart communication systems [31].

Cloud-based Cyber-Physical Systems (CCPS) use *Cyber-Physical Clouds (CPC)* for the virtualization of network components like sensors and actuators. These virtualized components function as conventional cloud resources that can be provisioned for cloud services. CPCs are prone to several identified security attacks, such as *HTTP and XML Denial of Service (HX-DoS)* attacks. HX-DoS combines HTTP and XML messages flooded at rates deliberately meant to overwhelm the cloud CPS infrastructures. This attack can be launched on any cloud service models, that is, IaaS, SaaS and PaaS. In [32], authors present a defense system called Pre-Decision, Advance Decision, Learning System (ENDER); this system is designed to identify the HX-DoS messages before they are received by the targeted system. Another common attack on CPC is Slowly-increasing Polymorphic DDoS Attack Strategy (SIPDAS), identified in [33]. SIPDAS mainly characterizes DoS attacks that dynamically modify their behavior to evade pattern detection algorithms. Possible control measure to this attack is to frequently check the consumption of computational resources, as well as the intensity of incoming requests [33].

16.4.3 Secure and Private Data Computation

With cloud computing, user data is usually stored in the CSP data centers, which are usually unknown to the user. The security of such user data is crucial in any network environments and even more critical in cloud computing, given that user data could more easily be moved to any location on the globe over the clouds. It is therefore the responsibility of the CSP to ensure that both their infrastructures, user data and applications are protected.

⁶ In a virtualized system, the hypervisor is the software or firmware that controls the system. In this case, it is the cloud operating system. More common hypervisors are VMware and VirtualBox.

⁷ Supervisory Control and Data Acquisition

Several possibilities of attacks exist in this realm. One of the most threatening is the *insider attack*; this is considered as one of the largest threats in cloud computing as a whole [34]. Insiders in this context are the CSP staff with access to the physical servers on which user data is stored. Possible ways to mitigate this risk is for the CSP to ensure well coordinated routine background checks for such employees. Other related security threats are:

- *Abuse and nefarious use of cloud*: where CSPs offer cloud access to anonymous users who may turn out to be criminals or malicious code authors;
- *Insecure interfaces and APIs*: where CSPs provide customers with porous APIs, e.g. APIs that allow for anonymous access or clear-text authentication;
- *Shared infrastructure*: where different CSPs share common infrastructures such as CPU caches and GPUs, hence extending user data security risks beyond the actual CSP;
- *Data loss or leakage*: could be in the form of deletion or alteration of data without adequate backup facilities for recovery [34].

16.4.4 Cloud Intrusion

Cloud intrusion affects the availability, confidentiality and integrity of cloud resources and services. Intrusion could come in various forms, depending on the level of sophistication of the intruder and the nature of loopholes and weak links on the cloud environments. Intrusion may range from hobbyist hackers, to organized crime, to corporate espionage, or even nation-state sponsored intrusions.

The most conventional means to mitigating the possibilities of cloud intrusion is by building *intrusion detection systems (IDS)* and other control mechanisms in the cloud computing environment. The IDS monitors the cloud environment for malicious activities and policy violations. IDS could be attached to any of the cloud service models – IaaS, SaaS, PaaS; it can also be extended to network hosts as well as the hypervisor. In [35], Cox discussed evolving and advanced IDSs required in the cloud environment; these include the hypervisor-based intrusion detection system, traditional host intrusion detection system (HIDS), and network intrusion detection system (NIDS), and he recommends HIDS to be deployed on both the front and back end of the cloud architecture, while NIDS and hypervisor-based intrusion detection system should be completely left to the back-end where the CSPs operate. Another workable but rather passive approach to addressing such challenges is building intrusion-tolerant cloud applications, which are capable of ignoring potential malevolent requests. The limitation to this approach is that intruders could easily circumvent the mechanisms of such controls over time; hence the CSP will need to keep pace with evolving intrusion patterns.

16.4.5 Access Control

The cloud environment is a large open distributed system; therefore migrating to cloud technologies implies certain levels of access sharing on both data and network infrastructures. The main function of access control is to ensure that only authorized users are granted access to data and network infrastructure. Additional functions may include

monitoring and recording of unauthorized users attempting to access the system. Several access control models have been identified, ranging from traditional Mandatory Access Control (MAC), to Discretionary Access Control (DAC) and Role Based Access Control (RBAC)[36].

These access control models are mainly based on user identity, where a unique identifier is applied to each user, and upon successful authentication, users are granted access to corresponding resources. In the cloud environment, there is a need for a flexible access control mechanism to support various kinds of domains and policies. Access control goes beyond controlling access to resources and the system itself; it also includes the management of users, files and other resources. Typical access control system consists of functions such as authentication, authorization and accountability. In [37], the authors discuss the loopholes in the above access control methods and why they may not be suitable for the cloud computing environment. The MAC model does not guarantee complete secrecy of information, since it does not support the separation of duties and privileges, moreover it does not always support dynamic activation of access rights for certain tasks. The DAC model, which tends to be more flexible, lacks adequate mechanisms for managing improper rights, hence less risk-aware, and this makes it unsuitable for the access control level require in cloud computing. RBAC seems more advantageous in many cases; however, it also lacks adequate dynamism to its access control methods, for instance, it lacks the ability to classify information according to sensitivity levels, which will be a major requirement in the cloud environment, given that certain information is less sensitive than others. RBAC also lacks delegation mechanisms needed in organizations for situations where certain staff members are absent.

16.5 Security Challenges in 5G MEC

The MEC network environment comprises multiples diverse technologies, including wireless networking, distributed computing, and the virtualization of networking equipment and computing servers all interoperating in an open ecosystem where service providers can deploy their applications. The heterogeneity and diversity of the MEC environment opens up a series of avenues for malicious attacks and privacy issues that could constitute a major threat to the entire MEC system. By extending IT services and cloud computing capabilities to the edge of mobile networks, the MEC platform tends to have a limited size of hosts at the mobile edge, which cannot enjoy the same level of protection as conventional large data centers, hence the need for more robust security measures to mitigate these security challenges on such network edges.

In addition, the MEC system is still in its infancy, hence the security concerns are mainly in the context of a *cloud-enabled* IoT environment. Security technologies are geared towards the MEC nodes, for example, MEC server and other IoT nodes. Threats such as *man-in-the-middle (MitM)* and *malicious mode problems* have been identified [2,38]. In [38], the authors present a broad threat description model for the MEC system; this section discusses the threat landscape of the MEC system and why security is one of the greatest challenges of the MEC system. In Section 16.7, we suggest workable mitigation techniques that could be used to avert these security challenges.

16.5.1 Denial of Service (DoS) Attack

DoS attack is an age-long threat in various computing and networking arenas. DoS creates an artificial scarcity or lack of online resources and network services. DoS attacks could happen in the form of distributed denial-of-service (DDoS) or wireless jamming and could be launched on both the virtualization and network infrastructures. In the case of MEC systems, DoS attacks have limited scope, as described in [38], as a DoS attack on the network edge will affect only the attacked vicinity and not the entire network. So also an attack on the core network infrastructure might not lead to a complete disruption in the functionality of the edge data centers. This is due to the autonomous and semi-autonomous nature of their protocols and services.

Another loophole in the MEC architecture that makes it vulnerable to DoS attack is the combination of multiple Virtual Machines (VMs) spread across several mobile edge hosts. This formation increases the possibilities of compromising multiple VMs simultaneously, leading to large-scale attacks such as DDoS [39]. On the pros side, the MEC mobile network infrastructure by virtue of its design is inherently suited for deploying extended defense perimeters capable of mitigating the DDoS attack within multiple fronts. Such defense mechanism allows for fragmented deployments at the network edge, hence capable of defending against smaller attacks before they get any further chances of escalating.

16.5.2 Man-in-the-Middle (MitM)

The MitM attack is characterized by the presence of a third malicious party interposed between two or more communicating parties or entities and secretly relaying or altering the communication between such parties; a common example is the MitM attack between a server and a client. For the MEC scenario, a MitM attack is categorized as an infrastructure attack [38], where the malicious attacker tries to hijack a certain segment of the network and begins to launch attacks, such as eavesdropping and phishing, on connected devices.

The potency of an MitM attack on mobile networks has been proven in various works and literature [40,41]. In these works, MitM attack was launched between 3G and WLAN networks. Such attacks would be even more threatening for the MEC scenario, given that MEC relies heavily on virtualization, hence launching an MitM attack on multiple VMs could very easily affect all other elements on both sides of the attack.

16.5.3 Inconsistent Security Policies

In mobile networks, the need to preserve user-security parameters when they roam from one operator network to another is of utmost importance. In the case of MEC, it is highly possible that all the security services are not updated very frequently and per-user basis. MEC servers can also be limited with resources, thus making it more challenging to facilitate such services. When a user moves from one operator network to another, latency sensitive services are utilized, and such user might be provided with these services through the visited operator MEC, but will his security or the security of the service he is using be ensured? This buttresses the need for security policy sharing among network operators on a much faster scale, to ensure that users traffic on the access networks are effectively attached to the MEC their services are migrating to.

Furthermore, inconsistent or irreconcilable ingress/egress firewall security policies among roaming partners can equally prevent some roaming traffic bound for the Internet from working correctly, if at all they happen to work. This is particularly important in 5G scenarios, where local breakouts will be much more prevalent. These local break-out scenarios will cause roaming IP traffic to be routed over a visited network, where firewall security policies will very likely differ from the same policies deployed in the home network. The reasons motivating the need for local break-out scenarios in 5G are due to the need to reduce latencies and the planned deployment of wireless functionality together with subscriber content at the edge of the network, such as in MEC.

16.5.4 VM Manipulation

This attack is typical for all virtualized and edge computing systems. In MEC, VM manipulation mainly affects the virtualization infrastructures [38]. The adversary in VM manipulation is mostly a malicious insider with enough privileges or a VM that has escalated privileges⁸. Such adversary tends to launch multiple attacks to the VMs running inside it.

VM manipulation opens up the affected VMs to numerous other potential attacks such as logic bombs, malware and other such malicious elements that could compromise the security of other data centers when such VM migrates to other physical location on the network [38]. In [42], authors present an attack called DKSM (Direct Kernel Structure Manipulation), which can effectively alter the existing VM introspection solutions into providing false information. VM introspection is a specialized technique for determining specific aspects of a guest VM execution from outside the VM; however, with DKSM attack, the VM introspection solution gets subverted.

16.5.5 Privacy Leakage

Illegitimate access to the MEC environment by an adversary could compromise the privacy of certain information on both the network and the service infrastructures. On the network infrastructure, the MEC paradigm limits the scope of privacy leakage through some specialized functions at the edge data center. The data center mainly stores information from local entities and is also able to extract more sensitive information regarding the user through context awareness; hence, it will be possible to detect a malicious adversary on the network edge [38].

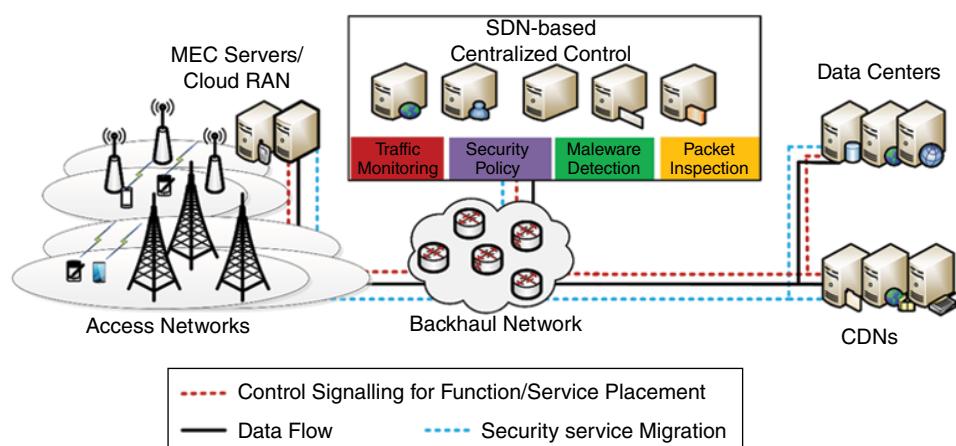
On the service infrastructure, the possibilities of adversaries accessing the information stored at the upper layer of the edge infrastructure could warrant substantial concerns for privacy leakage. However, similar to the case of network infrastructure, the potential damage of such a privacy breach is limited by the amount of information the adversary is able to gain access to. Typically, edge data centers allow for bypass of the central system and the exchange of information directly with each other, hence when information is processed at the lower layer, there are chances that the upper layer would only receive a subset of the processed information, thereby limiting the information made available to potential adversaries [38].

⁸ Privilege escalation happens when a malicious VM takes control of certain elements of the host.

16.6 Security Architectures for 5G Cloud and MEC

16.6.1 Centralized Security Architectures

The main benefit of centralization is the global visibility of network resource stats and policy synchronization. There are, of course, challenges with centralization, such as latency constraints, scalability and availability challenges; however, large systems that are not very latency sensitive, and centralized systems have more benefits as stated. Centralized control coupled with programmability enables run-time adoption of the network to the changing environment and business requirements. For example, if the traffic behavior changes, it will be beneficial to change the system behavior towards the traffic as well. SDN is one such candidate that enables centralized control of the network with programmability. The logically centralized control enabled by SDN brings dynamism in network security systems by harvesting intelligence from the network equipment through programmable Application Programming Interfaces (APIs). Using NFV, virtual security functions can be placed at any network perimeter, whenever the need arises using the programmable interfaces of SDN such as OpenFlow. In clouds and MEC scenarios, the security framework working as an SDN application can monitor the traffic or resource requests coming towards MEC or cloud infrastructures at a run time to validate the requests. In OpenFlow, each new request is forwarded to the centralized controller, where it can check the authenticity of the user. The controller can check the user credentials through the management plane or other integrity verification system such as HSS in the cellular domain. If the user is authentic, his request is served. If the user is malicious, his subsequent flows are either dropped or forwarded to another system for further analysis and counter actions [43]. Furthermore, the network administrator can prioritize and de-prioritize traffic, redirect or block traffic, change the security policies, and see the user behavior through software from the centralized control point without configuring individual devices. This minimizes the operational expenses (OpEx) for network operators. An integrated environment leveraging SDN-based centralized control framework is presented in Figure 16.6. With the centralized control platform, security services and



functions can be placed at the network edge through programmable APIs in the edge. This will enhance security of systems, such as C-RAN and MEC in the edge.

16.6.2 SDN-based Cloud Security Systems

The rise of SDN will change the dynamics around securing the data centers by offering opportunities to research for enhanced security [43,44]. Cloud computing systems consist of various resources, which are shared among users with the help of hypervisors. Hence, it provides an opportunity for adversaries to spread malicious traffic to erode the performance, consume more resources or stealthily access resource of other users. With the centralized and global view of network behavior and user activities, SDN provides cost-effective mechanisms to counter such threats. For example, the CloudWatcher [45], working as an SDN application, uses the SDN control platform to provide monitoring services to large and dynamic clouds. CloudWatcher provides mechanisms to control network flows, in order to guarantee their inspection through security systems. Similarly, the SnortFlow [46] uses the OpenFlow SDN model to provide intrusion detection and response systems for clouds. SnortFlow uses Snort-based Intrusion Prevention System (IPS) to detect intrusions and OpenFlow controller to generate actions for the detected flows. For data centers, the Automated Malware Quarantine (AMQ) [44] is an SDN-based solution that detects potential threats and isolates insecure network devices to stop them from adversely affecting the network. Using two modules on top of the SDN controller, such as the Bot Hunter and threat responder, the AMQ detects threats and isolates those threats using the SDN controller.

By coupling NFV with SDN, the above solutions can be used for securing any type of cloud, either in the edge or the data center. As shown in Figure 16.6, modules such as the Bot Hunter in AMQ, or the CloudWatcher, can be deployed in the network edge using the programmability offered by SDN and virtual function placement by NFV.

16.7 5GMEC, Cloud Security Research and Standardizations

The standards that will define 5G are yet to be outlined. The standardization of 5G is a multi-stakeholder process involving a huge numbers of operators, regulators, vendors, policy-makers and representatives of 5G users. However, research and standardization are currently ongoing in several technology areas related to 5G. A more detailed description of standardization activities in 5G is contained in Chapter 2 of this book. In this section, we discuss on standardization activities related to MEC and cloud computing:

- *European Telecommunications Standards Institute (ETSI)⁹:* ETSI was created in 1988 to be the main standardization organization in ICT within Europe. It comprises of several technical committees and Industry Specification Groups (ISGs). Its key partners are 3GPP¹⁰ and OneM2M¹¹. ETSI holds a crucial position in both research

⁹ ETSI. <http://www.etsi.org/>

¹⁰ 3rd Generation Partnership Project (3GPP). <http://www.3gpp.org/about-3gpp>

¹¹ One M2M. <http://www.onem2m.org/>

and standardization of technologies related to cloud computing and MEC. For instance, it was after ETSI launched the ISG for Mobile-Edge Computing in 2014 that MEC acquired its current meaning [47].

