UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA DELL'INFORMAZIONE

"Tecniche ultra-low-latency per le reti 5G: scenari applicativi e sfide tecnologiche"

Relatore: Prof. Leonardo Badia

Laureando: Leonardo Murari

ANNO ACCADEMICO 2020 – 2021

Data di laurea 15/11/2021

Abstract

Le reti di telecomunicazioni sono diventate, nel corso del tempo, una parte indispensabile dell'infrastruttura della nostra società e inoltre, con le reti di quinta generazione, l'ambizione è quella di ampliare ulteriormente la qualità e le tipologie di applicazioni con l'introduzione di tre servizi chiave: enhanced mobile broadband (eMBB), massive machine type communications (mMTC) e ultra-reliable and low latency communications (URLLC). Con il presente lavoro di tesi si intende fornire anzitutto una visuale del percorso compiuto, a partire dalla prima generazione fino ad oggi, nello sviluppo dei network e delle tecnologie di comunicazione. In seguito, si vuole dare uno sguardo al futuro, focalizzando l'attenzione sulle URLLC, presentando una serie di innovativi scenari d'utilizzo e applicazioni mission-critical che fanno direttamente affidamento su tale tipo di comunicazioni. Infine, guardando alla situazione presente, verranno investigate le varie fonti di ritardo end-to-end delle attuali reti wireless 4G Long Term Evolution. Conseguentemente verranno proposte alcune soluzioni tecnologiche, sia riguardanti l'architettura di rete, sia riguardanti varie tecniche per la riduzione della latenza, implementabili a livello protocollare, che permettano il soddisfacimento dei requisiti previsti per le URLLC.

Indice

Introduzione

Nel corso dei passati decenni, l'evoluzione dei sistemi di telecomunicazioni e la loro introduzione nel mercato è stata sostanzialmente guidata dalla necessità di migliorare la qualità dei servizi mobili a banda larga [1]. Tali servizi sono generalmente limitati all'utilizzo di smartphones e tablets, i quali generano la maggior parte del traffico dati. La rete Internet futura invece è chiamata a soddisfare le esigenze derivanti sia dai consumatori, sia dai vari settori dell'industria, oltre a dover supportare un elevatissimo e sempre crescente numero di dispositivi connessi. Con lo sviluppo dei sistemi di quinta generazione, perciò, l'obiettivo è quello di creare una tecnologia unica in grado di supportare unitariamente un'ampissima gamma di servizi e applicazioni, ciascuna avente requisiti differenti. Per fare ciò, come verrà spiegato nel seguito, il 5G si propone di introdurre tre servizi determinanti, ovvero una maggiorata larghezza di banda, la connettività Massive Internet of Things e la tipologia di comunicazioni Ultra Reliable Low Latency Communications (URLLC) [2]. In questo scritto verrà approfondito proprio quest'ultimo servizio, la cui ambizione è quella di supportare livelli di latenza e affidabilità senza precedenti. I requisiti in tale ambito, delineati dal Third Generation Partnership Project (3GPP) [3] e dal International Telecommunication Union (ITU) [4], richiedono infatti che la trasmissione di un pacchetto dati di 32 bytes avvenga correttamente almeno il 99.999% delle volte, con una latenza a livello di User Plane inferiore a 1 ms.

Il resto di questa tesi è organizzato come segue: nel Capitolo 2, con la finalità di dare al lettore una visione di insieme, verrà presentata la storia dello sviluppo dei vari sistemi di telefonia mobile. Per ciascuna generazione, perciò, verranno descritte le caratteristiche principali, fino ad arrivare all'avanguardia delle reti 5G, per le quali verranno presentati i requisiti caratterizzanti e le maggiori novità introdotte. Nel Capitolo 3, focalizzando l'attenzione sulla tipologia URLLC, verrà dato uno sguardo al futuro attraverso la descrizione di alcuni scenari applicativi e servizi facenti riferimento a tale tipo di comunicazioni. Tali casi di utilizzo, come i trasporti intelligenti e l'assistenza medica remota, saranno in grado di modificare profondamente vari settori industriali oltre che il panorama digitale sociale [5]. A seguire, nel Capitolo 4, per comprendere appieno le sfide poste dall'abilitazione di trasmissioni a latenza ultra-bassa, verranno anzitutto indagate quelle che sono le maggiori fonti di ritardo, a livello di Core Network e di accesso wireless, prendendo come riferimento una attuale rete 4G Long Term Evolution. Successivamente, quindi, verranno introdotte alcune soluzioni tecnologiche capaci di fare fronte al problema della minimizzazione dei ritardi di trasmissione. Tra queste,

come soluzione strutturale, si parlerà del paradigma di architettura di rete orientata al calcolo del Fog Computing, e in seguito di alcune altre tecniche e approcci, implementabili a livello protocollare in tale tipologia di rete, riportando inoltre alcuni risultati numerici che ne comprovino la validità.

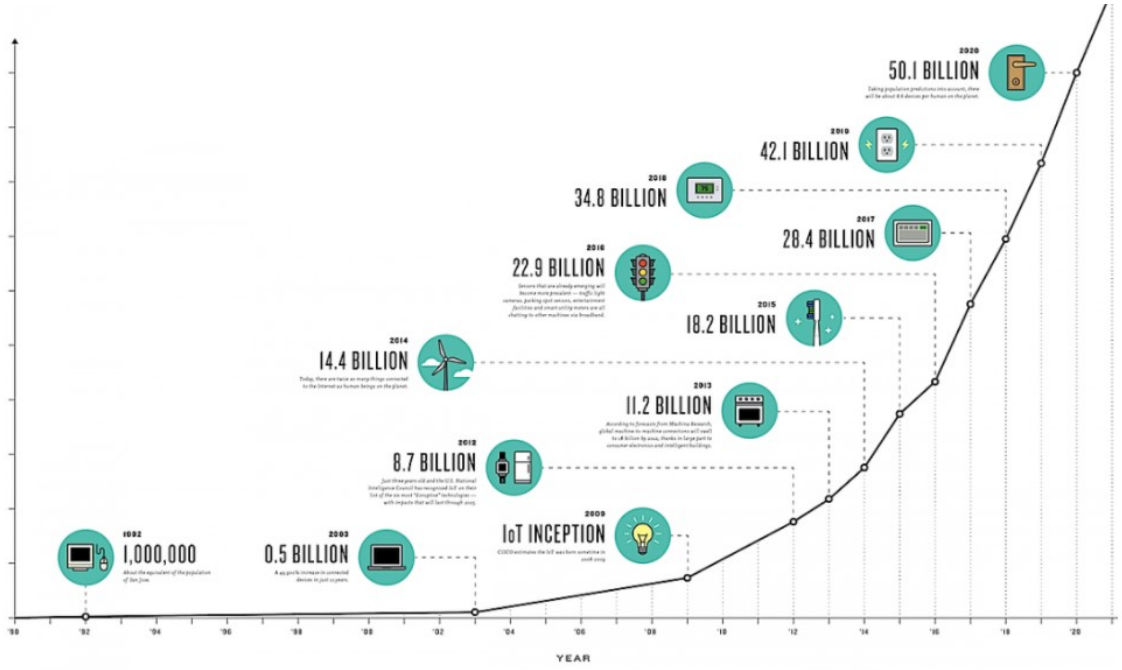


Figura 1: andamento nel tempo dal 1990 al 2020 del numero di dispositivi connessi ad Internet

1.1 Sistemi di prima generazione (1G)

Con il termine 1G si intende il primo standard di trasmissione di telefonia mobile, dunque la prima generazione di reti mobili. I sistemi cellulari di prima generazione erano quasi tutti composti da sistemi analogici e utilizzavano la tecnologia di modulazione di frequenza per la trasmissione radio. Nel 1979 la Nippol Telephone and Telegraph, a Tokyo, introdusse il primo sistema di rete cellulare nel mondo e, solo due anni dopo, l'epoca delle reti mobili era ormai avviata anche in Europa. I due protocolli principali di questa generazione erano il Nordic Mobile Telephones e il Total Access Communication Systems. Negli Stati Uniti, invece, fu lanciato nel 1982 l'Advanced Mobile Phone System (AMPS), che divenne presto il protocollo più popolare. Le reti cellulari di allora, tuttavia, non erano in grado di interagire tra loro in diversi continenti poiché impiegavano formati di segnalazione e controllo differenti e quindi incompatibili. Tipicamente si trattavano di apparati destinati a servire definite aree geografiche, aventi capacità di servizio molto limitate (alcune decine di canali con l'interruzione del servizio quando si usciva dall'area di copertura della Base Station Transceiver (BTS)). Poiché il numero massimo di Mobile Station (MS) attive all'interno di ogni area di copertura era limitato al numero dei canali (cioè delle frequenze) assegnate al servizio, ne risultava di conseguenza anche una bassa densità geografica di utenti. Gli apparati di prima generazione non garantivano grandi prestazioni nemmeno in termini di efficienza energetica o spettrale. Questi sistemi infatti prevedevano la suddivisione del territorio in zone geografiche autonome, ciascuna delle quali coperta da una BTS con l'impiego di tutte le frequenze assegnate al servizio. L'area di copertura della BTS doveva essere la più estesa possibile, e ciò implicava l'impiego di potenze di trasmissione adeguatamente elevate. La tecnica di accesso al mezzo, ovvero alla banda o porzione di banda di frequenze allocate al servizio radiomobile, era di tipo Frequency Division Multiple Access (FDMA).

1.2 Sistemi di seconda generazione (2G)

1.2.1 Standard GSM

Il 2G venne lanciato commercialmente nel 1991 in Finlandia con lo standard GSM (Global System for Mobile Communications) [6] le cui specifiche vennero pubblicate l'anno precedente

ad opera del gruppo francese Group Speciale Mobile in collaborazione con l'autorevole European Telecommunications Standards Institute (ETSI). Tale standard divenne ben presto il più diffuso. L'introduzione del GSM ha rappresentato una vera e propria rivoluzione nell'ambito dei sistemi di telefonia mobile, differenziandosi dalla prima generazione per numerosi aspetti. Anzitutto questa tecnologia offre interoperabilità tra sistemi, la quale ha permesso la stipulazione di accordi tra operatori di diversi paesi per l'effettuazione del roaming internazionale: era cioè possibile effettuare chiamate da un continente all'altro arginando così uno dei limiti principali dell'1G. Un'altra differenza fondamentale è rappresentata dal fatto che il 2G prevede una trasmissione di tipo digitale, la quale porta con sé importanti vantaggi e conseguenze. Tale tecnologia, infatti, aumenta la velocità di trasferimento fino a un data rate standard di 9.6 kbps, derivante dalle tecniche di compressione dati proprie della codifica di sorgente, migliorando allo stesso tempo l'affidabilità della trasmissione. Inoltre, permette di implementare funzioni di sicurezza in termini di cifratura della comunicazione. Il miglioramento della bit-rate a sua volta rende possibili nuovi e più ampi servizi quali il fax digitale, la posta elettronica, il trasferimento di chiamata, la teleconferenza e il servizio short message service (SMS) per l'invio di messaggi di testo.

Per il duplexing il GSM [7] utilizza la divisione in frequenza (frequency division duplex, FDD) attraverso due bande di 25 MHz, quella che va da 890 a 915 MHz usata per le trasmissioni in uplink, mentre quella che va da 935 a 960 MHz utilizzata per le trasmissioni in downlink. In seguito, tali bande verranno estese prima a 35 MHz, poi a 75 MHz dando luogo al cosiddetto sistema GMS *dual band*. Come protocollo di accesso radio invece viene impiegata una combinazione delle tecniche TDMA (Time Division Multiple Access) e FDMA (Frequency Division Multiple Access). Le due bande disponibili, cioè, sono suddivise in sottobande di 200 kHz ciascuna, ed ogni sottobanda viene utilizzata a divisione di tempo con 8 slot per trama. Inoltre, in alcune reti, per fare fronte ad attenuazioni rilevanti dovute alla presenza di notch nella risposta in frequenza del canale, si utilizza la tecnica del frequency-hopping, che consiste nel variare la frequenza portante della trasmissione in modo dinamico.

1.2.2 Standard GPRS ed EDGE

Nel corso degli anni, la tecnologia GSM venne continuamente perfezionata allo scopo di apportare sempre migliori servizi. In combinazione all'eccezionale crescita della telefonia mobile cellulare e di utenze Internet, ciò portò allo sviluppo di sistemi più avanzati che

prendono il nome di 2.5G. In particolare, associabile a quella che è l'evoluzione del GSM, è il General Packet Radio Service (GPRS) [8].

Il GPRS è una tecnologia radio che introduce la nuova classe di servizi basati sulla comunicazione wireless di dati, espandendo le funzionalità del GSM grazie all'introduzione della commutazione di pacchetto e della possibilità di accedere alla rete Internet. In particolare, il GPRS supporta le reti basate sia sull' Internet Protocol (IP) che sul protocollo X.25. Tra in vari servizi introdotti ci sono la messaggistica MMS (Multimedia Message Service), chiamate di gruppo, chiamate multicast e la connessione wireless a siti Web secondo il protocollo di accesso WAP (Wireless Application Protocol). Mentre i servizi voce vengono trasportati sfruttando l'architettura classica a commutazione di circuito, i dati sfruttano nuovi nodi di rete e la commutazione di pacchetto appena introdotta. Questi nuovi nodi che si aggiungono all'architettura precedente si definiscono GPRS Support Node (GSN) [9] e si dividono in due tipi:

- Serving GSN: responsabili della trasmissione dei pacchetti da e verso le MS svolgendo funzioni di gestione della mobilità e controllo degli accessi;
- Gateway GSN: funge da interfaccia tra backbone GPRS e reti dati a pacchetto esterne.

Il GPRS prevede quattro codifiche di canale: CS-1, CS-2, CS-3, CS-4 con data rate che vanno rispettivamente da 9.05 kbps a 21.4 kbps per slot. La scelta di una di queste dipende dalla quantità di rumore presente al trasmettitore: per esempio, se il canale è molto disturbato verrà selezionata la CS-1 mentre se è in condizioni perfette verrà adottata la CS-4. Anche se il numero massimo di time slot accorpabili per un utente è otto (arrivando così ad un data rate teorico di 171.2 kbps utilizzando il Code Scheme 4), è in realtà difficile che un singolo utente riesca ad ottenere tutti e 8 gli slot disponibili, per cui un valore più realistico si attesta tra 30-70 kbps. Ciò è dipeso anche da fattori quali il numero degli utenti connessi, la congestione della rete e il tipo di terminale GPRS utilizzato.

Prima dell'introduzione del 3G è stato sviluppato il sistema EDGE (Enhanced Data rate for GSM Evolution) [10]. Considerato come 2.75G si presta ad essere uno slancio verso la terza generazione offrendo servizi avanzati simili ma senza doverne implementare le tecnologie. Infatti, per aggiornare una cella, è necessaria solo l'installazione di una ricetrasmittente EDGE e un aggiornamento software remoto ricevibile da stazioni di base. La tecnologia EDGE consente maggiori velocità di trasferimento dati arrivando fino a 473.6 kbps teorici grazie all'introduzione della modulazione 8-PSK piuttosto che la normale GMSK utilizzata da GSM e GPRS. La larghezza di banda e la tipologia di accesso radio TDMA restano invariate.

1.3 Sistemi di terza generazione (3G)

Le tecnologie di terza generazione nascono con l'intento di unificare i vari tipi di media su un unico sistema di telecomunicazione che sia in grado di gestirli in modo flessibile, adattandosi alle esigenze dell'utente. Il 3G offre velocità di trasmissione più elevate con la possibilità di usufruire di molteplici servizi in contemporanea (ad esempio, comunicare al telefono mentre si scarica la posta elettronica). Vengono infatti supportati contemporaneamente i due tipi di connessione a commutazione di circuito e di pacchetto. La banda di utilizzo individuata dal World Administrative Radio Conference (WARC), è quella dei 2 GHz. Come verrà spiegato in seguito, queste frequenze vengono utilizzate senza limitazioni in Europa e in Asia con tecniche di accesso CDMA mentre negli Stati Uniti tali risorse radio devono essere condivise dal sistema Personal Communication Service (PCS) che già le utilizzava.

Le tecnologie di terza generazione sono basate sulle specifiche ITM-2000 stabilite dall'ITU (International Telecommunication Union). Originariamente pensato per essere uno standard unico, unificato a livello globale, il 3G è in realtà stato implementato in due standard principali:

- Universal Mobile Telecommunications System (UMTS), utilizzato in Europa, Asia e Giappone;
- cdma2000, sviluppato negli USA come evoluzione dello standard CDMA IS-95 sviluppato in ambito 2G, e per questo detto anche 3G IS-95.

1.3.1 Standard UMTS

L'UMTS [7] viene standardizzato in Europa dal 3GPP, mentre un altro gruppo, noto come 3GPP2, è stato costituito per sviluppare le specifiche globali per reti 3G, e quindi lo sviluppo di un Global Third Generation (G3G), in modo da consentire l'interoperabilità con le reti di tutto il mondo, non solo con quelle di nuova generazione, ma anche con i sistemi precedenti. I terminali UMTS, infatti, supportano anche sistemi 2G consentendo una maggiore mobilità tra reti attraverso semplici procedure di handover e ottimizzando l'utilizzo della copertura radio.

Le principali differenze con i sistemi di seconda generazione e rispettive fasi superiori (2.5 e 2.75 G), risiedono principalmente nella nuova tipologia di interfaccia radio: la tecnica wideband CDMA (W-CDMA) anziché TDMA/FDMA. L'introduzione di tale tecnologia consente di realizzare delle trasmissioni con efficienza spettrale maggiorata, generalmente più veloci e soprattutto differenziate in base al contesto in cui si trova ciascun utente: 144 kbps in condizioni

di mobilità veicolare (fino a 500 km/h), connessioni a 384 kbps in condizioni di mobilità pedestre, per arrivare fino a 2 Mbps in condizioni di ridotta mobilità (e.g. a casa o in ufficio). È dunque possibile offrire un'ampiezza di banda diversa a seconda del contesto. Altre differenze importanti con la seconda generazione sono il supporto di traffico asimmetrico sulle tratte di uplink e downlink, il multiplexing su un'unica connessione di servizi con requisiti di qualità diversi come voce e trasferimento dati. È inoltre presente la coesistenza tra le modalità di duplexing FDD e TDD. Alle sezioni di rete UMTS facenti uso della prima sono state allocate le bande di frequenza accoppiate di 1920÷1980 MHz per l'uplink, e 2110÷2180 MHz per il downlink, mentre per la seconda le bande non accoppiate da 1900 ÷ 1920 MHz e 2010 ÷ 2025 MHz. In entrambi i casi le bande sopra riportate vengono divise in portanti da 5 MHz.

L'accesso all'interfaccia radio per utenti mobili è fornito dalla rete di accesso detta UTRAN (UMTS Terrestrial Radio Access Network) [11]. L'UTRAN è composta da un gruppo di sottosistemi di rete radio (radio network subsystem, RNS), costituiti da un controllore di rete (radio network controller RNC) e da un insieme di stazioni radio base ricetrasmittenti chiamati Node-B. Ciascun Node-B gestisce una o più celle supportando trasmissioni in modalità sia FDD che TDD. La Core Network (CN) dell'UMTS invece è basata sull'architettura GSM aggiornata a GPRS e EDGE. Tali apparati, tuttavia, devono essere aggiornati per il funzionamento di servizi UMTS. I suoi elementi principali sono: il registro dei residenti Home Location Register (HLR) e quello dei visitatori, il Visitor Location Register (VLR), poi, per servizi a commutazione di circuito, il Mobile Switching Center (MSC) e per i servizi a commutazione di pacchetto, i serving GPRS support node (SGSN).