Back in 2013, ETSI delivered a report on cloud computing standards, where the Cloud Standards Coordination (CSC) initiative was launched [48], with an ambitious goal of creating 2.5 million new European jobs in the field of cloud computing by the year 2020. The goal of the CSC was to define the key roles of cloud computing, the potential use cases, standardization organizations, and the classification of respective activities for both users and service providers for the entire cloud life-cycle. Also in 2014, the ETSI MEC ISG was formed, with an aim of standardizing the MEC environment and also defining different possible service scenarios and technical requirements for MEC.

- *National Institutes of Standards and Technology (NIST)*: Founded in 1901 as a measurement standards laboratory in the United States and later became a part of the US department of defense, NIST provides measurement standards for a wide variety of technologies. NIST has also become a key player in the standardization efforts of 5G related technologies like cloud computing. For instance, earlier in July 2011, NIST CCSRWG¹² released what it called the *NIST Cloud Computing Standard Roadmap* [49]. This roadmap was designed to accelerate the secure adoption of cloud computing by the federal government through standard developments and guidelines in collaboration with other standard bodies.

The NIST Definition of Cloud Computing, identified cloud computing as:

...a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

is generally the most widely accepted definition of cloud computing. The NIST Cloud Computing Reference Architecture and Taxonomy Working Group has developed a reference architecture for high-level conceptual model of cloud computing, which are generally used in discussions regarding structures, requirements, and operations of cloud computing [49].

- *Next Generation Mobile Networks (NGMN) 5G Security Group*¹³: NGMN Alliance is a mobile telecommunications association that comprises of mobile operators, vendors, manufacturers and research institutes. This group has gained substantial recognition by both IEEE and 3GPP at different levels. NGMN Alliance announced the launch of a global initiative for 5G earlier in 2014, with the aim of spearheading the development of technologies and standards required for future communication networks [50]. The NGMN 5G security group focuses on expanding communication infrastructures through the use of integrated platforms to advance mobile services in 5G and LTE-advance communication networks. In the last quarter of 2016, the NGMN 5G security group released a comprehensive report on 5G security, MEC, low latency, and consistent user experience. This report provided more bases on how MEC and low latency

12 NIST Cloud Computing Standards Roadmap Working Group

13 NGMN. <https://www.ngmn.org/de/about-us/vision-mission.html>

combination will support varieties of new use cases and services on 5G. The other role of NGMN includes addressing spectrum requirements, establishing more transparent IPR regime, providing guidance to equipment developers and standardization bodies, establishing clear functionality and performance targets as well as providing an information exchange forum for the industry.

16.8 Conclusions

The large-scale adoption of 5G technologies does not only depend on its ability to deliver the anticipated performance and flexibility promises such as throughput, flexible RAN and latency, but more so on its ability to guarantee the security of all parties involved; users and service providers alike. With the amalgamation of myriads of technologies and use cases, emerging technologies like cloud computing and MEC would be the main targets of adversaries. In this chapter, we have presented the threat landscape for MEC and cloud computing in the context of 5G technologies. Threats such as manipulation of virtual machines, privacy leakage, DDoS, and MitM will become even more prevalent, given that 5G will rely heavily on virtualization and edge technologies. We have further proposed certain control measures and techniques that can mitigate these threats and provide highly guaranteed security on the network. This is particularly important, because in order for 5G to support the vast number of new applications and use cases that have been proposed, MEC and cloud technologies seem indispensable in any case, hence adequate security measures need to be put in place to affirm the confidence of both users and service providers in both technologies.

References

- 1 Security in cloud computing, *International Journal of Information Security*, 13(2), 95–96, 2014 [Online]. Available at: <http://dx.doi.org/10.1007/s10207-014-0232-2>
- 2 Vassilakis, V. et al. (2016) Security analysis of mobile edge computing in virtualized small cell networks. In: *Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer International Publishing, Thessaloniki, Greece.
- 3 NSN, Nokia (2013) Solutions and Networks: Looking Ahead to 5G. White paper, *Nokia Solutions and Networks Oy, Finland*.
- 4 NSN, Nokia (2013) 5G Use Cases and Requirements. White paper, *Nokia Solutions and Networks Oy, Finland*.
- 5 Wen, T. and Zhu, P. (2013) *5G: A Technology Vision*. Huawei.
- 6 Intel IT Center (2013) *Planning Guide: Virtualization and Cloud Computing Steps in the Evolution from Virtualization to Private Cloud Infrastructure as a Service*.
- 7 Bojanova, I., Zhang, J. and Voas, J. (2013) Cloud computing. *IT Professional*, 15(2), 12–14.
- 8 Mohamed, A. (2009) A history of cloud computing. *Computer Weekly*, 27.
- 9 Intel IT Center (2013) Planning guide. *Virtualization and Cloud Computing: Steps in the Evolution from Virtualization to Private Cloud Infrastructure as a Service*.

- 10 Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R. *et al.* (2010) A view of cloud computing. *Communications of the ACM*, 53(4), 50–58.
- 11 Eze Castle Integration (2013) *The History of Cloud Computing*. Boston, MA.
- 12 Mell, P. and Grance, T. (2011) *The NIST Definition of Cloud Computing*. NIST, Gaithersburg.
- 13 Jadeja, Y. and Modi, K. (2012) Cloud computing-concepts, architecture and challenges. *Proceedings of the International Conference on Computing, Electronics and Electrical Technologies (ICCEET)*, IEEE.
- 14 Savolainen, E. (2012) Cloud service models. *Seminar on Cloud Computing and Web Services*, vol. 10. University Of Helsinki, Department of Computer Science, Helsinki.
- 15 Tsai, W-T., Sun, X. and Balasooriya, J. (2010) Service-oriented cloud computing architecture. *Proceedings of the Seventh International Conference on Information Technology: New Generations (ITNG)*, IEEE
- 16 5GPPP Architecture Working Group (2016). *View on 5G Architecture*, June. Available at: <https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-5G-Architecture-WP-For-public-consultation.pdf>
- 17 Ericsson White Paper. Cloud RAN: *The Benefits of Virtualization, Centralization and Coordination*. No, 284 23-3271, September 2015.
- 18 Tullberg, H. *et al.* (2015) METIS SYSTEM CONCEPT: The shape of 5G to come. *IEEE Communications Magazine* (Online). Available at: https://www.metis2020.com/wp-content/uploads/publications/IEEE_CommMag_2015_Tullberg_etal_METIS-System-Concept.pdf
- 19 Hu, Y.C. *et al.* (2015) *Mobile Edge Computing – A Key Technology Towards 5G*. ETSI White Paper 11.
- 20 Patel, M. *et al.* (2014) *Mobile-edge Computing Introductory Technical*. White Paper, Mobile-edge Computing (MEC) industry initiative.
- 21 Mobile Edge Computing: The Edge is the Future. iGillott Research Inc. (2015). White Paper. Version 1.0. 12400 Austin TX 78738.
- 22 ETSI Group Specification (2016) *Mobile Edge Computing (MEC); Technical Requirements*, ETSI GS MEC 002 V1.1.1 (2016-03), March.
- 23 5G PPP. 5G Vision: The Next Generation of Communication Networks and Services (2015) [Online]. Available at: <https://5g-ppp.eu/wp-content/uploads/2015/02/5G-Vision-Brochure-v1.pdf#page=8>
- 24 ETSI, M. (2014) *Mobile-Edge Computing*. Introductory Technical White Paper.
- 25 Jarich, P. (2016) *Building 5G: Mobile Edge Computing*. Current Analysis Network Matter. Blog.
- 26 Tran, T.X., Hajisami, A., Pandey, P. and Pompili, D. (2016) *Collaborative Mobile Edge Computing in 5G Networks: New Paradigms, Scenarios, and Challenges* [Online]. Available at: <https://arxiv.org/pdf/1612.03184.pdf>
- 27 Embedded Intel Solutions (2015) *How Mobile Edge Computing is Helping Operators Face the Challenges of Today's Evolving Mobile Networks* [Online]. Available at: <http://eecatalog.com/intel/2015/08/17/how-mobile-edge-computing-is-helping-operators-face-the-challenges-of-todays-evolving-mobile-networks/>
- 28 Patel, M. *et al.* (2014) *Mobile-edge Computing Introductory Technical*. White Paper, Mobile-edge Computing (MEC) industry initiative.

- 29 Hart, J. (2016) What mobile Core Network Architecture Concepts will have an impact on 5G Evolution? *BT TSO – Research & Technology*. 5G NORMA Summer School, London KCL, June.
- 30 Lindstrom, P. *The Laws of Virtualization Security*. Baselinemag.com Driving Business Success with Technology [Online]. Available at: <http://www.baselinemag.com/c/a/Security/The-Laws-of-Virtualization-Security>
- 31 Puttonen, J., Afolaranmi, S.O., Moctezuma, L.G., Lobov, A. and Martinez Lastra, J.L. (2015) *Conference on Security in Cloud-Based Cyber-Physical Systems*, pp. 671–676. DOI:10.1109/3PGCIC.2015.30
- 32 Chonka, A. and Abawajy, J. (2012) Detecting and mitigating HX-DoS attacks against cloud web services. *Proceedings of the 15th International Conference on Network-Based Information Systems (NBiS)*, September, pp. 429–434.
- 33 Ficco, M. and Rak, M. (2015) Stealthy denial of service strategy in cloud computing. *IEEE Transactions on Cloud Computing*, 3(1), 80–94.
- 34 Hubbard, D. and Sutton, M. (2010) *Top Threats to Cloud Computing v1. 0*. Cloud Security Alliance.
- 35 Cox, P. *Intrusion Detection in a Cloud Computing Environment* [Online]. Available at: <http://searchcloudcomputing.techtarget.com/tip/Intrusion-detection-in-a-cloud-computing-environment>
- 36 Khan, A.R. (2012) Access control in cloud computing environment. *ARP Journal of Engineering and Applied Sciences* 7(5), 613–615.
- 37 Younis, Y.A., Kashif, K. and Madjid, M. (2014) An access control model for cloud computing. *Journal of Information Security and Applications*, 19(1), 45–60.
- 38 Roman, R., Javier, L. and Masahiro, M. (2016) *Mobile Edge Computing, Fog – A Survey and Analysis of Security Threats and Challenges*. Future Generation Computer Systems.
- 39 Liang, B. (2016) *Mobile Edge Computing*. University of Toronto, Canada. [Online]. Available at: http://paswkshop.comm.utoronto.ca/~liang/publications/Chapter_MEC_2016.pdf
- 40 Stojmenovic, I. et al. (2015) An overview of fog computing and its security issues. *Concurrency and Computation: Practice and Experience*, 10, 2991–3005.
- 41 Zhang, L. et al. (2010) A man-in-the-middle attack on 3G-WLAN interworking, vol. 1. *Proceedings of the International Conference on Communications and Mobile Computing (CMC)*, IEEE.
- 42 Bahram, S. et al. (2010) DKSM: Subverting virtual machine introspection for fun and profit. *Proceedings of the 29th IEEE Symposium on Reliable Distributed Systems*, IEEE.
- 43 Ahmad, I., Namal, S., Ylianttila, M. and Gurto, A. (2015) Security in software defined networks: a survey. *IEEE Communications Surveys & Tutorials*, 17(4), 2317–2346.
- 44 ONF, *SDN Security Considerations in the Data Center*. Open Networking Foundation, Palo Alto, CA [Online]. Available at: <https://www.opennetworking.org/images/stories/downloads/sdn-resources/solution-briefs/sb-security-data-center.pdf>
- 45 Shin, S. and Gu, G. (2012) CloudWatcher: Network security monitoring using OpenFlow in dynamic cloud networks (or: How to provide security monitoring as a service in clouds?), *Proceedings of the 20th IEEE ICNP*, October, pp. 1–6.
- 46 Xing, T., Huang, D., Xu, L., Chung, C.-J. and Khatkar, P. (2013) SnortFlow: A OpenFlow-based intrusion prevention system in cloud environment, *Proceedings of the 2nd GREE*, March, pp. 89–92.

- 47 ETSI (2014) *Mobile-Edge Computing Introductory Technical White Paper* [Online]. Available at: <http://www.etsi.org/technologies-clusters/technologies/mobile-edge-computing> (accessed 9 February 2017).
- 48 ETSI (2013) *Cloud Standards Coordination: Final Report v 1.0* [Online]. Available at: http://www.etsi.org/images/files/Events/2013/2013_CSC_Delivery_WS/CSC-Final-report-013-CSC_Final_report_v1_0_PDF_format-.PDF
- 49 Hogan, M. *et al.* (2011) *NIST Cloud Computing Standards Roadmap*. NIST Special Publication 35.
- 50 Zhang, N. *et al.* (2015) Cloud assisted HetNets toward 5G wireless networks. *IEEE Communications Magazine*, 53(6), 59–65.

17

Regulatory Impact on 5G Security and Privacy

Jukka Salo¹ and Madhusanka Liyanage²

¹ Nokia Networks, Espoo, Finland

² Centre for Wireless Communication, University of Oulu, Oulu, Finland

17.1 Introduction

The development path of any industry or economic sector is significantly affected by the opportunities provided by the available technologies, the particular characteristics of its markets and the directions and priorities of related government policies and regulations. These factors can be mutually supportive in stimulating growth and creating benefits, or they can conflict with one another, creating major blockages to development. Potential opportunities for development in the sector will arise from the interrelations among technologies, markets and policies. This will be true also with respect to the 5G telecommunication networks; especially so, because they will play a ubiquitous and pervasive role in the daily life of people.

The future 5G networks are a part of the critical information infrastructure, and it has to be ensured that they meet the requirements set for such infrastructures. It is clear that Security and Privacy will be of utmost importance for the critical infrastructures, and certainly they are among the key aspects to pay attention to when designing the new mobile network concepts and exploiting the related new technologies.

The concept of Privacy is one of the fundamental motivations for security [1]. Privacy is commonly understood as the right of individuals to control what information related to them may be collected and stored and by whom and to whom that information may be disclosed. By extension, Privacy is also associated with certain technical means (e.g. cryptography) to ensure that this information is not disclosed to any other than the intended parties, so that only the explicitly authorized parties can interpret the content exchanged among them.

Most commonly, privacy and confidentiality are used as the same term, but it should be noted that differentiation between Privacy and data confidentiality exists [2], the former relating to the protection of the association of the identity of users and the activities

performed by them (i.e. online purchase habits), and the latter relating to the protection against unauthorized access to data content. Encryption, access control lists, and file permissions are methods often used to provide data confidentiality.

The European Commission says in its press release related to the Telecoms Council, Brussels, 31 March 2009, that:

ICT systems, services, networks and infrastructures form a vital part of European economy and society, either by providing essential goods and services or by constituting the underpinning platform of other critical infrastructures. They are often called Critical Information Infrastructures as their disruption or destruction would have a serious impact on vital societal functions [3].

Therefore, Security and Privacy is a key discussion topic in the context of the regulation related to the 5G security.

The implications of the new cellular core network technologies on the Security and Privacy regulation was studied following the procedure that is also illustrated in Figure 17.1:

- 1) At first, the existing regulatory environment, and regulatory goals and regimes in Security and Privacy are presented to better understand the environment, where these topics are discussed today (Sections 17.2 and 17.3);
- 2) Then, the potential Security and Privacy issues in the context of new technologies will be identified (Sections 17.4.1–17.4.4);
- 3) Third, the relevance assessment of the identified Security and Privacy issues from the perspectives of regulatory goals will be made (Sections 17.5);
- 4) Then, the analysis of the potential regulatory approaches for solving the Security and Privacy issues will be made using the regulatory goals as the criteria (Sections 17.6);
- 5) In the final step, the direct impact of the new technologies on the Security and Privacy Regulation is assessed (Section 17.7).