La classificazione delle reti UMTS è passata da 3G a 3.5G mediante l'introduzione della famiglia di protocolli High Speed Packet Access (HSPA) [12] che ne potenzia le prestazioni. L' HSPA comprende due protocolli, lo High Speed Downlink Packet Access (HSDPA) che aumenta la velocità di trasmissione in downlink a 14.4 Mbps e lo High Speed Uplink Packet Access (HSUPA) che viceversa migliora la velocità in uplink fino a 5.76 Mbps. In seguito, l'HSPA viene migliorato con la 3GPP Release 7 nell'HSPA Evolution (HSPA+) in grado di raggiungere una bit-rate fino a 42 Mbps. Le tecnologie introdotte includono:

- Adaptive Modulation and Coding (AMC), cioè gli schemi di modulazione e codifica sono adattati dinamicamente rispetto alla qualità della connessione radio;
- meccanismo di HARQ (Hybrid Automatic Repeat Request) al livello LLC (Logical Link Control) che riduce il ritardo e aumenta l'efficienza di ritrasmissione;
- scheduling e ritrasmissioni veloci tramite l'elaborazione della richiesta direttamente dal Node-B;

- breve lunghezza frame per accelerare ulteriormente lo scheduling dei pacchetti per la trasmissione.

1.4 Sistemi di quarta generazione (4G)

Prima di poter parlare di una effettiva quarta generazione mobile, è necessario passare per quelle che sono state le tecnologie di transizione dal 3G al 4G, sulle quali poi si basano i nuovi sistemi. Tali tecnologie comprendono essenzialmente:

- Long-Term Evolution (LTE);
- Mobile WiMAX (Worldwide Interoperability for Microwave Access);

Tali standard sono definiti, appunto, "di transizione" o in alcuni casi 3.9 G poiché ancora non soddisfano le caratteristiche IMT-Advanced (International Mobile Telecomunications - Advanced) per sistemi 4G, definite da ITU. Il raggiungimento di questi requisiti avverrà con gli standard denominati appunto "Advanced".

1.4.1 LTE e LTE Advanced

L'LTE [13] è uno standard basato su GSM/EDGE e UMTS/HSPA, sviluppato dal 3GPP le cui specifiche sono state rilasciate nel Release 8 con alcuni minori miglioramenti nel Release 9. LTE si propone di incrementare la capacità e la velocità di trasmissione ad accesso wireless utilizzando nuove tecniche di elaborazione e modulazione digitale dei segnali. Un ulteriore obbiettivo è quello di ridisegnare e semplificare l'architettura di rete passando dal UMTS network a commutazione di pacchetto e di circuito ad un sistema basato completamente su protocollo IP per ridurre la latenza. Tale architettura è costituita dalla Evolved Packet Core (EPC), disegnata per prendere il posto della GPRS Core Network, e dalla Evolved UTRAN (E-UTRAN) che rappresenta l'interfaccia radio dell'LTE. Le sub componenti della EPC sono:

- MME (Mobility Management Entity): rappresenta il nodo di controllo chiave nell'accesso al network;
- SGW (Serving Gateway): svolge compiti riguardanti l'instradamento dei pacchetti dati da e verso l'eNB per servire lo UE;
- PGW (Packet Data Network Gateway): interfaccia le reti dati a pacchetto esterne e svolge funzioni IP come assegnazione di indirizzo e l'applicazione delle policy;

- HSS (Home Subscriber Server): rappresenta l'archivio centrale contenente le informazioni degli abbonati;

La E-UTRAN invece comprende una sola tipologia di nodo fisico: gli evolved Node-B (eNB). Essi comunicano direttamente con lo User Equipment (UE), inoltre sono connessi ai nodi di packet switching via interfaccia S1, e tra di loro secondo interfaccia X2. Rispetto ai semplici NB, non necessitano di un elemento di controllo separato e sono in grado di gestire il quadruplo dei dati.

Come per le reti UMTS, anche in questo caso vengono supportate entrambe le modalità di duplexing FDD e TDD, tuttavia le bande di frequenza impiegate sono differenti. In particolare, lo spettro di frequenze utilizzato è spezzettato in diversi intervalli. Ciò è dovuto alla necessità di convivere con differenti sistemi di comunicazione radio. Le bande di frequenze che vengono adoperate sono: 800 MHz, 850 MHz, 1800 MHz, 1900 MHz, 2100 MHz, 2600 MHz. A differenza dei sistemi precedenti in cui la banda veniva suddivisa in sottobande di larghezza fissa, LTE introduce una applicabilità flessibile all'utilizzo delle risorse radio riservando ad un utente un minimo di 1.25 MHz fino ad un massimo di 20 MHz di banda con la possibilità di variarlo nel tempo a seconda delle necessità. Altri due aspetti per cui LTE si differenzia dai suoi predecessori, e che gli consentono di raggiungere una efficienza spettrale tre volte superiori alla più evoluta versione dell'UMTS è:

1. l'utilizzo della modulazione OFDM (Orthogonal Frequency-Division Multiplexing) per le trasmissioni in download, che permette di mantenere un ottimo rapporto tra data-rate (fino a 326.4 Mbps) e robustezza alle interferenze, e l'utilizzo di Single-Carrier FDMA (SC-FDMA) per le trasmissioni in uplink raggiungendo un data-rate di 84.4 Mbps e permettendo di avere un rapporto picco/potenza media ridotto, quindi un minor consumo di potenza.
2. l'introduzione di sistemi Multiple Input Multiple Output (MIMO) attraverso l'utilizzo di antenne multiple, risolvendo così il problema relativo ai cammini multipli percorsi da un segnale, dovuto alla presenza di oggetti che ne provocano la riflessione. Apparecchiature MIMO permettono quindi di combinare tra loro i vari segnali ricevuti sfruttando il principio di interferenza costruttiva.

L'evoluzione dell'LTE, ovvero l'LTE Advanced, le cui caratteristiche tecniche sono state rilasciate principalmente nel 3GPP Release 10, è considerato il vero passo all'interno dei sistemi 4G. L'LTE Advanced riprende integralmente la precedente versione, migliorandone alcuni

aspetti seppur mantenendosi completamente compatibile con le sue apparecchiature e condividendone appieno le bande di frequenza. Le caratteristiche principali che implementa possono essere riassunte nei punti seguenti:

- carrier aggregation tramite la quale è possibile concatenare tra loro bande di frequenza contigue e non contigue per aumentare le prestazioni di picco della rete;
- larghezza di banda scalabile oltre i 20 MHz fino a 100 MHz;
- assegnazione asimmetrica della banda oltre che per la modalità TDD anche per FDD;
- accesso al mezzo ibrido con OFDMA e SC-FDMA in uplink;
- trasmissione e ricezione attraverso operazione di coordinazione multipunto (Coordinated Multi-Point, CoMP) grazie alla quale è possibile utilizzare celle vicine per trasmettere lo stesso segnale della cella servitrice, nel caso di perdita di robustezza del segnale;
- coordinazione MIMO tra eNB in entrambi uplink e downlink;
- utilizzo nella modulazione di segnale di 128-QAM (Quadrature Amplitude Modulation), la 256-QAM, inoltre, è stata inclusa nel Release 12;

Grazie a tutte queste migliorie, l'LTE-Advanced ha raggiunto le specifiche IMT per la quarta generazione di bit rate, in condizioni ottimali, di oltre 1 Gbps in downlink e 500 Mbps in uplink.

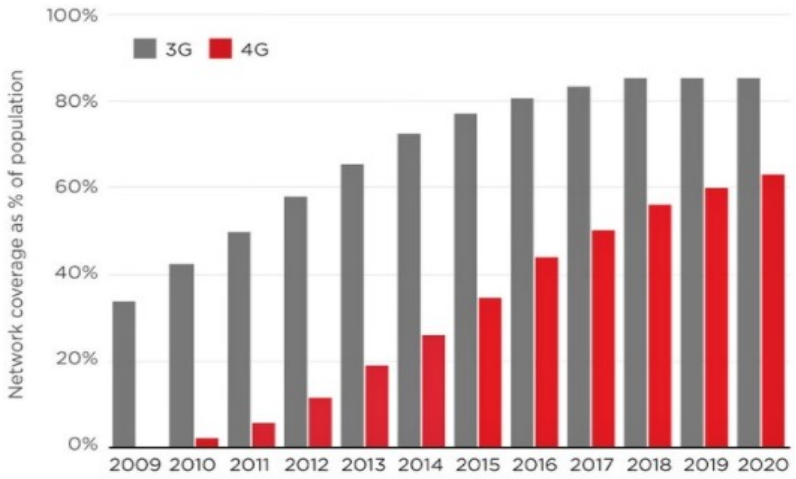
1.4.2 WiMAX e WiMAX 2

WiMAX [14] è una famiglia di standard di comunicazione wireless a banda larga, ratificata dal consorzio WiMAX Forum e basata originalmente su WirelessMAN (Metropolitan Access Network) standard IEEE 802.16 del 2001. L'obiettivo era quello di creare un sistema ad alta interoperabilità, che sappia supportare simultaneamente un grande numero di utenti con alte velocità di trasmissione (fino a 40 Mbps), e capace di fornire copertura per un ampio raggio (50 km) da ciascuna stazione base. La prima versione (802.16) era indirizzata all'uso del range di frequenze 10-66 GHz, successivamente con l'estensione 802.16a la banda operativa è stata ampliata anche al range 2-11GHz. Ciò è legato al fatto che inizialmente WiMAX [15] era sviluppato per applicazioni Line-of-Sight (LoS), per le quali sono adatte alte frequenze, mentre con l'introduzione di relativamente più basse frequenze venivano rese possibili anche comunicazioni Non LoS (NLoS), anche se ancora in modalità ad accesso fisso. Il supporto per la mobilità è stato introdotto con l'aggiornamento 802.16e nel dicembre 2005 il quale prevede stabilità di connessione fino a 122 km/h.