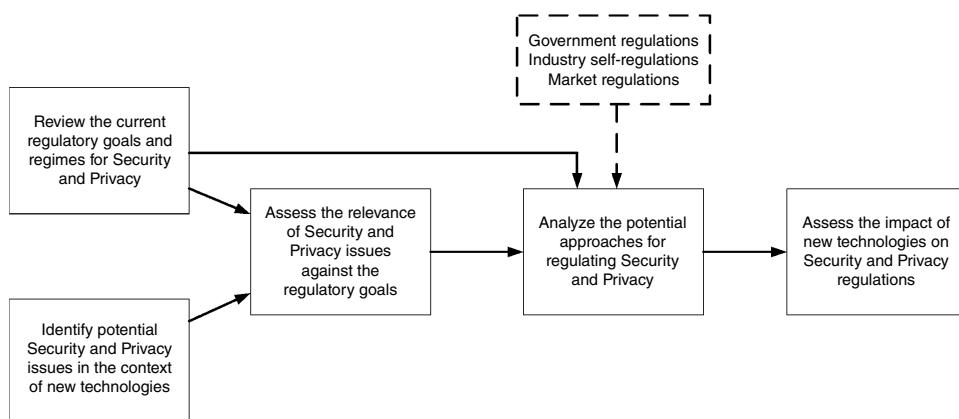


Figure 17.1 Approach to study the direct impact of new technologies on Security and Privacy Regulation.

17.2 Regulatory Objectives for Security and Privacy

It is clear that ensuring Security and Privacy will be the key objectives when designing the new information sharing concepts of the future 5G networks. To which level the new network concepts supports Security and Privacy is an issue that can be determined by regulators when setting the policies for operations.

To fulfil the regulatory requirements, it will not be enough to tackle it only with the technical security mechanisms; there will also be a need for laws that prohibit certain actions that would be impossible to find technical solutions to enforce.

In her speech in March 2014, Commissar Neelie Kroes said [4]:

Businesses working across the single market must sign 28 different telecoms contracts with 28 separate suppliers. The new services they use have to cope with all those different systems. No guarantee of quality, security, service. Not much good if you're considering a multi-million-euro investment in the cloud. Meanwhile networks and systems are insecure and unprotected. Deutsche Telekom reports 800 000 attacks every day – and that is just one company. Countries and companies that keep threats to themselves make the whole chain vulnerable. Companies who think this isn't happening to them are very naïve, or lying.

Today, the following target for a regulatory intervention is commonly accepted: Increasing/ensuring consumer welfare, which is ultimately determined by the structure and level of the retail prices paid by end consumers. The preconditions for reaching this target are the economic efficiency, rapid innovation, and efficient and timely investment.

17.2.1 Generic Objectives

In this section, the list of regulatory objectives for Security and Privacy in the context of the new mobile networks technologies has been created based on the Generic Regulatory Objectives [5–7] and on the Regulatory Objectives in Cloud Computing [8–10]. They are seen to be common for all new technologies under the scope of the 5G networks. These objectives are:

- 1) Promote the Digital Single Market to encourage efficient cross-border services; the harmonized implementation of all relevant Directives and legislative instruments are needed in the EU and in the global context;
- 2) Balance of interests in protecting privacy and in fostering the EU-wide and global use of services; Europe to fully realize the benefits of new technologies; Note: the current laws may discourage non-European users from using EU-based cloud computing providers or making use of European data centres, for instance;
- 3) Security and Privacy legislation has to be looked at in a global context and its compatibility with new technologies has to be ensured; For instance, Cloud Computing has to be facilitated in Europe and at a global level; Different jurisdictions/regions shall cooperate to develop interoperable requirements that facilitate information flows with appropriate Security and Privacy protection;

- 4) Foster interoperability and data portability; endorse technology neutrality and promote competition; avoid mandated standards or preferences that could frustrate, rather than promote, on-going interoperability efforts of the industry at large and among the vendors providing services and solutions;
- 5) The applicable law must be easy to define; A single set of rules on data protection, valid across the EU, shall be set up; A legal framework is needed that can be applied across borders, which gives users the means to exercise their rights across borders, which is based on the concept of accountability and draws on technological controls and self-regulatory codes and mechanisms as supported by Articles 17 and 27 of the Directive 95/46/EC;
- 6) The right to be forgotten, i.e. the right for the individual to request the deletion of his/her personal data;
- 7) Increased responsibility and accountability for those processing personal data.

17.3 Legal Framework for Security and Privacy

In the following sub-sections, the details of the legal framework for Security and Privacy, the Security and Privacy issues in the context of new network technologies, and their relevance assessment against the specific regulatory objectives, are presented.

17.3.1 General Framework

The networks today are in general more open than in the past and one weak link affects the integrity of the whole system. The growth of spam, viruses, spyware and other forms of malware, which is undermining users' confidence in electronic communications, is partly due to that openness. To ensure the security of these critical infrastructures and to protect the citizens' privacy, the EU has taken several measures for ensuring the security of these critical infrastructures and to protect the privacy of its citizens.

In the EU's: 1) Privacy Directive (EC Directive 2002/58/EC) [11]; and 2) Data Protection Directive (Directive 95/46/EC) [12], Privacy in the processing of personal data and the confidentiality of communications are recognized as fundamental rights that should be protected:

- 1) The Privacy Directive requires the Member States to harmonize and ensure an equivalent level of protection of the right to privacy with respect to personal data in the electronic communication sector. Regarding the confidentiality of communications, the Privacy Directive says that EU member states shall ensure the confidentiality of communications and the related data traffic through the national legislation. In particular, they shall prohibit listening, tapping, storage or other kinds of interception or surveillance of communications and the related traffic data by persons other than users, without the consent of the users concerned.
- 2) The Data Protection Directive prohibits the transfer of personal information to any country that does not have adequate privacy laws. As a result, EU member states

have implemented legislation that prohibits the transfer of personal information from the EU to third countries, unless such countries have adequate privacy protection in their laws [1].

On 25 January 2012, the European Commission proposed a comprehensive reform of the EU data protection rules. The new European Data Protection Regulation is meant to supersede the EU Data Protection Directive from 1995. According to the EC, the new rules will strengthen online privacy rights and boost Europe's digital economy. The reform of the outdated privacy rules reflects that technological progress and globalization have profoundly changed the way data are collected, accessed and used [13].

The European Parliament, having already voted in favor of the General Data Protection Regulation ("GDPR") [14], voted on 14 March 2014, on the proposed Network and Information Security ("NIS") Directive [15]. In line with the previous committee reports, the Parliament vote ensures that the Proposed Network and Information Security Directive focuses on protecting critical infrastructure in the energy, transport, financial services and health sectors. The EU legislative bodies will now enter into negotiations to agree a final text [16].

The Commission proposed the NIS Directive in February 2013. In addition to provisions aimed at Member State governments (e.g. to improve cyber security capabilities and cooperation to prevent and respond to cyber-attacks), the Directive targets private companies in the energy, transport, financial services and health sectors. These sectors are seen to be dependent on the correctly functioning network and information systems. The Commission draft also applied to "enablers of key internet services", such as providers of cloud computing services, app stores, e-commerce platforms, internet payment gateways, search engines and social networks.

The policy options for ensuring NIS have been assessed in the Impact Assessment of the NIS Directive [17]. Three policy options have been named as:

- 1) Option 1 – Business as usual;
- 2) Option 2 – Regulatory approach; and
- 3) Option 3 – Mixed approach.

The assessment covers, in addition to the level of security, the economic and social impacts of the three options.

For comparison, the United States does not provide adequate privacy protection from the European point of view [18]. To address this problem, the European Commission and the US Department of Commerce negotiated the Safe Harbor agreement, which is only applicable to transfers between the USA and the EU. Organizations outside the USA that have business operations within the EU, have to rely on different mechanisms to adhere to the Transborder Transfer principle from Directive 95/46/EC. This principle requires that personal identifiable information can only be transferred to those countries that are deemed to provide adequate security.

17.3.2 Legal Framework for Security and Privacy in Cloud Computing

In Europe, the processing of personal data is mainly regulated by the Data Protection Directive 95/46/EC [12], which is currently under revision. The Directive imposes stringent duties and obligations on the actors of such processing, mainly on the

“Controller”¹ but also on the “Processor”². The facts that personal data can be rapidly transferred by the Cloud Service Providers (CSPs) from one data centre to another and that the customer usually has no control or knowledge over the exact location of the provided resources (the “location independence” concept described in the article *Cloud Computing Legal Issues: An Overview (Part 1/2)*), stimulate customers’ concerns on data protection and data security compliance [19].

Article 4 of the Data Protection Directive [12] requires the Member States to apply the data protection rules to controllers who process personal data in the “context of the activities” of their EEA (European Economic Area) “establishment”, or who are not “established” in the EEA but, for purposes of processing personal data, “makes use of” equipment” (or “means”) situated in the EEA [20]. However, the application of article 4 to Cloud Computing is complicated by the fact that many cloud computing service providers do not own the data centres or equipment they use, and may well use the resources of other clouds. Those other Cloud Service Providers in turn may ultimately use data centres and servers rented by third parties. This means that the cloud users do not necessarily know in which data centres, or even countries, their data are stored or where their processing operations are run.

In addition, the data protection laws may differ between EU member states. There are also practical issues relating to whether the Directive can be enforced in non-EU countries. Clarification is therefore needed in the updated Directive on which country’s security requirements and other rules apply to a Cloud Computing user or provider [21].

The Governance models and processes need also to take into account the specific issues arising from the inherently global nature of the Clouds. Data is subject to specific legislative requirements that may depend on the location where they are hosted, and for what purposes they are processed. Different countries have different laws regarding which kind of data may be hosted and where and how it is to be protected. Within the Cloud, data/code may be hosted anywhere within the distributed infrastructure, that is, potentially anywhere in the world [8].

Clarification of applicable law governing the flow, processing and protection of data are desirable, so that both Cloud customers and Cloud Service Providers have a clear understanding about which rules apply where and how. While there is no question that the Privacy Directive, like other EU Directives, applies to Cloud services, questions do arise as to how and to what extent they apply (geographic and potential subject-matter limits), as well as how they *should* apply to maximize the potential benefits of those services, while still providing the appropriate level of personal data protection [8].

¹ Controller means the natural or legal person, public authority, agency or any other body, which alone or jointly with others determines the purposes and means of the processing of personal data. *Processing of personal data (Processing)* means any operation or set of operations, which are performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction.

² Processor means a natural or legal person, public authority, agency or any other body, which processes personal data on behalf of the controller.

17.3.3 Legal Framework for Security and Privacy in Software Defined Networking and Network Function Virtualization

The legal framework for Security and Privacy, as defined in Section 17.3.1, is assumed to be applied also for Software Defined Networking and Network Functions Virtualization. Also, the rules and policies which are being specified for Security and Privacy in Clouds are assumed to be applied for Network Functions Virtualization, because of the strong interrelation of those two technologies.

17.4 Security and Privacy Issues in New 5G Technologies

The following sub-sections deal with different Security and Privacy issues in 5G network context, that is, Cloud Computing, Network Functions Virtualization (NFV), and Software Defined Networking (SDN).

17.4.1 Security and Privacy Issues in Cloud Computing

The ITU focus group on Cloud Computing³ has identified the following security issues with respect to the cloud service users [22]:

- *Responsibility ambiguity*: the lack of a clear definition of responsibility among cloud service users and Providers may evoke conceptual conflicts. Also, the problem of which entity is the data controller stays open at an international scale;
- *Loss of governance*: for an enterprise, migrating a part of its own IT system to a cloud infrastructure implies to partially give control to the cloud service providers;
- *Loss of trust*: it is sometimes difficult for a cloud service user to recognize his Provider's trust level, due to the black-box feature of the cloud service. There is no measure of how to get and share the Provider's security level in a formalized manner. Furthermore, the cloud service users have no abilities to evaluate security implementation level achieved by the Provider;
- *Service Provider lock-in*: a consequence of the loss of governance could be the lack of freedom regarding how to replace a cloud provider with another;
- *Non-secure cloud service user access*: as most of the resource deliveries are through a remote connection, i.e. non-protected APIs, services are one of the easiest attack vectors. Attack methods such as phishing, fraud and exploitation of software vulnerabilities still achieve results;
- *Lack of information/asset management*: when applying to use cloud computing services, the cloud service user will have serious concerns on lack of information/asset management by Cloud Service Providers, such as the location of sensitive asset/information, the lack of physical control for data storage, and the reliability of data backup;
- *Data loss and leakage*: the loss of the encryption key or privileged access code will bring serious problems to the cloud service users.

³ The ITU-T Focus Group on Cloud Computing (FG Cloud) was established further to the ITU-T TSAG agreement at its meeting in Geneva, 8–11 February 2010, followed by the ITU-T study groups and membership consultation. It was successfully concluded in December 2011.

With respect to the Cloud Service Providers (CSP), the following security issues were identified by the Focus Group [22]:

- *Ambiguity in responsibility*: different user roles, such as the Cloud Service Provider, cloud service user, client IT admin and data owner, may be defined and used in a cloud system. Ambiguity of such user roles and the responsibilities definition related to the data ownership, access control, infrastructure maintenance, etc, may induce business or legal dissension;
- *Protection inconsistency*: due to the decentralized architecture of a cloud infrastructure, its protection mechanisms are likely to be inconsistent among distributed security modules;
- *Bylaw conflict*: depending on the bylaws of the hosting country, data may be protected by different applicable jurisdictions. For instance, the USA Patriot Act may authorize such seizures. The EU protects cloud service user's private data, which should not be processed in countries that do not provide a sufficient level of guaranteed protection. An international Cloud Service Provider may conflict with the bylaws of its local data centres, which is a legal threat to be taken into account;
- *Business discontinuity*: the "as a service" feature of cloud computing allocates resources and delivers them as a service. The whole cloud infrastructure, together with its business workflows, thus relies on a large set of services, ranging from hardware to application. However, the discontinuity of service delivery, such as a black-out or delay, may have a severe impact on the availability;
- *Shared environment*: cloud resources are virtualized and different cloud service users (possibly competitors) share the same infrastructure. Any unauthorized and violent access to cloud service user's sensitive data may compromise both the integrity and confidentiality;
- *Hypervisor isolation failure*: the hypervisor technology is considered as the basis of cloud infrastructure. Multiple virtual machines co-hosted on one physical server share both CPU and memory resources, which are virtualized by the hypervisor. This threat covers the failure of mechanisms to isolate attacks that could be launched on a hypervisor to gain illegal access to the memory of other virtual machines,
- *Service unavailability*: availability is not specific to the cloud environment. However, because of the service-oriented design principle, service delivery may be impacted while the cloud infrastructure is not available. Moreover, the dynamic dependency of cloud computing offers many more possibilities to an attacker. A typical denial of service attack on one service may blog the whole cloud system;
- *Abuse by Cloud Service Provider*: for a cloud service user, migrating a part of its own IT to a cloud infrastructure, implies to partially give control to the Provider. This may lead to a misconfiguration or malicious insider attack.

The Security and Privacy issues reported on NIST guidelines on Security and Privacy in Public Cloud Computing [23] are very much the same or similar to those listed above. The following first two issue descriptions can be seen to further explain the issues above. The third issue (Visibility) is complementary to the issues above:

- *Loss of control*: (see Loss of governance above). Transitioning to Cloud's architecture requires a transfer of responsibility and control to the cloud provider, over information as well as system components that were previously under the organization's

direct control. The transition is usually accompanied by the lack of a direct point of contact with the management of operations and influence over decisions made about the computing environment. This situation makes the organization dependent on the cooperation of the Cloud Service Provider to carry out activities that span the responsibilities of both parties, such as continuous monitoring and incident response;

- *Data Ownership:* (see Ambiguity responsibility above). The organization's ownership rights over the data must be firmly established in the service contract to enable a basis for the trust and privacy of data. The continuing controversy over privacy and data ownership rights for social networking users illustrates the impact that ambiguous terms can have on the parties involved. Ideally, the contract should clearly state that the organization retains exclusive ownership over all its data; that the cloud provider acquires no rights or licenses through the agreement, including intellectual property rights or licenses, to use the organization's data for its own purposes; and that the Cloud Service Provider does not acquire and may not claim any interest in the data due to security;
- *Visibility:* Knowledge of a Cloud Service Provider's security measures is also needed for an organization to conduct its risk management. For example, the process of identifying vulnerabilities should include an analysis of the system security features and the security controls used to protect the cloud environment. The Cloud Service Providers can be reluctant to provide details of their Security and Privacy measures and status; however, since such information is often considered proprietary and might otherwise be used to devise an avenue of attack.

Transparency in the way a Cloud Service Provider operates is a vital ingredient for effective oversight over system Security and Privacy by an organization. To ensure that policy and procedures are being enforced throughout the system's lifecycle, the service arrangements should include some means for the organization to gain visibility into the security controls and processes employed by the cloud provider and their performance over time. For example, the service agreement could include the right to audit controls via a third party.