L’architettura di rete minima consiste nella presenza di due tipi di dispositivi fissi: una Base Station (BS) ed almeno una Subscriber Station (SS). Le SS hanno il compito di inoltrare alle BS il traffico derivante dai terminali utenti a cui sono connessi. Le BS invece, distribuite in modo strategico per evitare lacune di copertura, si connettono da una parte con la rete fissa o con un’altra BS e dall’altra con le SS della loro cella. Esse, oltre a fungere da gateway per la connessione degli utenti alla rete, gestiscono i compiti relativi all’interfaccia radio tra cui link establishment, radio resource management e session management. Spesso, inoltre, per incrementare le performance del sistema in termini di margine di attenuazione, in particolare in condizioni NLoS, è previsto l’utilizzo di BS con antenne intelligenti (AAS, Adaptive Antenna System) le quali conformano la loro area di copertura alla posizione delle CPE (Customer Premises Equipment). Viene implementata la tecnologia MIMO e controllo d’errore HARQ. In questo caso la modalità di duplexing utilizzata è la TDD, la FDD invece verrà introdotta in WiMAX 2. Come in LTE, invece, la trasmissione si basa sulla tecnica OFDM con TDMA supportando diverse ampiezze di canale sia in uplink che in downlink fino a un massimo di 20 MHz. Lo schema di modulazione e codifica può comunque essere adattato dinamicamente in conformità alle condizioni del canale.

Lo sviluppo di WiMAX 2, evoluzione di WiMAX con la pretesa di entrare a far parte a tutti gli effetti della quarta generazione in conformità alle specifiche richieste da ITU, si basa sull’ aggiornamento alla versione 16m nel 2010 compiuto all’interno della comunità IEEE 802.16. L’aspettativa di WiMAX 2 è infatti quella di fornire più di 1 Gbps in mobilità fissa e oltre 100 Mbps in mobilità fino a 500 km/h. Ovviamente viene mantenuta la completa compatibilità con la versione precedente. 802.16m introduce tecnologie spesso simili a LTE Advanced, come ad esempio l’aggregation carrier per larghezze di banda di oltre 20 MHz, una migliorata efficienza spettrale (utilizzando sempre la modulazione di tipo OFDM) e subframes più brevi per ridurre la latenza.

In generale le valutazioni delle prestazioni dei due standard (WiMAX e LTE nella loro versione avanzata), per la somiglianza nelle tecnologie e nelle risorse impiegate, presentano simili performance. Nonostante questo, e nonostante WiMAX si sia sviluppato più velocemente di LTE, le maggiori compagnie di telecomunicazioni hanno scelto di basare i loro servizi 4G sul network LTE, il quale offre maggiore continuità e un più naturale sviluppo per le reti GSM/UMTS/HSPA maggiormente in uso. Nel 2020 infatti, più di un quarto di tutta l'utenza mobile utilizzava LTE e per la fine del 2021 si prevede l'aumento fino ad oltre la metà [16].



Mobile broadband coverage reach, 2009–2020
Source: GSMA Intelligence

1.5 Sistemi di quinta generazione (5G)

Il 5G rappresenta la nuova frontiera negli standard di comunicazione mobile, il suo sviluppo e definizione sono ancora in corso ad opera del 3GPP, il quale ne definisce caratteristiche, protocolli e metodi di funzionamento attraverso le varie Release, ad iniziare dalla 15 fino all'ultima più recente Release 18. Il 5G non si propone di essere una semplice evoluzione della precedente rete mobile con stesse caratteristiche e maggiori prestazioni ottenibili a costi inferiori, ma al contrario punta ad essere una piattaforma di rete innovativa capace di reinventare l'industria di innumerevoli settori con nuovi scenari, casi d'utilizzo, modelli di business e servizi innovativi, aprendo le porte verso una più ampia immagine di ciò che viene definito "ecosistema digitale". La continua crescita della domanda di throughput e capacità di trasmissione maggiore per il miglioramento dei servizi a banda larga sono solo uno degli obbiettivi che hanno incoraggiato e che trasportano lo sviluppo del 5G. In aggiunta, infatti, la nuova generazione ha l'obiettivo di abilitare tre tecnologie fondamentali [17]:

- *massive Machine Type Communications (mMTC):* tali comunicazioni sono caratterizzate da una totale automatizzazione nella generazione, nello scambio e nella processazione dei dati tra macchine intelligenti, senza quindi intervento umano. Ciò viene definito machine-centric anziché human-centric. Con la rapida diffusione di dispositivi con sistemi embedded, mMTC diventa un paradigma di comunicazione dominante per un ampissimo raggio di servizi, inclusa l'assistenza medica, i trasporti, o l'industria manifatturiera.
- *Ultra Reliable Low Latency Communications (uRLLC):* si tratta di una nuova categoria di comunicazioni che serve ad abilitare tutta quella serie di servizi che necessitano di requisiti molto stringenti nella latenza. Per sua natura, l'URLLC si differenzia dall'approccio tradizionale utility-based del network che fa affidamento a quantità statistiche medie (e.g. throughput medio, ritardo medio, tempo medio di risposta), imponendo dei valori che non sono più un'opzione ma diventano una necessità. Gli URLLCs si focalizzano su applicazioni che richiedono una consegna end-to-end di dati in modo sicuro, affidabile e veloce.
- *enhanced Mobile Broadband (eMBB):* una maggiorata larghezza di banda si assume l'obiettivo di soddisfare un crescente stile di vita digitale degli utenti che fanno uso di servizi e funzioni che implicano elevate capacità di trasmissione. eMBB si concentra quindi al supporto dell'aumento di data rate e capacità di sistema richiesti dagli utenti finali. Per fare ciò, eMBB introduce due fondamentali miglioramenti tecnologici: uno shift dello spettro di frequenze utilizzate nelle onde centimetriche e millimetriche (cmWave e mmWave), e array di decine o addirittura centinaia di antenne ricetrasmittenti che abilitano sistemi massivi MIMO (massive MIMO) e beamforming.

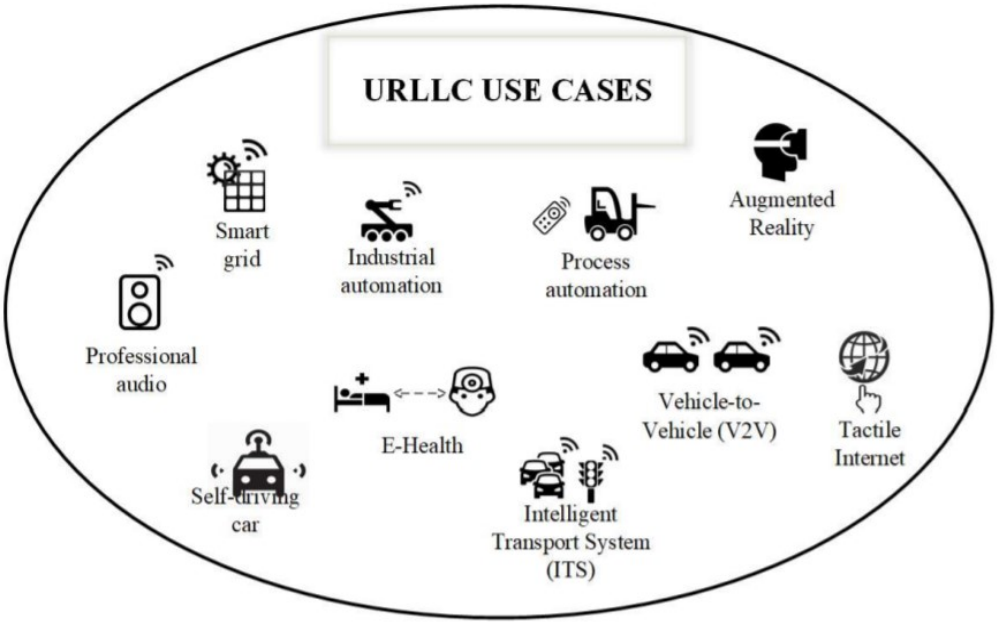
In particolare, per queste macro-tecnologie si è resa necessaria, oltre a una riconfigurazione dell'esistente EPC e di una nuova Core Network 5G, anche, e soprattutto, una nuova rete di interfaccia radio chiamata New Radio 5G. Queste sono state studiate e definite da parte dell'International Telecommunication Union Radiocommunications Standardization Sector (ITU-R) insieme al 3GPP, i quali ne hanno definito le specifiche tecniche presentate nella Tabella 1 [18].

KPI	**Value**
Peak Data Rate	20 Gbps DL – 10 Gbps UL
Peak Spectral Efficiency	30 bps/Hz DL - 15 bps/Hz UL
Bandwidth	400 MHz
Control Plane Latency	10 ms
User Plane Latency	0.5 ms URLLC - 4ms eMBB
Reliability	1-10−5 for 32 bytes con 1ms UL
Coverage	mMTC 164 dB con 160bps
Battery Life	>10 years (mMTC)
Connection Density	106 devices/Km2
Mobility	500 Km/h
Mobility Interruption Time	0 ms

Tabella 1: specifiche tecniche per la 5G NR

2. Scenari applicativi di URLLC

Con le nuove tecnologie e possibilità introdotte dal 5G, esso rappresenta l'ingrediente fondamentale per realizzare scenari e casi d'uso futuristici per qualsiasi settore e tipologia di industria. Gli scenari applicativi sono numerosissimi e alcuni sono già in stato di realizzazione. Di seguito verranno descritti cinque scenari [1], considerati tra i più ampi e di maggiore impatto, che fanno direttamente riferimento all'adempimento dei requisiti di latenza caratteristici delle reti di nuova generazione.



2.1 E-Health

Le reti di comunicazione in generale, e le reti mobili in particolare, si possono considerare la chiave per l'attuazione degli obbiettivi che l'industria medica sta cercando di raggiungere: soluzioni cloud-based che migliorino l'accessibilità di dati medici dettagliati, l'incremento della capacità per la trasmissione video in tempo reale ad alta definizione, il supporto per grandi numeri di dispositivi smart indossabili (e-health wearables) connessi e naturalmente comunicazioni a latenza ultra bassa. A tali fini, sono due le principali applicazioni considerabili in contesto 5G:

i) assistenza sanitaria da remoto e medicina di alto livello con l'utilizzo di bio-connettività;
ii) interventi di chirurgia a distanza con l'utilizzo di robot chirurgici controllabili da remoto.

Nel contesto della bio-connettività, esiste una tendenza verso la decentralizzazione degli ospedali per cui l'assistenza medica può essere fornita anche a casa o in viaggio (i.e. provvedimenti di emergenza in ambulanza), grazie a cartelle mediche elettroniche, analisi di dati per predire e prevenire patologie e all'utilizzo di sistemi incorporati per l'analisi farmaceutica. Nel contesto della chirurgia da remoto, l'obiettivo è quello di abbattere l'ostacolo dell'individuazione geografica dell'equipe medica per essere in grado di offrire assistenza di alto livello, anche per interventi e operazioni chirurgiche complesse. Questa rivoluzione dell'industria medica è in via di realizzazione tramite l'integrazione della connettività nel settore, e si può osservare nei seguenti strade percorribili:

- sviluppo di sistemi integrati che combinino documentazione clinica con differenti metodi di comunicazione (telefonico, radiofonico, satellitare, ecc.), e l'assistenza medica da remoto;
- reindirizzamento di costose visite mediche in ospedale con l'utilizzo della telemedicina;
- l'utilizzo di dispositivi indossabili che apportino vantaggi come la gestione e il monitoraggio di pazienti cronici o la possibilità di consultazione e collaborazione nel prendere decisioni;

I casi di utilizzo sopra menzionati possono chiaramente essere suddivisi tra quelli che dipendono fortemente dalla latenza del servizio a quelli che non ne risentono, o ne risentono in maniera inferiore. In casi, ad esempio, di tele-surgery, in cui l'intero trattamento viene eseguito da remoto, il network deve essere in grado di supportare una trasmissione tempestiva e affidabile di streaming audio e video. Inoltre, i requisiti sulla latenza si fanno ancora più stringenti nel caso di apparecchiature con feedback aptico, implementato da una serie di sensori situati sull'equipaggiamento chirurgico, grazie al quale il chirurgo è in grado di percepire quello che le braccia robotiche stanno toccando. I dispositivi cinestetici, infatti, lavorano in sistemi di controllo ad anello chiuso, per cui le due estremità (azione e reazione) devono operare in perfetta sincronia l'uno con l'altro. La tolleranza al ritardo in questo modo si riduce, necessitando di Round Trip Times (RTT) inferiori a 1ms. In termini di errore in trasmissione poi, è ovvia la necessità di disporre di sistemi estremamente affidabili con Block Error Rate inferiori a 10e(-9) poiché un qualsiasi piccolo errore potrebbe portare a risultati catastrofici.

2.2 Guida assistita e servizi di trasporto

Seguendo la definizione in [19] il termine Intelligent Transport Systems (ITS) fa riferimento all'utilizzo di tecnologie nell'ambito dell'Informazione, come sensori e sistemi di comunicazione wireless, in applicazioni riguardanti i trasporti con l'obbiettivo di offrire viaggi e spostamenti più fluidi ed efficienti, in entrambi gli ambienti sia pubblico che privato. In particolare, il mercato automobilistico si sta spostando verso un futuro in cui le macchine saranno fortemente connesse tra loro e alla rete, rendendo così possibili nuove esperienze d'utilizzo come la guida autonoma o assistita. Sarà così possibile incrementare la sicurezza, ridurre i consumi e il congestionamento stradale. Inoltre, l'utilizzo di servizi internet tradizionali, come la riproduzione di video o musica in streaming, permetterà di estendere le tipiche applicazioni da smartphone anche all'interno della vettura. Sono già in fase di sviluppo e ricerca numerose applicazioni, tra cui le più rappresentative sono [20]:

- guida autonoma, per la quale non è richiesto alcun intervento da parte dell'uomo. Le auto dovranno beneficiare di informazioni locali trasmesse dai veicoli circostanti e da infrastrutture stradali, le quali permetteranno di adattarsi e reagire a situazioni di traffico complesse. A questo contesto appartengono le applicazioni più popolari di automated overtake, cooperative collision avoidance e high density platooning;
- servizi di sicurezza ed efficienza stradale. Questi fanno affidamento al fatto che le auto connesse inviano e ricevono costantemente informazioni (di stato o di eventi circostanti) alle infrastrutture, agli altri veicoli e al network. Applicazioni possibili sono l'individuazione di Vulnerable Road User (VRU) come pedoni e ciclisti, oppure la conoscenza di ciò che succede oltre i veicoli contigui (applicazioni see-trough);
- digitalizzazione e logistica dei trasporti, le quali permettono di collezionare informazioni sul traffico ed utilizzarle in modo esteso per ottimizzare l'utilizzo delle strade in base alla scelta di vari itinerari, secondo vari criteri;
- sistemi di navigazione intelligente con realtà aumentata e real-time video per informazioni sul traffico;
- servizi di informazione e intrattenimento (i.e. infotainment), anche durante la marcia;
- l'incremento della capacità e della copertura della rete considerando le auto parcheggiate come nodi "portatili".

È chiaro che, di fronte a questi possibili scenari futuri in ambito automobilistico, la chiave abilitante è una massiccia comunicazione mobile. Una delle sfide maggiori, infatti, è l'enorme numero di dispositivi che sarà necessario connettere costantemente alla rete per lo scambio di

informazioni, principalmente di tipo Machine Type Communication (MTC), poiché sia i veicoli che le infrastrutture saranno disseminate di sensori e attuatori.

Visto il grande numero di applicazioni differenti nel settore, è naturale prendere in considerazione livelli di latenza specifici per ciascun caso. Per esempio, nel caso di sistemi di automated overtaking, è richiesta una tolleranza al ritardo end-to-end massima di approssimativamente 10 ms per ogni messaggio scambiato. Nel caso invece di applicazioni see-through con video real-time da 30 frames per secondo, queste possono essere supportate con capacità di trasmissione di 220 Mbps e latenza end to end di 50ms. In tali scenari inoltre è importante che il sistema di comunicazione si faccia carico di evitare un sovraccarico di MTC, visto l'elevato numero di segnali di controllo che vengono generati da tale tipo di comunicazioni anche per piccole quantità di dati. Nel caso della mobilità, che è certamente una protagonista negli scenari di connettività futura, oltre ad abilitare di per sé comunicazioni URLLC, è necessario anche ricercare soluzioni in termini di handovers e di ricerca preventiva del collegamento tra veicoli/dispositivi e le infrastrutture di interfaccia alla rete durante il movimento. Inoltre, nel caso di perdita di copertura, fare in modo di ristabilire rapidamente la comunicazione. Per fare ciò sarà necessario permettere una facile cooperazione tra molteplici tecnologie radio e, tra queste, un utilizzo efficiente delle risorse.

2.3 Intrattenimento

Il settore dell'intrattenimento ha subìto, e sta ancora subendo, grandi trasformazioni causate principalmente dal cambiamento delle abitudini e del comportamento degli individui. Le persone, infatti, non sono più solo meri consumatori ma interagiscono attivamente durante la fruizione di contenuti di intrattenimento e media. La modifica di tale industria è iniziata con l'incremento della disponibilità di reti ad alte velocità di trasmissione, sia per connessioni internet mobili che fisse. Ciò ha contribuito, congiuntamente ai data centers e al cloud computing, a far crescere la domanda di esperienze di qualità ed immersività maggiore. È ormai chiaro, infatti, che il consumatore in generale è molto influenzato dalle capacità dei dispositivi insieme alla disponibilità di servizi innovativi offerti dai vari content creator. Per fare un esempio, le persone preferiscono guardare video in streaming on demand piuttosto che trasmissioni programmate in televisione.

Le applicazioni specifiche che necessitano di comunicazioni URLLC possono essere classificate come segue:

- gaming multigiocatore
- utilizzo di sistemi di realtà aumentata con adattamento intelligente all'ambiente circostante
- esperienze multisensoriali super immersive e ad alta fedeltà tramite sistemi, ad esempio, di realtà virtuale

Con le esperienze full-immersion di intrattenimento, con particolare riferimento al gaming, si sono introdotti scenari in cui ciò che accade deve essere riprodotto simultaneamente a tutti gli utenti attivi, rendendo l'esperienza di gioco super immersiva. Tali applicazioni inoltre sono state arricchite di risoluzioni grafiche sempre crescenti per ottenere un livello di realismo il più alto possibile. Attualmente, gli sforzi maggiori dell'industria del gaming sono concentrati nel miglioramento dell'esperienza di gioco tramite l'inclusione di elementi di Virtual Reality o Augmented Reality (VR, AR) e di bio-sensing, i quali permetteranno ai giocatori di individuare persone all'interno del gioco in mondi reali o immaginari; tali elementi inoltre aggiungono la capacità di interagire con oggetti virtuali circostanti attraverso sistemi di riconoscimento dei movimenti, restituendo feedback di forza realistici. In tali applicazioni, una rete URLLC capace di provvedere all'integrazione di servizi multisensoriali attraverso video, audio e feedback tattili potrà migliorare ulteriormente l'esperienza d'utente per entrambi i settori di content delivery e gaming. Alcune delle sfide tecnologiche più importanti necessarie all'abilitazione di questi servizi sono estremamente simili ad altri casi di utilizzo. È necessario cioè garantire un Quality of Service (QoS) adeguato in termini di data-rate, latenza end-to-end, copertura e affidabilità. Ad ogni modo le applicazioni che richiedono i requisiti più stringenti sono relative sicuramente a VR e AR, per le quali sono necessari ritardi massimi consentiti tra i 7ms e 15ms per avere un'esperienza fluida e realistica.

2.4 Automazione industriale

L'introduzione del 5G è la chiave abilitante di quella che viene chiamata Industry 4.0. Con tale termine si indica la visione della Commissione Europea per la ristrutturazione e il reinventamento dei processi di produzione industriali d'Europa. È presente, infatti, una forte tendenza da parte dell'Unione Europea verso la digitalizzazione dell'industria, affiancando i processi di sviluppo, di produzione e la logistica a network intelligenti. In tal modo le organizzazioni europee hanno individuato la strada per giungere alla quarta rivoluzione industriale. In questo senso, le "fabbriche del futuro" (Factory of Future, FoF) non saranno delle entità chiuse e autonome ma faranno parte di un più ampio ecosistema.