17.4.2 Security and Privacy Issues in Network Functions Virtualization

It has to be noted here that the concept Network Virtualization is different from the Network Functions Virtualization (NFV). Network Virtualization, or a virtualized network, uses a single physical infrastructure to support multiple logical networks. Each logical network provides its users with a custom set of protocols and functionalities. An important aspect of Network Virtualization is that the three participating entities – network infrastructure providers, virtual network operators, and users – are independent and driven by different objectives. Thus, it cannot be assumed that they always cooperate to ensure that all aspects of the virtual network operate correctly and securely. Instead, each entity may behave in a non-cooperative or malicious way to gain benefits [34]. Security issues in virtualized network architectures impose significant challenges and require effective solutions. The problem of hosting network protocols and services on third-party infrastructures raises serious questions on the trustworthiness of the participating entities.

Physical network functions assume a tight coupling of the Network Functions software and hardware which, in most cases, is provided by a single vendor. In the Network Functions Virtualization (NFV) scenario, multiple vendors are expected to be involved in the delivery and setup of different virtualized elements (e.g. hardware resources, virtualization layer, virtualized network functions (VNF), virtualized infrastructure manager, etc.). As a result, due to the virtualization process, new security issues need to be addressed [24]. Examples are:

- The use of hypervisors may introduce additional security vulnerabilities. In general, to reduce the vulnerabilities of hypervisors in use, it is essential to follow the best practices on hardening and patch management;
- The usage of shared storage and shared networking may also add additional dimensions of vulnerability;
- The interconnectivity among the virtualized end-to-end architectural components exposes new interfaces that, unless protected, can create new security threats;
- The execution of diverse VNFs over the NFV infrastructure can also create additional security issues, if VNFs are not properly isolated from others.

According to Alcatel-Lucent's report on network virtualization [25], NFV will introduce several new security challenges:

- Due to the dispersion of virtual machines (VMs) that belong to a VNF across racks and datacenters, and due to the migration of VMs for the optimization or maintenance purposes, the physical perimeters of the network functions become blurred and "fluid", making it practically impossible to manually define and manage security zones;
- The introduction of hypervisors creates new attack surfaces that could result in, among other things, compromised isolation between VMs;
- Hardware, hypervisors, VNFs and cloud resource control solutions may be provided by different vendors, increasing the risk of security holes due to mismatched assumptions and expectations.

According to ETSI NFV ISG report on virtualization requirements [26], the NFV framework shall implement appropriate security counter measures to address:

- security vulnerabilities introduced by the virtualization layer;
- protection of data stored on the shared storage resources or transmitted via shared network resources;
- protection of new interfaces exposed by the interconnectivity among virtualized end-to-end architectural components, e.g. hardware resources, VNFs and management systems;
- isolation of distinct VNF sets executing over the NFV infrastructure to ensure security and separation between these VNF sets;
- secure management of VNF sets by other third-party entities (e.g. VNPaaS, enterprise virtual CPE, and virtual consumer home gateways).

The NFV Infrastructure shall be able to use standard security mechanisms wherever applicable to authentication, authorization, encryption and validation [26].

17.4.3 Security and Privacy Issues in Software Defined Networking (SDN)

Software-Defined Networking (SDN) provides a centralized intelligence and control model that is well suited to offer much-needed flexibility for network security deployments. Along with many benefits, SDN also poses new threats, particularly with the emergence of cloud and virtualized environments [27]. The central control of the SDN architecture could give an attacker the command over the entire network [28].

OpenFlow-enabled SDN offers a wide range of benefits for security implementation and management [27]:

- fine-grained enforcement and control of multiple simultaneous security policies throughout the data center;
- rapid response to threats, with the ability to rapidly steer or guarantee flows and VMs based on real-time network conditions;
- validation of security policies, and quick identification and resolution of any policy conflicts that may arise;
- efficient authentication of the flow rule producers through the use of digital signatures;
- incorporation of a trust model with live rule-conflict detection and resolution at the controller layer;
- synchronization of distributed policy insertion and removal;
- optimization of secure flow routing in a highly dynamic environment;
- dynamic assertion of extensions to the security policy when new threats are detected;
- provision of a mechanism for auditing and audit trails.

The separation of the different planes and aggregating the control functionality to a centralized system also opens up new challenges. The communication channels between the isolated planes can be targeted to masquerade one plane for attacking the other. The control plane can become a single point of failure and render the whole network down in the case of compromise. The malfunctioning or malicious software can compromise the whole network, having access granted to the control plane [29].

Basically, the security issues in SDN are concentrated around the main areas of:

- 1) application plane;
- 2) control plane;
- 3) data plane; and
- 4) communication security.

SDN enables applications to interact with and manipulate the behavior of network elements through the control layer. SDN has two properties which can be seen as attractive to malicious users: i) the ability to control the network by software; and ii) centralization of network intelligence in network controllers. Since, there are no standards or open specifications to facilitate open APIs for applications to control the network services and functions through the control plane, applications can pose serious security threats to the network resources, services and functions [32].

The Open Networking Foundation (ONF) organization has launched a study to determine how to make SDN more secure. The ONF is considering, for example, the idea of using distributed protocols, which are more resilient and harder to attack, simply because they are not concentrated [30].

Security needs to be everywhere within SDN. It needs to be built into the architecture, as well as delivered as a service to protect the availability, integrity and privacy of all connected resources and information [31].

17.4.4 Summary of Security and Privacy Issues in the Context of Technologies under Study (Clouds, NFV, SDN)

The future networks are a part of the critical information infrastructure, and it has to be ensured that they meet the requirements set for such infrastructures. It is clear that Security and Privacy will be of utmost importance for the critical infrastructures, and certainly they are among the key aspects to pay attention to when designing the new mobile network concepts and exploiting the related new technologies.

From the previous sections, we can conclude that the new mobile network concepts will open up numerous new Security and Privacy challenges. Those challenges, or issues, which are assumed to be the most relevant from the regulatory actions point of view, have been listed here for the further analysis. The focus is on the issues that can be resolved by the mutual agreements by the involved parties, or the intervention of the Regulatory Authority. Such issues, which are purely technical in nature, have been left out of the discussion.

Many of the issues listed here originate from the issues identified in the context of the Clouds concepts. However, they are also relevant to other concepts, as the implementation of those concepts are very much exploiting the clouds technologies.

Security and Privacy issues, which are assumed to be the most relevant in the context of regulation work in 5G networks, are listed below. Their importance against the regulatory objectives is further assessed in Section 17.4.4:

- A) *Responsibility ambiguity*: Different user roles, such as Cloud Service Provider, cloud service user, client IT admin and data owner, may be defined. Ambiguity of such roles and responsibilities may induce business or legal dissension.

Very similar to Responsibility ambiguity is Data ownership. The organization's ownership rights over the data must be firmly established in the service contract to enable a basis for the trust and privacy of data. The Cloud Service Provider shall not be able to acquire and may not claim any interest in the data due to security.

- B) *Bylaw conflict/Location of legal disputes*: Depending on the bylaw of the hosting country, data may be protected by different applicable jurisdiction. There are at least three possible locations to choose from: that of the victim; that of the offender; or that of the Service Provider:

- 1) Location of sensitive asset/information;
- 2) lack of physical control for data storage; and
- 3) reliability of data backup.

- C) *Shared environment*: The resources are virtualized and different Cloud Service Users (including MVNOs) – possibly competitors – share the same infrastructure. Any unauthorized and violent access to sensitive data may compromise both the integrity and confidentiality.

- D) *Different objectives for Trust:* The participating entities – network infrastructure providers, MVNOs, MVNEs and CSPs – are independent and driven by different objectives. They may not cooperate to ensure that all aspects of the network operate correctly and securely.
- E) *Interconnectivity:* The interconnectivity among new architectural components exposes new interfaces that, unless protected, can create new security threats. In the technology development phase, the standardization of protocols and network equipments should be considered.
- F) *Single point of failure:* The control plane in SDN can become a single point of failure and render the whole network down in case of compromise. The issue becomes even more complicated when the Controller is located in a CSP-operated Cloud.
- G) *Loss of governance:* For an enterprise, migrating a part of its own IT system to a cloud infrastructure, implies to partially give control to the Cloud Service Providers. This transition is accompanied by the lack of direct point of contact with respect to the management operations. The situation makes an organization dependent on the cooperation of the Cloud Service Provider to carry out activities that span the responsibilities of both parties.
- Similar to Loss of governance is Loss of control. Transitioning to Clouds architecture requires a transfer of responsibility and control to the Cloud Service Provider over information as well as system components that were previously under the organization's direct control. This situation makes the organization dependent on the cooperation of the Cloud Service Provider.
- H) *Service Provider lock-in:* A consequence of the loss of governance could be the lack of freedom regarding how to replace a cloud provider by another.
- I) *Visibility:* Knowledge of a Cloud Service Provider's security measures is also needed for an organization to conduct its risk management. The Cloud Service Providers can be reluctant to provide details of their Security and Privacy measures and status. Transparency in the way a Cloud Service Provider operates is a vital ingredient for effective oversight over system security and privacy by an organization.
- J) *Protection inconsistency:* Due to the decentralized architecture of a cloud infrastructure, its protection mechanisms are likely to be inconsistent among distributed security modules.

17.5 Relevance Assessment of Security and Privacy Issues for Regulation

Developing new technologies raises new issues and some of these issues require regulation as a solution. The new network technologies being studied in 5G Networks (Clouds, NFV, SDN) are no different in this aspect. In this section, the relevance of the issues described and summarized in Section 17.4.4 has been assessed against the regulatory objectives for Security and Privacy. Those objectives have been summarized in Section 17.2, and they are considered to be relevant for the technologies under the scope of 5G Networks.

The different regulatory objectives are influenced by the regulatory issues of high relevance, as summarized in the Table 17.1 below:

Table 17.1 Summary of regulatory targets with Security and Privacy issues of high impact.

| Regulatory objectives | Issues of high relevance and impact |
|-------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1) Promote Digital Single Market | Responsibility ambiguity/Data ownership, By law conflict/Location of legal disputes, Different objectives for Trust, Interconnectivity, Loss of governance/Loss of control, Service Provider lock-in, Visibility |
| 2) Balance of interest | By law conflict/Location of legal disputes, Shared environment, Interconnectivity, Loss of governance/Loss of control |
| 3) Global context | Responsibility ambiguity/Data ownership, By law conflict/Location of legal disputes, Shared environment, Different objectives for Trust, Interconnectivity, Visibility |
| 4) Foster interoperability and data portability | Shared environment, Interconnectivity, Loss of governance/Loss of control, Service Provider lock-in |
| 5) Applicable law must be easy to define | Responsibility ambiguity/Data ownership, By law conflict/Location of legal disputes, Shared environment, Different objectives for Trust, Interconnectivity, Loss of governance/Loss of control, Visibility |
| 6) Right to be forgotten | Responsibility ambiguity/Data ownership, By law conflict/Location of legal disputes, Different objectives for Trust, Loss of governance/Loss of control |
| 7) Increased responsibility and accountability | Responsibility ambiguity/Data ownership, By law conflict/Location of legal disputes, Shared environment, Loss of governance/Loss of control, Service Provider lock-in, Visibility |

17.6 Analysis of Potential Regulatory Approaches

There are at least three levels at which Security and Privacy could be regulated, each with benefits and drawbacks [32]:

- Government regulation;
- Industry self-regulation; and
- Consumer or market regulation.

The most obvious place to regulate Security and Privacy is at the governmental level. The governments are responsible for writing laws and regulations, and people look to their governments to lay down such rules that prevent harms to the public.

Another level at which to regulate privacy is at the industrial level. Industries can develop principles and practices that reflect consensus on the best approach to privacy. In “industry self-regulation”, a network of leading companies may require their business partners to meet industry standards on privacy.

Finally, there is consumer or market regulation. Consumers are in the best position to know their desires with respect to privacy, and they are in the best position to enforce the terms of their desires through their choices in the marketplace.

The different approaches have been analysed from the regulatory objectives' perspective in Table 17.2. The objectives were introduced in Section 17.2. The most relevant Security and Privacy issues to keep in mind in this analysis have been elaborated on in Sections 17.4.4.

The table summarizes how the different regulatory approaches would contribute to the different regulatory objectives. The statement "Yes" in the table means that there is positive impact on the regulatory objective in question, and the statement "No" means that there is no impact.

The Government regulation would promote many of those targets and Industry self-regulation, also many of them, at least slightly. Clearly, the Consumer or market regulation would have the most difficulties in promoting any of the targets; in this approach, the Service Providers, which offer such a level of Security and Privacy that pleases consumers and corporate users, would succeed, but most of the issues would remain unresolved.

17.7 Summary of Issues and Impact of New Technologies on Security and Privacy Regulation

Several Security and Privacy issues related to the new network technologies have been identified both on the Service Provider side and on the customer side. These issues may be complicated, especially arising from the inherently global nature of the Clouds. The wide-scale deployment of Cloud Computing, Network Function Virtualization and Software Defined Networking can trigger a number of security and data protection risks, stemming mainly from the new interfaces, shared environments, new players with different views and objectives on Security and Privacy, and from the more complicated value networks.

Data is subject to specific legislative requirements that may depend on the location where they are hosted, and for what purposes they are processed. Different countries have different laws regarding which kind of data may be hosted where, and how it is to be protected. Clarification of applicable law governing the flow, processing and protection of data are desirable, so that both the Service Providers and customers (private and corporate) have a clear understanding about which rules apply where and how [33].

On 14 March 2014, the European Parliament voted on the proposed Network and Information Security (NIS) Directive [15]. In addition to provisions aimed at Member State governments (e.g. to improve cyber security capabilities and cooperation to prevent and respond to cyber-attacks), the Directive targets private companies in the energy, transport, financial services and health sectors. These sectors are seen to be dependent on the correctly functioning network and information systems.

The policy options for ensuring NIS have been assessed in the Impact Assessment of the NIS Directive (Section 17.4.2). The three policy options identified are:

- 1) Business as usual;
- 2) Regulatory approach; and
- 3) Mixed approach.

Table 17.2 Regulatory approaches for Security and Privacy.

| Criteria (Regulatory objectives) | Government regulation | Industry self-regulation | Consumer or market regulation |
|--------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1) Promote the Digital Single Market | <p>Yes. The responsibilities have to be defined in the same way across country borders.</p> <p>Yes. Agreements between governments are needed to push the European-wide standards and practices. This is especially important for enabling new entries in the market.</p> <p>Yes. Citizens expect that governments lay down rules that prevent harms to the public.</p> | <p>Yes. The traditional telecom payers can reach consensus, e.g. on the standards for interoperability. Standardization work is already ongoing.</p> <p>No. New players are emerging in the telecoms business. They do not have such experience/ tradition on co-operation as the traditional telecom vendors do have.</p> <p>No. Industrial players cannot agree, e.g. on the single set of rules for managing security and privacy across different regions. The different players may be driven by different objectives.</p> <p>No. A Service Provider dominating in the market would like to apply its own standards for interconnection and portability, for instance.</p> | <p>No. Different standards and rules for managing security and privacy may exist, depending on the Service Provider.</p> |
| 2) Balance of interests | <p>Yes. Balancing the interests in protecting Security and Privacy, on the one hand, and fostering EU-wide services on the other hand, can be agreed at least at the EU level.</p> <p>No. Balancing the interests across regions (Europe, America, Asia) is almost impossible.</p> <p>No. To control many types and uses of information in balance with encouraging the commercial exploitation of new systems and business models will be challenging.</p> | <p>Yes/No. Industrial players try to realize the benefits of the new technologies. However, operating, e.g. in the shared environment needs more trust between parties. And that may depend on the service in question.</p> <p>Yes. Consumers seem not to care so much about Privacy, in practice.</p> <p>No. There will be variations of balance in different regions and countries.</p> <p>No. Commercial interests may lead to breaches in using information and, thus, breaking the balance.</p> | <p>Yes. Consumers and corporate users can choose whether to deal with businesses who promise them a given level of Security and Privacy. Service Providers who offer Security and Privacy that pleases consumers and corporate users succeed.</p> <p>No. Where some consumers may have strict senses of privacy, others have fewer reservations about revealing personal information and receiving benefits of participation on commercial life.</p> <p>Yes. Most consumers seem not to care about Privacy, in practice.</p> |

| | | |
|-------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| | | |
| 3) Global context | <p>Yes. Security and Privacy legislation and its compatibility with new technologies can be agreed at the European level.</p> <p>No. The compatibility of Security and Privacy legislation with the new technologies is difficult to reach across regions.</p> <p>No. The different levels of privacy are difficult to regulate. Consumers may have:</p> <ul style="list-style-type: none"> ● strict senses of privacy; or ● less reservations about revealing personal information. | <p>No. Local players may have very different views and interests about Security and Privacy.</p> <p>Yes. Big international players will have interests to agree on rules that are applied globally.</p> |
| 4) Foster Interoperability and data portability | <p>Yes. Agreements between governments are needed to push the European-wide standards and practices. This is especially important for enabling new entries in the market.</p> | <p>No. A dominant player may want to push their own closed standards on Interoperability.</p> <p>Yes. The telecom players have a tradition to agree on the Interoperability.</p> |
| 5) Applicable law must be easy to define | <p>Yes. The governments (EU Commission) can agree on the applicable law in Europe.</p> <p>Yes. Responsibility and accountability of those storing and processing data can be defined.</p> <p>No. The agreement on the applicable law across regions would be difficult to reach (see Global context)</p> | <p>No. The Industrial players have no mandate to agree on the applicable law.</p> <p>No. Depending on the bylaw of the hosting country, data may be protected by different applicable jurisdiction. This may lead to reduced use of new services.</p> |
| | | No. Market regulation is not allowed in some regions. |
| | | No. Consumers or corporate users have no power to push Interoperability and data portability. |
| | | No. Consumers have no mandate to agree on the applicable law. |

(Continued)

Table 17.2 (Continued)

| Criteria (Regulatory objectives) | Government regulation | Industry self-regulation | Consumer or market regulation |
|------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| 6) Right to be forgotten | Yes. The governments can enforce the rule for "Right to be forgotten". | Yes. The consensus on the right to be forgotten can be reached between the companies of good reputation. No. Rogue companies, that do not want to follow industry standards, may take the opportunity to charge. | No. Consumers cannot enforce to be forgotten by the Cloud Service Provider. |
| 7) Increased responsibility and accountability | Yes. Governments define and enforce the responsibility and accountability for Service Providers. This may be challenging in case of Cloud Computing, however, because the data may be located anywhere. | No. Definition of responsibility and accountability across regions and countries is challenging for different kinds of industrial players. Lack of clear definition of responsibility among Service Providers and users may evoke conflicts. | No. Consumers have no power to define nor to enforce responsibility and accountability within and across regions and countries. |

In that assessment, Options 1 and 3 are not considered viable for reaching the policy objectives, and are therefore not recommended. The reasons not to recommend are that their effectiveness would depend on whether the voluntary approach would actually deliver a minimum level of NIS and, regarding Option 3, it would depend on the good will of the Member States to set up capabilities and co-operate cross border.