Nel passato, i compiti di monitoraggio sono stati eseguiti con l'utilizzo di reti di sensori wireless (Wireless Sensor Network, WSN), le quali interconnettono un dato numero di sensori in modo che eseguano operazioni di rilevamento e controllo. L'evoluzione di queste WSN ha contribuito molto allo sviluppo dell'IoT nell'ambito delle comunicazioni mobili, e al giorno d'oggi si sono raggiunte applicazioni industriali che includono molto più del semplice monitoraggio e tracciamento, e che possono integrare tutte quelle capacità introdotte dalla digitalizzazione delle FoF. In questo contesto, le applicazioni possibili spaziano da operazioni time-critical al controllo remoto delle attrezzature di fabbrica:

- applicazioni di controllo da remoto basate su realtà aumentata per fornire supporto in produzione e manutenzione
- ottimizzazione e controllo di processi time-critical: ricezioni istantanee di informazioni dai sistemi di monitoraggio, controllo remoto di operazioni robotiche, apparati di robot collaborativi operanti in sistemi di controllo ad anello chiuso. Tale famiglia di applicazioni è caratterizzata dalla necessità di latenze che possono essere inferiori ad 1ms;
- comunicazioni non time-critical, comprendenti applicazioni come controllo di qualità, localizzazione e raccolta dati dai sensori;
- comunicazioni fluide tra diversi apparati lungo la catena di produzione.

Per supportare tali applicazioni, come il coordinamento e l'intervento in tempo reale, la rete deve essere in grado di offrire scenari di connettività multipla ed eterogenea, in cui ogni dispositivo sia capace di comunicare efficacemente anche negli scenari industriali più complessi. Inoltre, diventa necessaria una riconfigurazione del controllo del QoS e delle domande di traffico in modo che questi siano sufficientemente veloci ed affidabili da permettere un rapido adattamento della rete alle necessità che si presentano. In merito al supporto dei requisiti di ritardo della rete, oltre alla latenza stringente, anche il jitter, ovvero la fluttuazione del ritardo nella trasmissione dei pacchetti, deve essere contenuto. Alcune applicazioni, poi, prevedono grandi trasmissioni di dati, come ad esempio video 3D o contenuti per AR, altre applicazioni invece interessano piccole trasmissioni di dati da molti sensori differenti. Allo stesso modo, un requisito fondamentale per tutti questi casi di utilizzo è la gestione efficiente delle comunicazioni MCT.

2.5 Smart Grids

L'accessibilità a energia sostenibile e sicura dovrà essere il paradigma di sviluppo dei futuri sistemi energetici con l'obiettivo di preservare il pianeta e i suoi abitanti. L'elettrificazione degli apparati industriali, di mobilità, cittadini o, più in generale, l'elettrificazione dei consumi energetici è senza dubbio un passo fondamentale da attuare per raggiungere l'ambizioso obiettivo di carbon neutrality nel 2050 previsto del Green Deal del marzo 2020 presentato dalla Commissione Europea. Per fare ciò, l'utilizzo diffuso e distribuito delle fonti di energia rinnovabile e meccanismi di integrazione intelligente nelle reti elettriche, sono degli strumenti fondamentali. Per poter permettere l'utilizzo in modo efficacie ed efficiente di tali fonti energetiche, per loro natura non programmabili, sarà quindi essenziale l'ammodernamento delle reti elettriche in ottica Smart Grids. Si tratta, in accordo con la definizione della IEA (International Energy Agency), di un sistema di reti elettriche che utilizza la tecnologia digitale per monitorare e gestire il trasporto di elettricità da tutte le fonti di generazione per soddisfare le diverse richieste di energia elettrica degli utenti finali. In questo modo sarà possibile [21]:

- consentire un utilizzo coordinato ed efficiente delle risorse distribuite consentendo una effettiva valorizzazione delle stesse e una minimizzazione dei costi;
- abilita l'integrazione di quote crescenti di fonti rinnovabili con benefici evidenti in termini di impatto ambientale, apportando inoltre una maggior creazione di valore per tutta la filiera energetica;
- rende più flessibile il rapporto produttore-consumatore aprendo a nuovi meccanismi di offerta basati sul coinvolgimento attivo dell'utente (consumatore e prosumer) e favorendo alla formazione di comunità energetiche sia locali che diffuse;
- massimizza l'affidabilità, la stabilità e la resilienza della rete.

Con le Smart Grids viene richiesto alle tradizionali reti elettriche monodirezionali di funzionare come reti bidirezionali, passando da elementi passivi di distribuzione e trasporto ad elementi attivi in grado di ripatire e ridistribuire l'energia in modo intelligente in base alla necessità. Per fare ciò, si presenta una discreta varietà di requisiti in latenza, che possono andare da RTT di 2-3ms a 20ms, dipendenti da diversi fattori quali la tipologia di griglia, l'architettura della rete e la tecnologia impiegata o l'applicazione di riferimento. Il controllo remoto delle luci stradali o dei semafori, ad esempio, ha requisiti più laschi rispetto all'ottimizzazione del voltaggio della tensione in un apparato di automazione industriale o alla telelettura e telegestione dei contatori di energia.

3. Soluzioni tecnologiche

Nei precedenti capitoli è stata data una visione generale del percorso tecnologico evolutosi fino ad oggi e si è dato uno sguardo ad alcuni dei numerosissimi scenari che la nuova generazione ha l'ambizione di rendere accessibili. In questo capitolo, invece, si guarderà alla situazione presente attraverso una analisi delle fonti di latenza dell'attuale network e verranno presentate delle possibili soluzioni che siano in grado abilitare, dal punto di vista della latenza, i casi d'uso introdotti. In particolare, verrà descritto anzitutto il paradigma di architettura di rete orientato al calcolo del Fog Computing, e in seguito verranno presentate alcune altre tecnologie, implementabili all'interno di questo modello architetturale.

3.1 Componenti di ritardo nelle reti attuali

Le reti cellulari sono sistemi complessi, comprensivi di una molteplicità di livelli e rispettivi protocolli necessari al corretto funzionamento. Tali protocolli e le loro interazioni, rappresentano delle significative fonti di ritardo per la trasmissione dei pacchetti dati all'interno del network. La latenza, inoltre, dipende anche da numerosi altri parametri, quali la distanza e il mezzo rappresentato dal canale tra trasmettitore e ricevitore, l'architettura di rete, la tipologia di tecnologia wireless utilizzata come anche il numero di utenti attivi connessi. Le componenti della latenza nelle reti 4G standard LTE sono state attentamente analizzate e quantificate dal 3GPP nella Release 14 [22], sia per trasmissioni in uplink che in downlink. In particolare, le componenti di ritardo più rilevanti, derivanti da protocolli e algoritmi di accesso radio, dallo user equipment al gateway (uplink) e ritorno (downlink) sono:

- *Grant Acquisition Process:* una volta che lo UE ha creato i dati e li ha "pacchettizzati", questo è pronto per trasmetterli alla base station, o all'eNB. Per fare ciò, prima viene inviato un messaggio di Scheduling Request (SR) all'eNB, il quale deve rispondere all'utente generando lo Scheduling Grant (SG), (cioè il messaggio di controllo che conferma la possibilità di ricevere la trasmissione), infine la trasmissione dei pacchetti può avere inizio. È importante notare che la SR può essere inviato solo una volta che l'utente è connesso e allineato alla base station, ovvero è presente un SR-valid Physical Uplink Control Channel (PRCCH). Una volta che lo UE riceve e decodifica lo SG, la trasmissione dei pacchetti ha luogo invece nel Physical Uplink Shared Channel (PUSCH). In questa modalità di gestione dell'allocazione delle risorse, la limitazione

principale consta nel fatto che lo UE deve attendere la disponibilità non solo del canale di trasmissione dei dati effettivi, ma anche del PRCCH, e ciò dipende dalla sua periodicità. In base alle funzionalità LTE del Release 8, il tempo medio di attesa per il PRCCH con periodo di10 ms è di 5 ms.

- *Random Access Procedure:* a differenza del Grant Acquisition (GA) Process, questa procedura si applica agli utenti che non sono ancora connessi e allineati alla base station. Non essendoci il canale di controllo PRCCH, lo UE inizia un processo di accesso random al canale (Random Access Channel, RACH) che funge da uplink GA, conosciuto come random access-based SR. Nella procedura totale di RACH si includono i tempi di trasmissione, detection e processing da ambo i lati trasmettitore e ricevitore per un totale di 9.5 ms di ritardo.
- *Trasmission Time Interval (TTI):* qualsiasi trasmissione di pacchetti, che siano di informazione o controllo, viene eseguita in subframe di 1 ms di durata. Questa è, dunque, l'unità minima di trasmissione in LTE.
- *Signal Processing:* si tratta del tempo impiegato per la processazione dei pacchetti (ad esempio per la codifica e decodifica). In generale tale valore si attesta sui 3 ms.
- *Packet Retransmission:* nel caso di uplink Hybrid Automatic Repeat Request (HARQ), il Round Trip Time (RTT) è di 8 ms (nel caso di frequency division duplex). Nel caso di downlink non è direttamente specificato poiché lo schema è asincrono.

Ci sono poi una serie di altre rilevanti componenti di ritardo causate da diversi fattori provenienti dalla trasmissione dei pacchetti all'interno del Core Network/ Internet. Tali componenti sono difficilmente quantificabili per via della loro ampia variabilità e sono:

- *Propagation Delay:* ovvero il ritardo di propagazione del segnale, dipendente dal mezzo fisico e dalle distanze in gioco.
- *Serialization Delay:* ovvero il tempo necessario per incanalare una unità di informazione, come un pacchetto dati, all'interno del canale, come ad esempio un cavo in fibra ottica. Tale ritardo è dipendente dalla quantità di informazioni da "serializzare". Nelle correnti reti ad alta velocità a banda larga, si tratta di quantità dell'ordine del microsecondo.
- *Protocol Delay:* protocolli orientati alla connessione, come ad esempio il Transmission Control Protocol (TCP), possono incrementare i tempi di ritardo end-to-end con meccanismi di ritrasmissione e congestion avoidance.
- *Switches & Routers:* nodi di routing e switching ad alte prestazioni posso aggiungere ritardi che, in generale, rappresentano approssimativamente il 5% della latenza end-to-end totale.

- *Queuing & Buffer Management:* l'accodamento dei pacchetti, derivante da situazioni di congestionamento della rete, posso apportare fino a 20 ms di ritardo.

Delay component	**Time [ms]**
Grant Acquisition Process	5
Random Access Procedure	9.5
Transmission Time Interval	1
Signal Processing	3
Packet Retransmission	8
Core Network/ Internet	variabile

Tabella 2: varie fonti di ritardo in uplink e downlink all'interno di una rete LTE

È facile vedere che le due fonti di latenza più critiche del network di accesso radio sono quelle riguardanti l'instaurazione del collegamento (procedimento Grant Acquisition o la procedura di accesso random) e le ritrasmissioni dei pacchetti, dovute alla congestione del canale o ad errori di trasmissione.

3.2 Fog Computing

Per comprendere le sfide che si presentano nell'abilitazione di servizi a latenza ultra-bassa, è utile iniziare considerando quella che è la tipica architettura di rete attraverso la quale viaggia l'informazione. L'attuale EPC è caratterizzata da un piccolo numero di elementi di rete ad alta capacità ed affidabilità, come ad esempio i SGW, PGW e MME, i quali fanno affidamento a costosi hardware specifici per le applicazioni da svolgere. A causa del costo e della difficoltà della gestione di tali nodi, gli elementi di rete EPC, insieme ai core data center, sono tipicamente distribuiti in modo altamente centralizzato e geograficamente dispersivo. Il collocamento limitato di questi NE (Network Element) e degli Internet Exchange Points (IXP), pongono un importante problema nella capacità di offrire servizi a latenza ultra-bassa ai limiti della rete. I pacchetti, infatti, prima di essere instradati verso l'utente finale, devono entrare attraverso un router IXP per poi essere inoltrati attraverso nodi PGW e SGW. Tale sistema rispecchia il paradigma di calcolo utilizzato fino a ora, ovvero il Cloud Computing (CC). Con CC si intende un modello computazionale centralizzato nel quale la maggioranza dei calcoli vengono eseguiti

nel cloud per via della sua elevata potenza di calcolo e capacità di storage. Ciò significa che tutti i dati e le richieste di servizio devono essere trasmesse al Cloud, ovvero al centro della rete. Combinando tale sistema con l'enorme quantità di dati generati previsti con introduzione del 5G e degli scenari IoT che ne conseguono, e combinato ad un incremento non altrettanto apprezzabile delle capacità di trasmissione, ne deriva un ritardo dato dal trasferimento e dall'elaborazione dei dati inaccettabile per il supporto di numerosissime applicazioni latency-sensitive di nuova generazione. Per questi casi di utilizzo, infatti, molto spesso alcune decisioni possono essere prese localmente, senza la necessità di appoggiarsi al Cloud. Inoltre, anche se alcune decisioni devono essere elaborate e prese nel Cloud, è spesso inefficiente e non necessario inoltrare tutti i dati per la loro processazione e memorizzazione poiché non tutti i dati, in realtà, sono utili all'analisi e al processo decisionale. Il Fog/Edge Computing si inserisce proprio in questo contesto come tassello mancante utile al Cloud per coprire anche queste aree di utilizzo. Con Fog Computing (FC) [23] si intende una architettura di rete orientata al calcolo geograficamente distribuita dove dispositivi differenti e molto eterogenei tra loro sono costantemente connessi all'edge della rete per offrire elasticamente servizi di elaborazione, memorizzazione e comunicazione dati. La caratteristica principale del FC è quella di estendere i tipici servizi del Cloud verso i bordi del network. In questo modo, accorpando e aggiungendo risorse locali, tali servizi diventano molto più vicini all'utente finale e quindi più facilmente e soprattutto velocemente fruibili. Il tempo di trasferimento dati infatti, insieme al numero di trasmissioni totale nella rete viene ampiamente ridotto. È per tale motivo che il paradigma del FC si presenta come una promettente strada percorribile nella ricerca del soddisfacimento della domanda di applicazioni real-time e latency-sensitive.

Il modello concettuale che descrive l'architettura Fog inserisce tutti gli apparati che estendono le funzionalità del cloud in un livello intermedio dedicato in connessione ai dispositivi finali di produzione dei dati. Gli agenti attivi propri di questo modello di piattaforma architetturale sono detti *Fog nodes*: questi hanno la caratteristica di essere lontani dal data center Cloud, ma molto prossimi ai dispositivi utente; sono inoltre in numero elevato e distribuiti convenientemente sul territorio in modo da fornire copertura al maggior numero di utenti possibile. Da una parte quindi i nodi Fog sono connessi, principalmente in modalità wireless, con i dispositivi perimetrali della rete a cui danno supporto, dall'altra possono essere connessi col Cloud in modo da poterne usufruire appieno le risorse. Tali nodi, anche essendo meno performanti delle apparecchiature delle piattaforme Cloud, hanno un comparto tecnologico adatto all'esecuzione di compiti complessi e alla memorizzazione dei dati; inoltre sono scalabili e adattabili a diversi tipi di deployment. È importante enfatizzare il fatto che il FC si pone come un'espansione del

CC e non come un sostituto. I dati prodotti dall'utenza dei servizi Cloud, cioè, passano dal Fog che seleziona le computazioni che richiedono tempi stringenti ed esegue il lavoro che può essere fatto localmente mentre il Cloud viene sollecitato per richieste che il livello Fog non è in grado di eseguire o che ha eseguito parzialmente, e in generale viene interpellato per richieste che prevedono vincoli temporali poco stringenti, come, ad esempio, l'elaborazione statistica dello storico del sistema o l'analisi e lo storage a lungo termine di Big Data. Lo stato, e i dati in generale, custodito dai nodi Fog è di dimensioni ridotte rispetto a quello che deve conservare il Cloud e mediamente è preservato per un tempo inferiore, quello utile all'esecuzione del servizio IoT real-time.

3.2.1 Obiettivi

Una volta data quella che è una visione generica del paradigma del FC, prima di specificare le caratteristiche e l'architettura di rete di tale modello, è bene definire quali sono gli obbiettivi fondamentali [24]:

- Latenza: è vitale che una piattaforma FC sia in grado di garantire all'utente finale i requisiti di bassa latenza richiesti dalle varie applicazioni e servizi.
- Efficienza: l'utilizzo efficiente di risorse ed energia è necessario sia per motivi diretti che indiretti. Come motivazioni immediate si ha il fatto che i nodi Fog sono limitati in risorse computazionali e di memoria, e il fatto che molti di questi, insieme agli UE, sono spesso alimentati a batteria. Con motivazioni indirette invece si intende il fatto che la ricerca dell'efficienza in tali ambiti va ad impattare anche sulla capacità di offrire basse latenze.
- Generalità: causa l'alta eterogeneità dei nodi Fog e dell'utenza della rete, è necessaria la capacità di lavorare con piattaforme e protocolli differenti.

3.2.2 Caratteristiche e vantaggi

Il FC nasce per affrontare compiti di calcolo, comunicazione e storage per i dispositivi vicini all'edge del network. Sebbene questa sia la caratteristica fondamentale e più immediata, ci sono molti altri vantaggi che possono essere riassunti come segue [25]:

1) Bassa latenza e interazioni real-time

   I nodi Fog ai margini della rete acquisiscono localmente i dati generati da sensori e dispositivi, li processano e li memorizzano a seconda della necessità. Ciò riduce significativamente la quantità di trasmissioni e movimenti dei dati attraverso Internet e permette l'uso di servizi localizzati ad alta velocità e qualità. Come già detto, è possibile così abilitare basse latenze che soddisfino i requisiti di applicazioni real-time e time-sensitive.

2) Risparmio in larghezza di banda

   FC estende le capacità computazionali e di storage ad un livello intermedio tra l'utente finale e il Cloud tradizionale. Alcuni compiti, tra cui rimozione della ridondanza, pulizia e filtraggio, estrazione dell'informazione, il pre-processamento dei dati, sonno eseguiti localmente. Così, solo la parte di dati utili viene trasmessa al Cloud. In alcuni scenari applicativi inoltre, la presa di decisioni può essere presa direttamente nei nodi Fog piuttosto che eseguita dal cloud.

3) Supporto alla mobilità

   L'architettura Fog è in grado di offrire servizi location-based e capace di controllare, comunicando sempre con l'utente finale, di controllare da e verso dove i dispositivi mobili si stanno muovendo, quindi come farli accedere all'informazione. In questo modo vengono migliorate le performance del sistema e la qualità del servizio.

4) Distribuzione geografica e analisi di dati decentralizzata

   Rispetto al modello centralizzato del CC, il FC consiste in un grande numero di nodi ampiamente distribuiti in una architettura decentralizzata. Invece di processare e memorizzare le informazioni in data center centralizzati molto distanti dall'utente finale, tale lavoro può essere fatto completamente o in parte in questi nodi. In tal modo è possibili supportare una più veloce analisi di big data, servizi location-based migliori, capacità di decision-making in real-time più potenti.

5) Eterogeneità e interoperabilità

   I nodi Fog sono molto diversi tra loro sia nell'ambiente in cui sono distribuiti sia nella loro forma fisica o virtuale. Possono essere infatti server ad alte prestazioni, gateways, edge routers, stazioni base, access points ecc. Tali piattaforme hardware possiedono quindi caratteristiche fisiche, software e di sistema operativo anche molto diverse. Sono usate inoltre anche differenti tipi di connessioni, da link ad alte velocita cablate coi data center a connessioni wireless di tipo WiFi, WLAN, 3G, 4G ecc. Data la natura eterogenea del FC, ciò significa che i vari dispositivi e nodi in generale derivano da differenti costruttori e provider. Ciò significa che FC deve essere in grado di

interoperare e cooperare tra questi nella condivisione di informazioni e risorse per garantire i servizi richiesti.