Option 2 is the preferred one, given that under this Option the protection of EU consumers, business and Governments against NIS incidents, threats and risks would improve considerably. The analysis supports the Impact Assessment of the NIS Directive: Government regulation that would best promote the targets for Security and Privacy.

The technology (Clouds, NFV and SDN) implications on Security and Privacy regulations can be summarized as (in a random order):

- New technologies will allow new types of market structures with new types of players in the telecommunications value chain. Different players have different views, practices and objectives for securing Security and Privacy. The responsibilities and accountabilities for securing end-to-end Security and Privacy, and the ownership rights over the data, have to be clearly and firmly defined across country borders;
- Due to the de-centralized architecture of the Clouds-based implementations, the protection mechanisms are likely to be inconsistent. Regulators have to push for the consistent systems with respect to Security and Privacy;
- Due to the de-centralized nature of architecture (Clouds, Virtualization) the user related data may be located where-ever. In this new, more complicated environment, the governments have to also enforce the rule for “right to be forgotten”;
- A loss of governance may lead to a Service Provider lock-in (one kind of loss of Security). Regulators need to take action to foster interoperability and data portability;
- Balancing of interests in protecting Security and Privacy, on the one hand, and fostering EU-wide (and global) services, on the other hand, shall be agreed at least at the EU level. However, different levels of Privacy may be difficult to regulate;
- The new Security and Privacy issues in the context of new technologies are more of a global nature than ever before, and EU-wide (global) approaches are needed in Regulation. EU-wide standards and practices are also needed. This is also important for enabling new entries to the market. The location of legal disputes has to be clear and agreed.

References

- 1 Security in Telecommunications and Information Technology (2003) ITU-T. Available at: <http://www.itu.int/itudoc/itu-t/85097.pdf>
- 2 ITU-T Recommendation X.805 (2003) Security architecture for systems providing end-to-end communications. Available at: <http://www.itu.int/rec/T-REC-X.805-200310-I/en>
- 3 Telecoms Council, Brussels (2009) 31 March 2009. Available at: <http://europa.eu/rapid/pressReleasesAction.do?reference=MEMO/09/139&format=HTML&aged=0&language=EN&guiLanguage=en/>

- 4 Commissioner Kroes speech on the Digital Economy and exchange of views on Internet Governance in Parliament: Why the digital economy matters. Düsseldorf, 26 March 2014.
- 5 Hyland, N. (2011) Loss of Personal Data is Sufficient to Advance Privacy Lawsuit [Online]. Available at: <http://www.cyberlawcurrents.com/?p=1394> (accessed on 9 May 2011).
- 6 Wong, N. (2006) Response to the DoJ motion, The Official Google Blog [Online]. Available at: <http://googleblog.blogspot.com/2006/02/response-to-doj-motion.html> (accessed on 9 May 2011).
- 7 Gilbert Tobin Lawyers (2007) Economic study on IP interworking, February 2007. Available at: http://www.gsmworld.com/documents/IP_Interconnection_Economic_Study_on_IP_Networking.pdf
- 8 Industry Recommendations to Vice President Neelie Kroes on the Orientation of a European Cloud Computing Strategy (2011). Available at: http://ec.europa.eu/information_society/activities/cloudcomputing/docs/industryrecommendations-ccstrategy-nov2011.pdf
- 9 Industry joint paper on the review of the EU Legal Framework for data protection. Available at: http://www.orange.com/en_EN/group/european_policy/privacy/att00022388/1010_Industry_Joint_Paper_on_Data_Protection.pdf
- 10 Article 29 Data Protection Working Party, Opinion 05/2012 on Cloud Computing (adopted July 1, 2012).
- 11 Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications). Available at: <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32002L0058:en:HTML>
- 12 Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 24 October 1995. Available at: http://ec.europa.eu/justice_home/fsj/privacy/docs/95-46-ce/dir1995-46_part1_en.pdf
- 13 Eurescom mess@ge (2012) The magazine for telecom insiders, 1/2012.
- 14 Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Available at: <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0011:FIN:EN:PDF>
- 15 Commission Proposal for a Directive on Network and Information Security COM (2013) 48 final. Available at: http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=1666
- 16 Press Release: The EP successfully votes through the Network & Information Security (NIS) directive. Available at: http://europa.eu/rapid/press-release_STATEMENT-14-68_en.htm
- 17 SWD (2013) 31. Executive Summary of the Impact Assessment. Accompanying the document: Proposal for a Directive of the European Parliament and of the Council Concerning measures to ensure a high level of network and information security across the Union. Available at: http://eeas.europa.eu/policies/eu-cyber-security/cybsec_impact_ass_res_en.pdf

- 18 Ruiter, J. and Warnier, M. *Privacy Regulations for Cloud Computing Compliance and Implementation in Theory and Practice*. Available at: http://www.iids.org/aigaion/indexempty.php?page=actionattachment&action=open&pub_id=316&location=spcc10.pdf-9869f6a896824ba29d27ad19e6da5585.pdf
- 19 Cloud Computing Legal Issues: When does Directive 95/46/EC Apply? Available at: <http://common-assurance.com/blog/files/2cf981cc32595f347ba14371ea17643f-11.html>
- 20 Computer World UK, Blog: Cloud computing and EU data protection law, 28 September 2011. Available at: <http://blogs.computerworlduk.com/cloud-vision/2011/09/cloud-computing-and-eu-data-protection-law/index.htm>
- 21 University of London, Centre for Commercial Law Studies, News: Data protection law creates cloud of uncertainty for cloud computing, 21 November 2011, Cloud of Unknowing papers. Available at: <http://www.ccls.qmul.ac.uk/news/2011/59982.html>
- 22 ITU Focus Group on Cloud Computing. FG Cloud Technical Report, Part 5. Available at: <http://www.itu.int/en/ITU-T/focusgroups/cloud/Pages/default.aspx>
- 23 NIST, Draft Special Publication 800-144: Guidelines on Security and Privacy in Public Cloud Computing. Available at: https://cloudsecurityalliance.org/wp-content/uploads/2011/07/NIST-Draft-SP-800-144_cloud-computing.pdf
- 24 NFV 0010, Network Functions Virtualization; Architectural Framework.
- 25 Alcatel-Lucent: Network Functions Virtualization! Challenges and Solutions. Strategic White Paper. Available at: <http://www.tmcnet.com/tmc/whitepapers/documents/whitepapers/2013/9377-network-functions-virtualization-challenges-solutions.pdf>
- 26 NFV 0012, Network Functions Virtualization; Virtualization Requirements.
- 27 Open Network Foundation (ONF), SDN Security Considerations in the Data Center.
- 28 McAfee Blog Central, Software Defined Networking Promises Greater Control While Increasing Security Risks. Available at: <http://blogs.mcafee.com/mcafee-labs/2014-threats-predictions-software-defined-networking-promises-greater-control-while-increasing-security-risks>
- 29 SIGMONA Internal Report IR4.1, State-of-the-Art in Mobile Transport Networks.
- 30 SDN security challenges alongside the potential of a new technology. Available at: <http://searchsdn.techtarget.com/tip/SDN-security-challenges-alongside-the-potential-of-a-new-technology>
- 31 SDN Central. Security Challenges in SDN (Software-defined Networks). Available at: <http://www.sdncentral.com/security-challenges-sdn-software-defined-networks/>
- 32 Privacilla.org. Available at: <http://www.privacilla.org/business/howtoregulate.html>
- 33 FP7-ICT-2009-5-257448-SAIL (2012) *Deliverable D-2.8 – Evaluation of Business Models*.
- 34 Natarajan, S. and Wolf, T. (2012) *Security Issues in Network Virtualization for the Future Internet*. Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA. Available at: <http://www.ece.umass.edu/ece/wolf/pubs/icnc2012.pdf>

Index

Note: Page numbers in *italic* denote figures, those in **bold** denote tables.

- 1G cellular systems 6–8
 - advanced mobile phone service (AMPS) 7
 - security 7–8, 60, 61–62
 - threats and protection 61–62
- Total Access Communication Systems (ETACS) 7
- 2G cellular systems 8–15
 - code division multiple access (CDMA) 10, 11
 - enhanced data rate for GSM evolution (EDGE) 10
 - global system for mobile communications 8–9
 - Groupe Spéciale Mobile (GSM) 8–15
 - GSM packet radio systems (GPRS) 9–10
 - security 10–15, 60, 62–63
 - short messaging service (SMS) 8
 - subscriber identity module (SIM) 11
 - threats and protection 62–63
- 3G cellular systems 15–20
 - CDMA 2000 15
 - HSPA (high-speed packet access) 17
 - security 17–20, 60, 63
 - threats and protection 63
 - UMTS network architecture 16
 - UMTS WCDMA 15–16
- 3GPP activities, 5G standardization activities 49–50
- 4C business model, cyber security in 5G 104–105, 114
- 4G cellular systems 22–27
 - bandwidth management 24
 - carrier aggregation 24
- Cooperative Multipoint Transmission and Reception for LTE-Advanced 23–24
- enhanced MIMO 23
- Evolved Packet Core (EPC) 24–25
- key technologies 23–24
- LTE (Remote Access Network) domain security 65
- LTE-advanced E-UTRAN architecture 24–25
- LTE core network domain security 65
- LTE security model 26–27
- LTE UE (User Equipment) domain security 64–65
- Multiple-Input Multiple-Output (MIMO) 23
- network architecture 24–25
- relays 24
- requirements 22
- security 25–27, 61, 63–66
- security threat analysis 65–66
- spectrum management 24
- threats and protection 63–66
- WiMAX security 27
- 5G characteristics definitions, NGMN Alliance white paper 347–348
- 5G cloud computing architecture 379–380
- 5G cyber security *see* cyber security in 5G
- 5G end-to-end system
 - 5G mobile networks 46–48
 - Management and Orchestration (MANO) 47–48
- 5G- LiFi security challenges 160
- device connectivity, large number 160

- 5G mobile networks
see also Mobile Virtual Network Operators
- 5G end-to-end system 46–48
 - 5G mobile core network 44–46
 - 5G research communities 52–54
 - advance malware **69**
 - APTs (Advance Persistent Threats) 68
 - cloud computing 46
 - cost efficiency 36, 37
 - critical infrastructure threats **70**
 - customer perspective 32
 - defining 31
 - enabling technologies 37–44
 - energy efficiency 36, 37
 - enhanced mobile broadband (eMBB) 32–33
 - evolved security model 68–69
 - flexibility 36, 37
 - high availability 35, 37
 - high data rate 34–35, 37
 - Internet of Things (IoT) botnets **70**
 - Internet of Things (IoT) threat
 - landscape 68, 195–196
 - massive connectivity 35, 37
 - Massive Machine-type Communications (mMTC) 33
 - mobile OS and app patch
 - management 71–72
 - Network Function Virtualization (NFV) 44–46
 - paradigm shift 31–32
 - privacy 36–37, 37
 - programmability 36, 37
 - ransomware 67, 68, **69**
 - reliability 35, 37
 - requirements 33–44
 - seamless mobility 35, 37
 - secure device management 71
 - security 36–37, 37, 61, 66–73
 - security failures, key 70–71
 - security lifecycle functions 70–73
 - security monitoring 72–73
 - security threat analysis 69–70
 - security threat analysis and assessment 72
 - Software Defined Network (SDN) 44
 - spectrum 32
 - spectrum efficiency 36, 37
 - standardization activities 48–52
 - system architecture 32
 - technologies, enabling 37–44
- threats and protection 61, 66–70
- ultra low latency 34–35, **37**
 - ultra-reliable and low latency communications (URLLC) 33
 - use cases 32–33
 - user perspective 32
 - Virtual Network Functions (VNF) 44–46
 - zero-day attacks **70**
- 5G networks
- cloud computing 374–381
 - Multi-Access Edge Computing (MEC) 381–385
 - Software Defined Monitoring (SDM) 231–242
- 5G positioning 281–312
- 5G localization chain 282–283
 - Angle of Arrival (AOA) **284**, 287, 288, 297, 298
 - backup systems 298–299
 - Brands and Chaum's cryptographic distance-bounding 301–302
 - cryptographic authentication 300–301
 - cryptographic distance-bounding 301–302
 - cryptographic techniques 299–304
 - Direction or angle of Arrival (DOA) 281, **284**, 286–287, **286**
 - European projects **311–312**
 - features summary **284**
 - fingerprinting 289, 297
 - interference signals 297–298
 - International projects **311–312**
 - legislation on user location privacy in 5G 304–310
 - mechanisms summary **286**
 - outdoor vs. indoor technologies 283, **284**
 - overview 285–290
 - passive positioning vs. active positioning 283–285
 - positioning-communication dichotomy in 5G 285
 - privacy cryptographic techniques 299–304
 - privacy-preserving location-based services 303–304
 - Received Signal Strength (RSS) 281, **284**, **286**, 288–289, 295–298
 - security cryptographic techniques 299–304

- Time Difference of Arrival (TDOA) 281, 284, **284**, 286, **286**, 287, 296, 298
- Time of Arrival (TOA) 281, **284**, 286–289, **286**, 296–297
- trustworthiness enhancement 299
- 5G projected capacity 195
- 5G Radio Access Network (RAN) 38–44
- Cloud-based Radio Access Network (Cloud-RAN) 42
 - D2D communication 40–42
 - Device-to-Device (D2D) communication 40–42
 - Fog Computing (FC) 42–43
 - M2M communication 40, 41
 - Machine Type Communication (MTC) 40, 41
 - massive MIMO 38–39
 - mmWave Communication 38
 - Mobile Edge Computing (MEC) 42–44
 - small cell technology 39–40
 - ultra-dense small cells 39–40
- 5G research communities 52–54
- 5G security 75–95
- see also* cyber security in 5G
 - access network security 79
 - challenges 76–81
 - control layer 80–81, 85–86
 - core network 80–81
 - Denial of Service (DoS) attacks 79–80
 - design principles 75–95
 - Distributed DoS (DDoS) attacks 79–80
 - flash network traffic 78
 - future directions 94–95
 - ITU-T recommendations 77
 - mandated security in the network 78
 - network access security 79
 - Network Function Virtualization (NFV) 81–83, 231–232, 351
- Next Generation Mobile Networks (NGMN) 76, 78–79, 81–82
- novel technologies 81–87
- overviews 76–81
- radio interface keys 78
- recommendations 76–81
- Software Defined Networking (SDN) 81–82, 83–94
- subscriber level security policies 78–79
- user plane integrity 78
- Virtual Network Functions (VNF) 81–83
- 5G security and privacy, regulation 399–417
- 5G standardization activities 48–52
- 3GPP activities 49–50
 - European Telecommunications Standards Institute (ETSI) 50–51
 - Institute of Electrical and Electronic Engineers (IEEE) 51
 - International Telecommunication Union (ITU) 48–49
 - Internet Engineering Task Force (IETF) 52
- 5G use cases and requirements, Internet of Things (IoT) 247–248
- 5G-WiFi interoperability 143–146
- 5G-WLAN security 143–149
- application layer 149
 - MAC layer 147
 - network architecture 146–149
 - network layer 147–148
 - transport layer 148–149
- 5G-WiFi security 150–156
- anonymity gains 151
 - architectural design 156–157, 160–161
 - confidentiality attacks 151
 - Denial of Service (DoS) attacks 151
 - device connectivity, large number 151
 - impersonation attacks 151
 - increased device connectivity 151
 - Network Function Virtualization (NFV) 150
 - network security flaws 151
 - non-cryptography-based security 150–151
 - packet switched networks **154**
- 5G-WLAN security 143–161
- 5G- LiFi security challenges 160
- architectural design 156–157, 160–161
 - LiFi-5G networks interoperability 159–160
 - LiFi networks 158–159
 - security challenges 150–156
 - WiFi-5G networks
- interoperability 143–146
- a**
- A3 algorithm, Groupe Spéciale Mobile (GSM) 12–13
- A5 algorithm, Groupe Spéciale Mobile (GSM) 13, 14
- A8 algorithm, Groupe Spéciale Mobile (GSM) 13

- access control
 cloud computing 387–388
 Software Defined Networking (SDN) 91
- access network security, 5G security 79
- accidental action, structures 174, 179
- active positioning vs. passive positioning, 5G positioning 283–285
- Active Probing, monitoring technique 232
- advanced mobile phone service (AMPS), 1G cellular systems 7
- advance malware, 5G mobile networks 69
- Advance Persistent Threats (APTs), 5G mobile networks 68
- ALARP (As Low As Reasonably Practicable) principle, risk assessment/evaluation 190
- American 5G related activities, 5G research 54
- AMPS (advanced mobile phone service), 1G cellular systems 7
- ancillary items, structures 178
- Angle of Arrival (AOA), 5G positioning 284, 287, 288, 297, 298
- anonymity-based approaches, privacy 272
- anonymity gains, 5G-WiFi security 151
- AOA *see* Angle of Arrival
- application layer
 5G-WiFi interoperability 149
 LiFi-5G networks interoperability 159
 Software Defined Networking (SDN) 84–85
- application plane security, Software Defined Networking (SDN) 86–87
- application security, Telecommunication network as a Service (Taas) 333
- Apps (smartphone applications), privacy 270
- APTs (Advance Persistent Threats), 5G mobile networks 68
- architectural design
see also structural design
 5G-WiFi security 156–157, 160–161
 5G-WLAN security 156–157, 160–161
 LiFi-5G networks 160–161
 security consideration 160–161
 WiFi-5G networks 156–157, 160–161
- architectural resilience, cyber preparedness 105–106
- architecture
 centralized security architectures 391–392
 cloud computing 376–377, 379–380, 391–392
 mobile network 197, 198
- Multi-Access Edge Computing (MEC) 391–392
- SDN-based cloud security systems 392
- Software Defined Monitoring (SDM) 235–238
- Asian 5G related activities, 5G research 53–54
- As Low As Reasonably Practicable (ALARP) principle, risk assessment/evaluation 190
- b**
- B2B (business-to-business) customers, cyber security in 5G 107
- B2C (business-to-customer) customers, cyber security in 5G 107
- B2G (business-to-government) customers, cyber security in 5G 107
- backup systems, 5G positioning 298–299
- bandwidth management, 4G cellular systems 24
- bluetooth, security attacks 155
- Bot-detection, Realm Gateway (RGW) 213, 220–222
- Brands and Chaum's cryptographic distance-bounding, 5G positioning 301–302
- business model approach
 4C business model 104–105, 114
 cyber preparedness 105–106
 cyber security in 5G 103–106
 layered 4C model 104–105, 114
- business model options, cyber security in 5G 112–114
- c**
- carrier aggregation, 4G cellular systems 24
- CDMA (code division multiple access)
 2G cellular systems 10, 11
 Supplemental Code Channel (SCH) 10
- CDMA 2000
 3G cellular systems 15
 security 17–18
- Celtic-Plus, 5G research 53
- centralized security, cyber security in 5G 109–110
- CES *see* Customer Edge Switching
- CETP *see* Customer Edge Traversal Protocol
- China: IMT-2020 5G Promotion Group, 5G research 54

- cleaver cracks, structural design
 - failures 184–185
- Cloud-based Radio Access Network (Cloud-RAN), 5G mobile networks 42
- cloud computing 373–394
 - 5G cloud computing
 - architecture 379–380
 - 5G mobile networks 46
 - 5G networks 374–381
 - access control 387–388
 - architecture 376–377, 379–380, 391–392
 - centralized security
 - architectures 391–392
 - cloud deployment models 377–378
 - cloud intrusion 387
 - cloud security research 392–394
 - cloud service models 378
 - community cloud 378
 - Cyber-Physical System (CPS) security 386
 - cyber security in 5G 111
 - defining 393
 - Extreme Mobile Broadband (xMBB)
 - 380–381
 - history 375–376
 - hybrid cloud 378
 - Infrastructure as a Service (IaaS) 378
 - legal framework for security and privacy 403–404
 - Massive Machine-type Communications (mMTC) 381
 - National Institutes of Standards and Technology (NIST) 406–407
 - overview 375–376
 - Platform as a Service (PaaS) 378
 - private cloud 377
 - private data computation 386–387
 - public cloud 377–378
 - scenarios 380–381
 - SDN-based cloud security systems 392
 - secure data computation 386–387
 - security and privacy 405–407, 410–411, 417
 - security architectures 391–392
 - security challenges 385–388
 - Software as a Service (SaaS) 378
 - Software Defined Networking (SDN) 392
 - standardizations 392–394
 - Ultra-reliable Machine-type Communications (uMTC) 381
 - use cases 380–381
 - virtualization security 385–386
- cloudification of the network operators
 - availability requirements 334–335
 - Mobile Virtual Network Operators (MVNO) 325–326, 334–335
- Telecommunication network as a Service (TaaS) 325–326
- cloud intrusion, cloud computing 387
- cloud security research
 - European Telecommunications Standards Institute (ETSI) 392–393
 - Multi-Access Edge Computing (MEC) 392–394
 - National Institutes of Standards and Technology (NIST) 393
 - Next Generation Mobile Networks (NGMN) 393–394
- code division multiple access (CDMA), 2G cellular systems 10
- Codes of Practice, structures 187, 188
- communications
 - Mobile Virtual Network Operators (MVNO) 263–265
 - robots 263, 265
- communications failures 165
- community cloud, cloud computing 378
- COMP128, Groupe Spéciale Mobile (GSM) 14
- confidentiality, cf. privacy 399–400
- confidentiality attacks, 5G-WiFi security 151
- consumer/market regulation 414–416
- control channels security, Software Defined Networking (SDN) 89–91
- control layer
 - 5G security 80–81, 85–86
 - Software Defined Networking (SDN) 85–86
- control plane security, Software Defined Networking (SDN) 87
- Cooperative Multipoint Transmission and Reception for LTE-Advanced, 4G cellular systems 23–24
- core network, 5G security 80–81
- corporate gateway, Customer Edge Switching (CES) 225–226
- cost efficiency, 5G mobile networks 36, 37
- cost of cyber-attacks 102–103
- CPS (Cyber-Physical System) security, cloud computing 386

- critical information protection, cyber preparedness 105–106
- critical infrastructure threats, 5G mobile networks **70**
- cryptographic authentication, 5G positioning 300–301
- cryptographic distance-bounding, 5G positioning 301–302
- cryptographic techniques, 5G positioning 299–304
- Customer Edge Switching (CES) 196–229 advantages 196–197, 229 CES testbed 213–214 CETP policy-based communication 206–210, 214–217 corporate gateway 225–226 deployment 197, 203–205 deployment in 5G networks 222–228 Domain Name System (DNS) to initiate communication 205–206 features 196–197 Industrial Internet (II) 227–228 Internet classical weaknesses 196–197, 209–210 key performance indicators (KPIs) 213–214 mobile broadband 224–225 national CERT centric trust domain 226 Network Address Translation (NAT) 197, 203–205 cf. Network Address Translation (NAT) 204 policy architecture 209 Realm Gateway (RGW) 210–212 reliability 225, 226, 227, 228 scalability 225, 226, 227, 228 security benefits 224–225, 226, 227, 228 security evaluation 213–222 security framework 203–213 security mechanisms 209–210, 229 security testing 214–217 Software Defined Networking (SDN) 200, 223, 229 state-of-the-art in mobile networks security 197–203 Customer Edge Traversal Protocol (CETP) 204–206 CETP policy-based communication 206–210, 214–217 customer groups, cyber security in 5G 107 cyberactivism 100–102 cybercrime 100–102 cyber-espionage 100–102 Cyber-Physical System (CPS) security, cloud computing 386 cyber preparedness, business model approach 105–106 cyber security in 5G 99–114 4C business model 104–105, 114 B2B (business-to-business) customers 107 B2C (business-to-customer) customers 107 B2G (business-to-government) customers 107 business case 106–111 business model approach 103–106 business model options 112–114 centralized security 109–110 cloud computing 111 context 100–103 cost of cyber-attacks 102–103 customer groups 107 cyberactivism 100–102 cybercrime 100–102 cyber-espionage 100–102 cyber preparedness 105–106 cyberterrorism 100–102 cyberwarfare 100–102 delivering 110–111 device-driven security 113 device-to-device (D2D) security 109–110 distributed security 109–110 edge network 111 general-purpose technologies (GPTs) 106 home carrier 110 infrastructure security 109–110 interconnect 111 IoT (Internet of Things) sensors 111 issues 108–109 landscape 109–110 layered 4C model 104–105, 114 location-driven security 113 M2M gateway 111 new revenue generation vs. level of security 107–108 platform-driven security 112–114 radio access 111 radio terminals 111 roaming carrier 111

- scenarios 109–110, **111**
- types of threat 101–102
- users 108–109
- cyberterrorism 100–102
- cyberwarfare 100–102

- d**
- D2D *see* Device-to-Device
- data decomposition, watermark-based blind physical layer security (WBPLSec) 122, 124
- data integrity, WiFi-5G networks 156–157
- data link security, Software Defined Networking (SDN) 88–89
- data plane security, Software Defined Networking (SDN) 87
- data privacy 269–271
- Data Protection Directive (Directive 95/46/EC) 304–306, **309**, 402–404
- EU legal instruments **309**
- legal framework for security and privacy 304–305, 402–404
- data security, Telecommunication network as a Service (TaaS) 327–328
- DDoS (Distributed DoS) attacks, 5G security 79–80
- Deaves and Harris model, structures 175
- definitions
 - 5G characteristics, NGMN Alliance white paper 347–348
 - 5G mobile networks 31
 - cloud computing 393
- Denial of Service (DoS) attacks
 - 5G security 79–80
 - 5G-WiFi security 151
 - Multi-Access Edge Computing (MEC) 389
 - physical layer security 119
- device connectivity, large number
 - 5G- LiFi security 160
 - 5G-WiFi security 151
- device-driven security, cyber security in 5G 113
- device identity confidentiality, WiFi-5G networks 156
- device identity privacy 273
- Device Polling, monitoring technique 233
- Device-to-Device (D2D)
 - communication, 5G mobile networks 40–42
 - security, cyber security in 5G 109–110
- Direction or angle of Arrival (DOA), 5G positioning 281, **284**, 286–287, **286**
- Distributed DoS (DDoS) attacks, 5G security 79–80
- distributed network security services, Network Function Virtualization (NFV) 366
- distributed security
 - cyber security in 5G 109–110
 - Internet of Things (IoT) 254–259
- DNS (Domain Name System)
 - communication initiation, Customer Edge Switching (CES) 205–206
 - Realm Gateway (RGW), DNS floods 217–222
- DOA *see* Direction or angle of Arrival
- Domain Name System (DNS)
 - Customer Edge Switching (CES), communication initiation 205–206
 - Realm Gateway (RGW) 211–212
 - Realm Gateway (RGW), abuse prevention 213
 - Realm Gateway (RGW), DNS floods 217–222
- DoS *see* Denial of Service attacks
- DPD *see* Data Protection Directive
- drones 248–249, 252, 253–254, 259, 260–263, 262, 264

- e**
- early development, evolution of cellular systems 4–6
- EDGE (enhanced data rate for GSM evolution), 2G cellular systems 10
- edge network, cyber security in 5G 111
- eMBB (enhanced mobile broadband), 5G mobile networks 32–33
- encryption, physical layer security 119–120
- end-system security *see* Customer Edge Switching
- energy efficiency, 5G mobile networks 36, **37**
- enhanced data rate for GSM evolution (EDGE), 2G cellular systems 10
- enhanced MIMO, 4G cellular systems 23
- enhanced mobile broadband (eMBB), 5G mobile networks 32–33

- EPC (Evolved Packet Core), 4G cellular systems 24–25
- ePrivacy Directive *see* Privacy Directive (EC Directive 2002/58/EC)
- ETACS (Total Access Communication Systems), 1G cellular systems 7
- ETSI *see* European Telecommunications Standards Institute
- EU FP7, 5G research 52
- EU H2020 program, 5G research 52–53
- EU legal instruments
- Data Protection Directive (Directive 95/46/EC) **309**
 - General Data Protection Regulation (GDPR) (EU) **309**
 - legislation on user location privacy in 5G **309**
 - Privacy Directive (EC Directive 2002/58/EC) **309**
- Eurocodes, structures 170, 173, 180
- European Telecommunications Standards Institute (ETSI) 48
- vs. Network Address Translation (NAT) 50–51
 - Network Function Virtualization (NFV) 348–350
 - standardizations 392–393
- evolution of cellular systems
- 1G cellular systems 6–8
 - 2G cellular systems 8–15
 - early development 4–6
- Evolved Packet Core (EPC), 4G cellular systems 24–25
- evolved security model, 5G mobile networks 68–69
- Extreme Mobile Broadband (xMBB), cloud computing 380–381
- f**
- fatigue, structural design failures 183, 185
- FBG (Fibre Bragg Gratings), structures 171, 190
- FC (Fog Computing), 5G mobile networks 42–43
- FEA (finite element analyses), structures 179–180
- Fibre Bragg Gratings (FBG), structures 171, 190
- fingerprinting, 5G positioning 289, 297
- finite element analyses (FEA), structures 179–180
- firewalls, Software Defined Networking (SDN) 92
- flash network traffic, 5G security 78
- flexibility, 5G mobile networks 36, 37
- Flow Collection, monitoring technique 233
- Fog Computing (FC), 5G mobile networks 42–43
- future directions
- 5G security 94–95
 - Mobile Virtual Network Operators (MVNO) 340–341
- future scenarios, legislation on user location privacy in 5G 309–310
- g**
- GBP (group-based policy), Network Function Virtualization (NFV) 367–368
- General Data Protection Regulation (GDPR) (EU)
- EU legal instruments **309**
 - legislation on user location privacy in 5G 304, 305–307, 308–309
- general-purpose technologies (GPTs), cyber security in 5G 106
- geometrical and material nonlinear imperfect analysis (GMNIA), structures 179
- Gi/SGi interface, common threats 197–203
- global system for mobile communications, 2G cellular systems 8–9
- global trust operator (GTO)
- mobile networks 202–203
 - trust processing 202–203
- GMNIA (geometrical and material nonlinear imperfect analysis), structures 179
- government regulation **414–416**
- GPRS (GSM packet radio systems), 2G cellular systems 9–10
- GPTs (general-purpose technologies), cyber security in 5G 106
- group-based policy (GBP), Network Function Virtualization (NFV) 367–368
- Groupe Spéciale Mobile (GSM)
- 2G cellular systems 8–15
 - A3 algorithm 12–13
 - A5 algorithm 13, 14
 - A8 algorithm 13

- authentication process 12
 COMP128 14
 International Mobile Subscriber Identity (IMSI) 11–12, 62–63
 Ki 12
 network architecture 9–10
 security 11–14
 Temporary Mobile Subscriber Identity (TMSI) 11
 growth of communication services 3, 4
GSM *see* Groupe Spéciale Mobile
 GSM packet radio systems (GPRS), 2G cellular systems 9–10
 GTO (global trust operator)
 mobile networks 202–203
 trust processing 202–203
- h**
- high availability, 5G mobile networks 35, 37
 high data rate, 5G mobile networks 34–35, 37
 high-speed packet access (HSPA), 3G cellular systems 17
 historic perspective, evolution of cellular systems 4–6
 home carrier, cyber security in 5G 110
 horizontal network slicing, Network Function Virtualization (NFV) 354
 HSPA (high-speed packet access), 3G cellular systems 17
 HSPA+ 20
 hybrid cloud, cloud computing 378
 hypervisor security, Telecommunication network as a Service (TaaS) 328–329
- i**
- ice, structural design failures 183
 identity issues 267–278
 background 268–269
 identity management 273–274
 International Mobile Subscriber Identity (IMSI) 268–269, 273–274
 Universal Subscriber Identity Module (USIM) 274, 275–277
 identity privacy 272–273
 IDS (Intrusion Detection Systems), Software Defined Networking (SDN) 91
 IEEE (Institute of Electrical and Electronic Engineers), 5G standardization activities 51
 IETF (Internet Engineering Task Force), 5G standardization activities 52
 II *see* Industrial Internet
 iJAM protocol
 physical layer security 121–122
 secrecy capacity 129–131
 cf. WBPLSec 130–131, 139
 impersonation attacks, 5G-WiFi security 151
 IMSI (International Mobile Subscriber Identity), Groupe Spéciale Mobile (GSM) 11–12, 62–63
 inconsistent security policies, Multi-Access Edge Computing (MEC) 389–390
 increased device connectivity, 5G-WiFi security 151
 indoor vs. outdoor technologies, 5G positioning 283, 284
 Industrial Internet (II) 196, 197, 223
 Customer Edge Switching (CES) 227–228
 road traffic and transport 227–228
 industry self-regulation 414–416
 Infrastructure as a Service (IaaS)
 cloud computing 378
 Network Function Virtualization (NFV) 354–355
 Telecommunication network as a Service (TaaS) 326, 336, 337–338, 339, 340
 infrastructure layer, Software Defined Networking (SDN) 86
 infrastructure security
see also structures
 cyber security in 5G 109–110
 Institute of Electrical and Electronic Engineers (IEEE), 5G standardization activities 51
 interconnect, cyber security in 5G 111
 interference signals, 5G positioning 297–298
 International Mobile Subscriber Identity (IMSI)
 Groupe Spéciale Mobile (GSM) 11–12, 62–63
 identity issues 268–269
 privacy 268–269
 trust issues 268–269
 International Telecommunication Union (ITU)
 5G standardization activities 48–49
 ITU-T recommendations, 5G security 77

Internet, Industrial *see* Industrial Internet
 Internet classical weaknesses 196–197
 Customer Edge Switching (CES)
 196–197, 209–210
 Internet Engineering Task Force (IETF), 5G
 standardization activities 52
 Internet of Things (IoT)
 5G use cases and requirements 247–248
 botnets, 5G mobile networks **70**
 devices attacks 250–253
 distributed security 254–259
 drones 248–249, 252, 253–254, 259,
 260–263, 262, 264
 literature overview 249–254
 Mobile Cloud Robot (MCR) 247–248,
 259–263
 physical layer security 120
 privacy 270
 research motivation 253–254
 robot attack classification 255–256
 robot data classification 254
 robot security platform 256–259
 robot threats classification 255–256
 security 247–265
 security threats 250, 252–253, 253
 sensors, cyber security in 5G 111
 services attacks 250–253
 threat landscape, 5G mobile networks 68,
 195–196
 use case evolution 250–253
 Intrusion Detection Systems (IDS), Software
 Defined Networking (SDN) 91
 Intrusion Prevention Systems (IPS), Software
 Defined Networking (SDN) 91
 IoT *see* Internet of Things
 IPS (Intrusion Prevention Systems), Software
 Defined Networking (SDN) 91
 IS-95, security 14–15
 ITU (International Telecommunication
 Union)
 5G standardization activities 48–49
 ITU-T recommendations, 5G security 77

j
 jamming, physical layer security 121
 jamming channel, watermark-based blind
 physical layer security (WBPLSec)
 123–124
 jamming receiver

outage probability 131–136
 physical layer security 131–136
 secrecy capacity 131–136
 watermark-based blind physical layer
 security (WBPLSec) 126
 Japan: 5GMF Forum, 5G research 54

k
 key management, WiFi-5G networks 157
 Keystone, OpenStack 362, 363
 Ki, Groupe Spéciale Mobile (GSM) 12

l
 layer-based SDN architecture, Software
 Defined Networking (SDN) 329–331
 layered 4C model, cyber security in 5G
 104–105, 114
 legacy monitoring systems 233–234, 238–240
 see also Software Defined Monitoring
 vs. Software Defined Monitoring (SDM)
 238–240
 legal framework for security and privacy
 402–405
 cloud computing 403–404
 Data Protection Directive (Directive 95/46/
 EC) 304–305, 402–404
 Privacy Directive (EC Directive
 2002/58/EC) 402–405
 Software Defined Networking (SDN) 405
 legislation on user location privacy in 5G
 304–310
 see also regulation
 challenges 309–310
 Data Protection Directive (Directive 95/46/
 EC) 304–306, **309**, 402–404
 EU legal instruments **309**
 EU policy 304–306
 future scenarios 309–310
 General Data Protection Regulation
 (GDPR) (EU) 304, 305–307, 308–309
 international issues 308–309
 legal framework 304–306, 309–310
 location data processing 306
 policy 309–310
 Privacy Directive (EC Directive 2002/58/EC)
 305–308
 privacy protection by design and
 default 306–307
 security protection 307

- steps 310
- LiFi-5G networks, architectural design 160–161
- LiFi-5G networks interoperability 159–160
- application layer 159
 - MAC layer 159
 - network layer 159
 - transport layer 159
- LiFi- 5G security challenges 160
- LiFi networks, 5G-WLAN security 158–159
- Light Fidelity (LiFi) 160
- LiFi-5G networks
 - interoperability 159–160
 - LiFi- 5G security challenges 160
 - LiFi networks, 5G-WLAN security 158–159
- Limit State Design (LSD) principles, structures 171, 173
- literature overview, Internet of Things (IoT) 249–254
- Load and Resistance Factor Design (LRFD), structures 171, 173
- Location Based Services (LBS), privacy 271–272
- location-driven security, cyber security in 5G 113
- location privacy 271–272
- Log Analysis, monitoring technique 233
- LRFD (Load and Resistance Factor Design), structures 171, 173
- LSD (Limit State Design) principles, structures 171, 173
- LTE 21
- LTE-advanced E-UTRAN architecture, 4G cellular systems 24–25
- LTE authentication process 26, 26
- LTE network architecture 21–22
- LTE security model, 4G cellular systems 26–27
- m**
- M2M communication, 5G mobile networks 40, 41
- M2M gateway, cyber security in 5G 111
- machine learning for NFV-based security services, Network Function Virtualization (NFV) 369–370
- Machine Type Communication (MTC), 5G mobile networks 40, 41
- MAC layer
- 5G-WiFi interoperability 147
- LiFi-5G networks interoperability 159
- maintenance failures, structural design failures 183–186
- Management and Orchestration (MANO)
- 5G end-to-end system 47–48
 - Open Platform for NFV (OPNFV) 363–364
- mandated security in the network, 5G security 78
- Man-in-the-Middle (MitM), Multi-Access Edge Computing (MEC) 389
- MANO *see* Management and Orchestration
- massive connectivity, 5G mobile networks 35, 37
- Massive Machine-type Communications (mMTC)
- 5G mobile networks 33
 - cloud computing 381
- massive MIMO, 5G mobile networks 38–39
- MCR *see* Mobile Cloud Robot
- MEC *see* Mobile Edge Computing; Multi-Access Edge Computing
- MIMO (Multiple-Input Multiple-Output), 4G cellular systems 23
- MitM (Man-in-the-Middle), Multi-Access Edge Computing (MEC) 389
- mMTC (Massive Machine-type Communications), cloud computing 381
- mMTC (massive machine type communications), 5G mobile networks 33
- mmWave Communication, 5G mobile networks 38
- mobile broadband, Customer Edge Switching (CES) 224–225
- Mobile Cloud Robot (MCR) 249, 254, 258
- Internet of Things (IoT) 247–248, 259–263
 - security scenarios 259–263
- Mobile Edge Computing (MEC), 5G mobile networks 42–44
- mobile networks
- see also* 5G mobile networks
 - architecture 197, 198
 - challenges 200–201
 - global trust operator (GTO) 202–203
 - security framework principles 200–201
 - trust domains 202–203
 - trust processing 202–203

- mobile networks security
see also security
 OpenFlow 88–94
 robots 257–263
 Software Defined Networking (SDN)
 88–94
 mobile OS and app patch management, 5G
 mobile networks 71–72
 Mobile Virtual Network Operators (MVNO)
 5G mobile networks 46
 cloudification of the network
 operators 325–326, 334–335
 data security, Telecommunication network as a Service (TaaS) 327–328
 future directions 340–341
 hypervisor security, Telecommunication network as a Service (TaaS) 328–329
 Network Function Virtualization (NFV) 323–325, 327, 331–333, 334–335
 robots 263–265
 security 323–342
 security benchmark 336, 337–338
 Software Defined Networking (SDN)
 329–331
 Telecommunication network as a Service (TaaS) 331–332
 virtual machines (VM), security,
 Telecommunication network as a Service (TaaS) 328–329
 Mobile WiMAX 20
 monitoring ability, network slicing 195
 monitoring structures 171, 172
 monitoring techniques
see also Software Defined Monitoring
 Active Probing 232
 Device Polling 233
 existing 232–235
 Flow Collection 233
 improvements needed 235
 legacy monitoring systems 233–234, 238–240
 limitations 233–234, 238–240
 Log Analysis 233
 Netflow (sFlow) 232
 Packet Analysis 233
 Remote Monitoring (RMON) 232
 Security information and event management (SIEM) 233
 Simple Network Monitoring Protocol (SNMP) 232
 use of monitoring in 5G 234–235
 Multi-Access Edge Computing (MEC) 373–374
 5G networks 381–385
 architecture 391–392
 centralized security
 architectures 391–392
 cloud security research 392–394
 Denial of Service (DoS) attacks 389
 inconsistent security policies 389–390
 Man-in-the-Middle (MitM) 389
 overview 381–383
 privacy leakage 390
 SDN-based cloud security systems 392
 security architectures 391–392
 security challenges 388–390
 standardizations 392–394
 use cases 384–385
 VM manipulation 390
 multi-antenna technique 21
 Multiple-Input Multiple-Output (MIMO), 4G cellular systems 23
 multi-tenancy
 Network Function Virtualization (NFV) 354, 359–360
 XaaS models 359–360
 mutual authentication, WiFi-5G networks 157
 MVNO *see* Mobile Virtual Network Operators
- n**
- NAT *see* Network Address Translation
 national CERT centric trust domain, Customer Edge Switching (CES) 226
 National Institutes of Standards and Technology (NIST)
 cloud computing 406–407
 cloud security research 393
 standardizations 393
 Netflow (sFlow), monitoring technique 232
 network access security, 5G security 79
 Network Address Translation (NAT)
 Customer Edge Switching (CES) 197, 203–205
 cf. Customer Edge Switching (CES) 204
 state-of-the-art in mobile networks
 security 199–200

- network architecture
 4G cellular systems 24–25
 Groupe Spéciale Mobile (GSM) 9–10
 WiFi-5G networks
 interoperability 146–149
- Network Descriptor, Software Defined Monitoring (SDM) 241, 242
- Network Function Virtualization (NFV)
 5G mobile networks 44–46
 5G security 81–83, 231–232, 351
 5G-WiFi security 150
 competitive landscape 353–354
 distributed network security services 366
 European Telecommunications Standards Institute (ETSI) 348–350
 group-based policy (GBP) 367–368
 horizontal network slicing 354
 Infrastructure as a Service (IaaS) 354–355
 introduction to NFV 348–351
 machine learning for NFV-based security services 369–370
 Mobile Virtual Network Operators (MVNO) 323–325, 327, 331–333, 334–335
 multi-tenancy 354, 359–360
 network security as a service 366
 NFV-based network security 365–366
 NFV drivers 353–355
 One Cloud 355
 Open Platform for NFV (OPNFV) 327, 332–333, 334–335, 360–364
 OpenStack 240–241, 360–364
 OpenStack Congress 368–369
 opportunity cost 353–354
 Platform as a Service (PaaS) 354–355
 policy-based security services 366–367
 rapid delivery service 354
 security and privacy 407–408, 410–411, 417
 security services 347–370
 simple view 348–349
 Software as a Service (SaaS) 354–355
 Software Defined Networking (SDN) 351–353
 technology curve 353
 telco cloud 351–353
 Telecommunication network as a Service (TaaS) 323–333, 336
 Virtual Network Functions (VNF) 348–351
- virtual security appliances 365–366
 VNF security in operation 358–359, 365–366
 VNF security lifecycle 355–358
 VNF trust relationship 355–358
 XaaS models 354–355, 359–360
- network layer
 5G-WiFi interoperability 147–148
 WiFi-5G networks interoperability 159
- network resilience, Software Defined Networking (SDN) 91–92
- network security as a service, Network Function Virtualization (NFV) 366
- network security automation, Software Defined Networking (SDN) 92–94
- network security flaws, 5G-WiFi security 151
- network slicing 195
 5G end-to-end system 46–47
 horizontal network slicing, Network Function Virtualization (NFV) 354
 monitoring ability 195
- Neutron, OpenStack 362, 363
- new technologies
 regulation 400, 401, 413–417
 security and privacy 405–411, 413–417
- Next Generation Mobile Networks (NGMN)
 5G security 76, 78–79, 81–82
 5G Security Group, standardizations 393–394
 cloud security research 393–394
 NGMN Alliance white paper, 5G characteristics definitions 347–348
 trust models 276
- NFV *see* Network Function Virtualization
- NGMN *see* Next Generation Mobile Networks
- NIST *see* National Institutes of Standards and Technology
- non-cryptography-based security, 5G-WiFi security 150–151
- Nova, OpenStack 361, 362
- novel technologies, 5G security 81–87
- O**
- OFDM (Orthogonal Frequency Division Multiplexing) 21
- One Cloud, Network Function Virtualization (NFV) 355

- OpenFlow
 mobile networks security 83–85, 87, 88–94
 Software Defined Networking (SDN)
 83–85, 87, 88–94
- Open Platform for NFV (OPNFV)
 Management and Orchestration (MANO) 363–364
 Network Function Virtualization (NFV) 332–333, 334–335, 360–364
 Telecommunication network as a Service (TaaS) 327
- OpenStack
 Keystone 362, 363
 Network Function Virtualization (NFV) 240–241, 360–364
 Neutron 362, 363
 Nova 361, 362
 OpenStack Congress 368–369
 Software Defined Monitoring (SDM) 240–241
- OpenStack Congress
 Network Function Virtualization (NFV) 368–369
 OpenStack 368–369
- OPNFV *see* Open Platform for NFV
- Orthogonal Frequency Division Multiplexing (OFDM) 21
- outage probability
 jamming receiver 131–136
 physical layer security 131–136
 secrecy capacity 131–136
- outdoor vs. indoor technologies, 5G
 positioning 283, 284
- p**
- Packet Analysis, monitoring technique 233
 packet switched networks, 5G-WiFi
 security 154
- paint degradation, structural design
 failures 184–185
- passive positioning vs. active positioning, 5G
 positioning 283–285
- perimeter defence, cyber preparedness 105–106
- permanent action, structures 174, 179
- pervasive agility, cyber preparedness 105–106
- physical layer security 119–139
 encryption 119–120
 iJAM protocol 121–122
- innovative process 122
 Internet of Things (IoT) 120
 jamming 121
 jamming receiver 131–136
 motivation 121–123
 outage probability 131–136
 protocol for analysis 121
 related work 121
 secrecy capacity 131–136
 watermark-based blind physical layer security (WBPLSec) 121–131
- Platform as a Service (PaaS)
 cloud computing 378
 Network Function Virtualization (NFV) 354–355
 Telecommunication network as a Service (TaaS) 326, 336, 339, 340
- platform-driven security, cyber security
 in 5G 112–114
- policy-based security services, Network Function Virtualization (NFV) 366–367
- privacy 267–278
see also Data Protection Directive; General Data Protection Regulation (GDPR) (EU); legislation on user location privacy in 5G
 5G mobile networks 36–37, 37
 anonymity-based approaches 272
 Apps (smartphone applications) 270
 background 268–269
 cf. confidentiality 399–400
 data privacy 269–271
 device identity privacy 273
 European projects 311–312
 identity privacy 272–273
 International Mobile Subscriber Identity (IMSI) 268–269
 International projects 311–312
 Internet of Things (IoT) 270
 Location Based Services (LBS) 271–272
 location privacy 271–272
 subscriber identity privacy 272–273
 third party/service providers 270–271
 user privacy 269–273
- privacy cryptographic techniques, 5G
 positioning 299–304
- Privacy Directive (EC Directive 2002/58/EC)
 EU legal instruments 309
 legal framework for security and privacy 402–405

- legislation on user location privacy in 5G 305–308
- privacy leakage, Multi-Access Edge Computing (MEC) 390
- privacy-preserving location-based services 5G positioning 303–304
cryptographic techniques 303–304
- private cloud, cloud computing 377
- private data computation, cloud computing 386–387
- programmability, 5G mobile networks 36, 37
- public cloud, cloud computing 377–378
- r**
- radio access, cyber security in 5G 111
- Radio Access Network (RAN), 5G *see* 5G Radio Access Network
- Radio Frequency Identifier (RFID), security attacks 156
- radio interface keys, 5G security 78
- radio terminals, cyber security in 5G 111
- ransomware, 5G mobile networks 67, 68, **69**
- Realm Gateway (RGW)
address allocation 212
Bot-detection 213, 220–222
Customer Edge Switching (CES) 210–213
DNS (Domain Name System) floods 217–222
Domain Name System (DNS) 211–212
Domain Name System (DNS) abuse prevention 213
name server classification 212
security evaluation 217–222
security mechanisms 212–213
TCP-splice 213
- Received Signal Strength (RSS), 5G positioning 281, **284**, **286**, 288–289, 295–298
- regulation
see also legislation on user location privacy in 5G
5G security and privacy 399–417
consumer/market regulation **414–416**
government regulation **414–416**
industry self-regulation **414–416**
legal framework for security and privacy 402–405
new technologies 400, 401, 413–417
- objectives for security and privacy 401–402
regulatory approaches 412–413, **414–416**
relevance assessment 411–412
- relays, 4G cellular systems 24
- reliability, 5G mobile networks 35, 37
- Remote Monitoring (RMON), monitoring technique 232
- research
see also cloud security research
5G research communities 52–54
- research motivation, Internet of Things (IoT) 249–254
- response awareness, cyber preparedness 105–106
- RFID (Radio Frequency Identifier), security attacks 156
- RGW *see* Realm Gateway
- risk assessment/evaluation
ALARP (As Low As Reasonably Practicable) principle 190
structures 189–190
- RMON (Remote Monitoring), monitoring technique 232
- road traffic and transport, Industrial Internet (II) 227–228
- roaming, robots 263–265
- roaming carrier, cyber security in 5G 111
- robots
attack classification 255–256
attack scenarios 263, 264
communications 263, 265
data classification 254
Mobile Cloud Robot (MCR) 247–248, 249, 254, 258, 259–263
mobile network security 257–263
Mobile Virtual Network Operators (MVNO) 263–265
roaming 263–265
security platform 256–259
SIMcard robots 259, 260
SIMless robots 260–263
threats classification 255–256
- RSS *see* Received Signal Strength
- s**
- SC-FDE (Single Carrier Frequency Equalization) 21
- SC-FDMA (multiple version of SC-FDE) 21

- SCH (Supplemental Code Channel), code division multiple access (CDMA) 10
- SDM *see* Software Defined Monitoring
- SDN *see* Software Defined Networking
- seamless mobility, 5G mobile networks 35, 37
- secrecy capacity
- iJAM protocol 129–131
 - jamming receiver 131–136
 - outage probability 131–136
 - physical layer security 131–136
 - simulation scenario 134–136
 - watermark-based blind physical layer security (WBPLSec) 128–129
- secrecy metrics, watermark-based blind physical layer security (WBPLSec) 126–127
- secure data computation, cloud computing 386–387
- secure device management, 5G mobile networks 71
- security 59–73
- 1G cellular systems 7–8, 60, 61
 - 2G cellular systems 10–15, 60, 62–63
 - 3G cellular systems 17–20, 60, 63
 - 4G cellular systems 25–27, 61, 63–66
 - 5G mobile networks 36–37, 37, 61, 66–73
 - 5G security 75–95
 - breaches 59
 - CDMA 2000 17–18
 - Groupe Spéciale Mobile (GSM) 11–14
 - IS-95 14–15
 - mobile networks 59–73
 - UMTS network architecture 18–20
- security architectures
- cloud computing 391–392
 - Multi-Access Edge Computing (MEC) 391–392
- security challenges, Software Defined Networking (SDN) 84–86
- security cryptographic techniques, 5G positioning 299–304
- security failures, key, 5G mobile networks 70–71
- security framework, Customer Edge Switching (CES) 203–213
- security framework principles, mobile networks 200–201
- Security information and event management (SIEM), monitoring technique 233
- security lifecycle functions, 5G mobile networks 70–73
- security monitoring, 5G mobile networks 72–73
- security systems, Software Defined Networking (SDN) 92
- security threat analysis
- 4G cellular systems 65–66
 - 5G mobile networks 69–70
- security threat analysis and assessment, 5G mobile networks 72
- serviceability limit states (SLS), structures 181
- sFlow (Netflow), monitoring technique 232
- shield effect, structural design failures 183
- short messaging service (SMS), 2G cellular systems 8
- SIEM (Security information and event management), monitoring technique 233
- SIM *see* subscriber identity module
- Simple Network Monitoring Protocol (SNMP), monitoring technique 232
- simulation scenario, secrecy capacity 134–136
- Single Carrier Frequency Equalization (SC-FDE) 21
- SLS (serviceability limit states), structures 181
- small cell technology, 5G mobile networks 39–40
- SMS (short messaging service), 2G cellular systems 8
- SNMP (Simple Network Monitoring Protocol), monitoring technique 232
- Software as a Service (SaaS)
- cloud computing 378
 - Network Function Virtualization (NFV) 354–355
 - Telecommunication network as a Service (TaaS) 326, 336, 339, 340
- Software Defined Monitoring (SDM) 231–242
- 5G networks 231–242
 - advantages 238–240
 - architecture 235–238
 - challenges 240–242
 - interfaces 237
 - vs. legacy monitoring systems 238–240
 - modules 236–237

- Network Descriptor 241, 242
- OpenStack 240–241
- Software Defined Networking (SDN)
 - 5G mobile networks 44
 - 5G security 81–82
 - access control 91
 - application layer 84–85
 - application plane security 86–87
 - cloud computing 392
 - control channels security 89–91
 - control layer 85–86
 - control plane security 87
 - Customer Edge Switching (CES) 200, 223, 229
 - data link security 88–89
 - data plane security 87
 - firewalls 92
 - infrastructure layer 86
 - Intrusion Detection Systems (IDS) 91
 - Intrusion Prevention Systems (IPS) 91
 - layer-based SDN architecture 329–331
 - legal framework for security and privacy 405
 - mobile networks 88–94
 - Mobile Virtual Network Operators (MVNO) 329–331
 - Network Function Virtualization (NFV) 351–353
 - network resilience 91–92
 - network security automation 92–94
 - OpenFlow 83–85, 87, 88–94
 - SDN-based cloud security systems 392
 - security, Telecommunication network as a Service (TaaS) 329–331
 - security and privacy 409–411, 417
 - security challenges 84–86
 - security solutions 86–87
 - security systems 92
 - telco cloud 351–353
 - Telecommunication network as a Service (TaaS) 329–331
 - traffic monitoring 91
 - Transport Layer Security (TLS) 86, 89, 90
 - South Korea: 5G Forum, 5G research 53–54
 - spectrum, 5G mobile networks 32
 - spectrum efficiency, 5G mobile networks 36, 37
 - spectrum management, 4G cellular systems 24
- standardization activities, 5G *see* 5G standardization activities
- standardizations
 - cloud computing 392–394
 - European Telecommunications Standards Institute (ETSI) 392–393
 - Multi-Access Edge Computing (MEC) 392–394
 - National Institutes of Standards and Technology (NIST) 393
 - Next Generation Mobile Networks (NGMN) 5G Security Group 393–394
- standards
 - European Telecommunications Standards Institute (ETSI) 48, 50–51
 - structures 188–189
- state-of-the-art in mobile networks
 - security 197–203
 - Network Address Translation (NAT) 199–200
- steel corrosion, structural design
 - failures 184–185
- steel design verifications, structures 180–181
- structural design 165–190
 - see also* architectural design
- structural design failures 166, 182–187
 - cleaver cracks 184–185
 - fatigue 183, 185
 - ice 183
 - maintenance failures 183–186
 - paint degradation 184–185
 - shield effect 183
 - steel corrosion 184–185
 - structures 182–187
 - water traps 185
 - weld imperfections 185, 185–186
 - wind issues 182–183
- structures 165–190
 - accidental action 174, 179
 - actions 174–179
 - ancillary items 178
 - Codes of Practice 187, 188
 - Deaves and Harris model 175
 - design codes 170
 - design philosophy 171–181
 - environmental aspects 169
 - Eurocodes 170, 173, 180
 - examples 167

- structures (*cont'd*)
 Fibre Bragg Gratings (FBG) 171, 190
 finite element analyses (FEA)
 179–180
 geometrical and material nonlinear
 imperfect analysis (GMNIA) 179
 historical development 168–171
 infrastructure security, cyber security in
 5G 109–110
 Limit State Design (LSD) principles
 171, 173
 Load and Resistance Factor Design
 (LRFD) 171, 173
 LRFD (Load and Resistance Factor
 Design) 171, 173
 LSD (Limit State Design) principles
 171, 173
 monitoring 171, 172
 opportunities 188–190
 outlook 170–171
 performance-based design philosophy
 181–182
 permanent action 174, 179
 problems 181–187
 protection 182
 recommendations 188–190
 risk assessment/evaluation 189–190
 serviceability limit states (SLS) 181
 standards 188–189
 steel design verifications 180–181
 structural analysis 179–180
 terrorism 186–187
 typology 167, 168–169
 ultimate limit states (ULS) 180–181
 vandalism 186–187
 variable action 174, 179
 wind issues 166, 174–178
 subscriber identity module (SIM)
 2G cellular systems 11
 SIMcard robots 259, 260
 SIMless robots 260–263
 Universal Subscriber Identity Module
 (USIM), identity issues 274, 275–277
 subscriber identity privacy 272–273
 subscriber level security policies, 5G
 security 78–79
 Supplemental Code Channel (SCH), code
 division multiple access (CDMA) 10
 system architecture, 5G mobile networks 32
- t**
 TaaS *see* Telecommunication network as
 a Service
 TCP-splice, Realm Gateway (RGW) 213
 TDOA *see* Time Difference of Arrival
 technologies, enabling, 5G mobile
 networks 37–44
 technologies, new *see* new technologies
 telco cloud
 Network Function Virtualization
 (NFV) 351–353
 Software Defined Networking
 (SDN) 351–353
 Telecommunication network as a
 Service (TaaS)
 application security 333
 cloudification of the network
 operators 325–326
 data security 327–328
 deployment security 338–340
 hypervisor security 328–329
 Infrastructure as a Service (IaaS) 326,
 336, 337–338, 339, 340
 Network Function Virtualization
 (NFV) 323–333, 336
 Platform as a Service (PaaS) 326, 336,
 339, 340
 security benchmark 336, 337–338
 security classification 326–327
 Software as a Service (SaaS) 326, 336,
 339, 340
 Software Defined Networking (SDN)
 329–331
 virtual machines (VM) security 328–329
 Temporary Mobile Subscriber Identity
 (TMSI), Groupe Spéciale Mobile
 (GSM) 11
 terrorism, structural failure 186–187
 third party/service providers,
 privacy 270–271
 Time Difference of Arrival (TDOA), 5G
 positioning 281, 284, 284, 286, 286,
 287, 296, 298
 Time of Arrival (TOA), 5G positioning 281,
 284, 286–289, 286, 296–297
 TLS *see* Transport Layer Security
 TMSI (Temporary Mobile Subscriber
 Identity), Groupe Spéciale Mobile
 (GSM) 11

- TOA *see* Time of Arrival
- Total Access Communication Systems (ETACS), 1G cellular systems 7
- traffic monitoring, Software Defined Networking (SDN) 91
- transport layer
- 5G-WiFi interoperability 148–149
 - LiFi-5G networks interoperability 159
- Transport Layer Security (TLS), Software Defined Networking (SDN) 86, 89, 90
- trust domains, mobile networks 202–203
- trust issues 267–278
- background 268–269
 - International Mobile Subscriber Identity (IMSI) 268–269
 - Next Generation Mobile Networks (NGMN) 276
 - threats, potential 274, 275, 277
 - trust models 274–277
 - Virtual Mobile Network Operator (VMNO) 274–276
- trust processing
- global trust operator (GTO) 202–203
 - mobile networks 202–203
- trustworthiness enhancement, 5G
- positioning 299
- u**
- ultimate limit states (ULS), structures 180–181
- ultra-dense small cells, 5G mobile
- networks 39–40
- ultra low latency, 5G mobile
- networks 34–35, 37
- ultra-reliable and low latency communications (URLLC), 5G mobile networks 33
- Ultra-reliable Machine-type Communications (uMTC), cloud computing 381
- UMTS network architecture
- 3G cellular systems 16
 - vs. LTE network architecture 21–22
 - security 18–20
- UMTS Terrestrial Radio Access Network (UTRAN) 16
- UMTS WCDMA, 3G cellular systems 15–16
- Universal Subscriber Identity Module (USIM),
- identity issues 274, 275–277
- URLLC (ultra-reliable and low latency communications), 5G mobile
- networks 33
- user identity confidentiality, WiFi-5G
- networks 156
- user plane integrity, 5G security 78
- user privacy 269–273
- USIM (Universal Subscriber Identity Module),
- identity issues 274, 275–277
- v**
- vandalism, structural failure 186–187
- VANET, security attacks 155
- variable action, structures 174, 179
- virtualization security, cloud computing 385–386
- virtual machines (VM), security
- Telecommunication network as a Service (TaaS) 328–329
- Virtual Mobile Network Operator (VMNO),
- trust models 274–276
- Virtual Network Functions (VNF)
- 5G mobile networks 44–46
 - 5G security 81–83
 - Network Function Virtualization (NFV) 348–351
 - VNF security in operation 358–359, 365–366
 - VNF security lifecycle 355–358
 - VNF trust relationship 355–358
- virtual security appliances, Network Function Virtualization (NFV) 365–366
- VM *see* virtual machines
- VM manipulation, Multi-Access Edge Computing (MEC) 390
- VMNO (Virtual Mobile Network Operator),
- trust models 274–276
- VNF *see* Virtual Network Functions
- w**
- watermark-based blind physical layer security (WBPLSec)
- 5G networks 136–138
 - data decomposition 122, 124
 - cf. iJAM protocol 130–131, 139
 - jamming channel 123–124
 - jamming receiver 126
 - physical layer security 121–131
 - secrecy capacity 128–129
 - secrecy metrics 126–127
 - system model 123–131
 - transmitter 124–126

- water traps, structural design failures 185
WBPLSec *see* watermark-based blind physical layer security
- weld imperfections, structural design failures 185, 185–186
- WiFi (Wireless Fidelity) 143–144
 interoperability with 5G networks 144
 security 144–146
- WiFi-5G networks
 architectural design 156–157, 160–161
 data integrity 156–157
 device identity confidentiality 156
 key management 157
 mutual authentication 157
 user identity confidentiality 156
- WiFi-5G networks interoperability
 5G-WLAN security 143–149
 network architecture 146–149
- WiMAX, Mobile 20
- WiMAX security
 4G cellular systems 27
 Mobile WiMAX 20
 security attacks 155
- wind issues
 structural design failures 182–183
 structures 166, 174–178
- Wireless mesh Networks (WMNs), security attacks 155
- Wireless sensor Networks (WSNs), security attacks 155
- WMNs (Wireless mesh Networks), security attacks 155
- Worldwide Interoperability for Microwave Access (WiMAX) 20
 4G security 27
 Mobile WiMAX 20
 security attacks 155
- WSNs (Wireless sensor Networks), security attacks 155

X

- XaaS models
 multi-tenancy 359–360
 Network Function Virtualization (NFV) 354–355, 359–360
- xMBB (Extreme Mobile Broadband), cloud computing 380–381

Z

- zero-day attacks, 5G mobile networks **70**

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley's ebook EULA.