6) Sicurezza

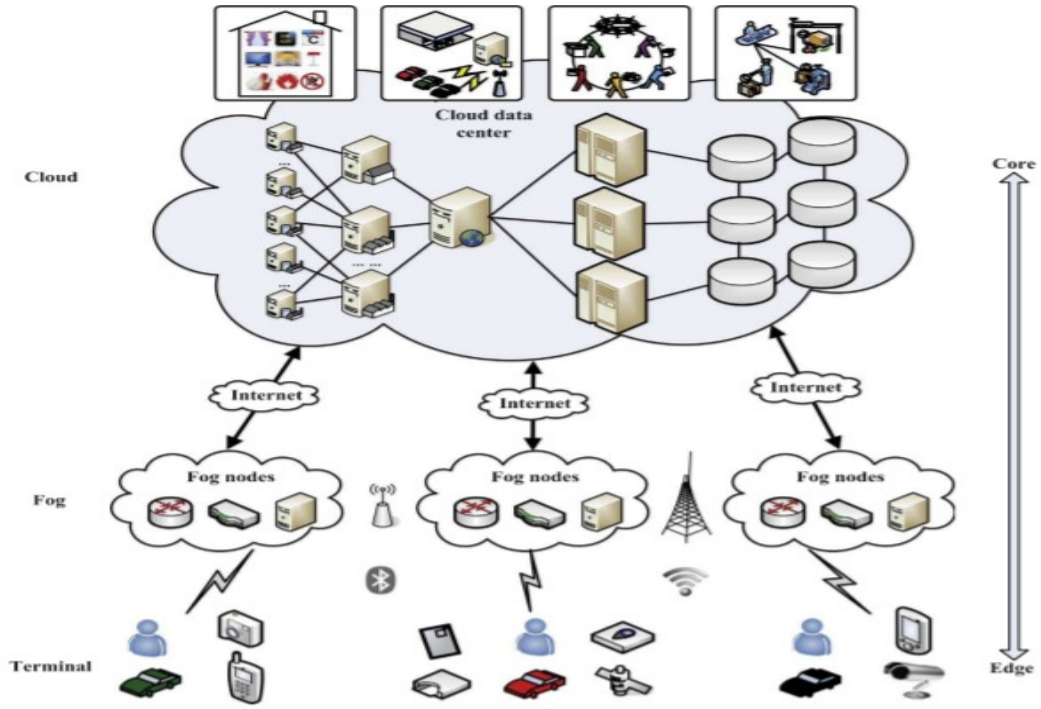
   La vicinanza dell'hosting dei servizi all'utente finale presenta particolari vantaggi in fatto di privacy e sicurezza riducendo la quantità di dati sensibili che transitano nella rete e isolando logicamente le informazioni tra le aree dove vengono allocati i dati ed eseguiti i vari processi.

7) Bassi consumi energetici

   L'alta dispersione dei nodi di rete FC permette che questi non generino grandi quantità di calore dovute alla loro concentrazione, non necessitando così di potenti sistemi di raffreddamento. Oltre a questo, la modalità di comunicazione short range e altre politiche di gestione dell'energia dei nodi mobili riducono il consumo di potenza nella comunicazione. [26] dimostra con alcuni risultati sperimentali che il consumo medio di energia adottando il paradigma FC si riduce del 40.48% rispetto all'architettura di rete tradizionale.

3.2.3 Architettura del Fog Computing

Il modello di riferimento di un ambiente computazionale facente uso del paradigma FC è composto di tre livelli [27]:



- *Livello Cloud:* il livello CC consiste di molteplici server, database e dispositivi ad alte prestazioni computazionali e di storage, i quali forniscono servizi di vario genere. Con nodi di questo tipo vengono offerte potenti risorse di calcolo e capacità di storage adatte per l'analisi, l'elaborazione e la memorizzazione di enormi quantità di dati. In termini di latenza end-to-end, il tempo da pagare è dell'ordine di grandezza dei secondi.
- *Livello Fog:* tale livello si colloca vicino al margine della rete. È composto da un grande numero di nodi Fog, i quali includono router, gateway, access point, base station e altri specifici server Fog. Questi nodi sono ampiamente distribuiti tra i dispositivi terminali e il Cloud e possono essere fissi oppure mobili, quindi montati su supporti in movimento. I dispositivi finali, infatti, possono accedere convenientemente a tali nodi per ottenere servizi. I nodi Fog hanno capacità computazionali, trasmissive e di memorizzazione limitate, ma sufficienti per compiere analisi real-time e supportare applicazioni time-sensitive all'interno del livello. Inoltre, i nodi Fog sono connessi con il Cloud attraverso la core network con cui possono interagire e collaborare per ottenere maggiori risorse.
- *Livello Terminali:* questo livello è il più vicino all'utente finale e all'ambiente fisico di utilizzo. Consiste in dispositivi IoT di vario genere, per esempio smartphones, veicoli smart, sensori, attuatori, carte intelligenti, lettori ecc. Tali dispositivi sono largamente distribuiti geograficamente e sono responsabili del rilevamento di oggetti fisici, eventi e in generale della creazione dei dati da inoltrare poi ai livelli superiori per essere elaborati e memorizzati. È possibile notare che smartphones e veicoli smart rappresentano delle eccezioni nel senso che posseggono delle capacità di calcolo autonome.

3.2.4 Piattaforma Fog Computing sperimentale: simulazione e risultati

Per verificare la validità del paradigma del Fog Computing come via percorribile nell'abilitazione di servizi con requisiti di bassa latenza, verrà di seguito proposta una piattaforma FC sperimentale sulla quale verranno poi eseguite alcune simulazioni che comprovino la capacità di tale architettura di ridurre efficacemente i ritardi rispetto al tradizionale approccio basato sul Cloud Computing. Riprendendo l'esperimento proposto in [24], la piattaforma in esame consiste di due sottosistemi, ciascuno composto di un router e tre server. I router sono connessi al cloud Amazon EC2 attraverso connessione WAN, inoltre sono

interconnessi tra di loro attraverso connessioni LAN. Su ciascun sottosistema è installato OpenStack e, in particolare, i quattro moduli Keystone, Glance, Nova, and Cinder. Keystone serve per le funzioni di autenticazione e autorizzazione, Glance per la gestione dei dispositivi virtuali, Nova come modulo computazionale per semplici funzionalità del network e Cinder come modulo di storage. I router sono inoltre integrati con funzioni Wireless AP in modo tale che i dispositivi mobili possano accedere al cloud Amazon EC2 attraverso di essi.

In seguito, vengono riportati alcuni risultati sperimentali effettuati sulla piattaforma sopra descritta. In Figura 2 viene presentata la comparazione tra i risultati in termini di latenza e larghezza di banda fornita da entrambe le piattaforme Fog e Cloud. Come metrica per la latenza viene utilizzato il Round Trip Time (RTT) mentre la misura della larghezza di banda è data sia per uplink che downlink. Come è evidente da tali risultati preliminari, il FC presenta forti vantaggi in entrambe le metriche prestazionali.

	RTT (ms)	Up/Down-link Bandwidth (Mbps)
Fog	1.416	83.723/101.918
Cloud	17.989	1.785/1.746

Figura 2: confronto tra latenza e larghezza di banda (in UL e DL) tra Fog e Cloud

Per comprendere meglio i benefici apportati dall'utilizzo del FC, viene poi sperimentato l'utilizzo di un servizio di riconoscimento facciale che possa essere eseguita da uno smartphone, attraverso sia il Fog che il Cloud. Il funzionamento di tale servizio prevede l'installazione di una app sullo smartphone dell'utente, quest'ultimo quindi, tramite app, potrà prendere una immagine contenente il volto di una persona e trasmetterlo ad un server remoto presente nel Fog o nel Cloud. Il server procederà al riconoscimento del volto attraverso una procedura di matching con le immagini di un database locale. Nell'implementazione utilizzata vengono prese immagini di 384x286 pixel e un database di 1521 foto. Facendo eseguire la stessa richiesta dal cloud Amazon EC2 e dalla piattaforma Fog si ottengono i risultati di Figura 3.

	Face recognition time on the server (ms)	Response Time (ms)
Fog	2.479	168.769
Cloud	2.492	899.970

Figura 3: performance di riconoscimento facciale eseguito su Fog e Cloud

Il tempo di risposta ("Response Time") misura il periodo di tempo che intercorre tra il momento in cui lo smartphone inizia ad effettuare l'upload dell'immagine e il momento in cui riceve il risultato dal server remoto. Da questi ultimi risultati si vede che il tempo necessario al lavoro computazionale è sostanzialmente identico; sapendo inoltre dalla Figura 2 che la differenza tra Fog e Cloud è inferiore a 20 ms, se le deduce che la larghezza di banda è il fattore che contribuisce per la maggiore nella differenza riscontrabile tra i due tempi di risposta. Anche in questo caso i tempi risultanti dal sistema Fog sono chiaramente migliori.

4.2.3 Tecnologie utilizzate

La realizzazione e l'effettiva utilità del FC dipendono dall'implementazione di alcune tecnologie che ne supportino la distribuzione e le applicazioni. In questo paragrafo verranno presentate alcune di queste tecnologie, selezionate per via della rilevanza e dell'impatto che queste hanno nel soddisfacimento dei requisiti di latenza per trasmissioni URLLC.

a) Resource management

Come si è visto nella precedente analisi delle fonti di latenza nelle attuali reti, la categoria riguardante le risorse radio (comprendente il Grant acquisition process e il Random Access procedure) costituisce la maggior origine del ritardo in una trasmissione. Per tale motivo, la ricerca di metodi di scheduling e di gestione delle risorse per l'abilitazione di applicazioni a bassa latenza in network Fog assume una elevata priorità. Tra queste, due possibili tecniche proposte in [28] sono:

a.1*) Non-Orthogonal Multiple Access (NOMA)*

Nelle esistenti reti wireless LTE, le risorse radio sono allocate ortogonalmente agli utenti, in tal modo si richiede alla stazione radio base, prima di procedere all'elaborazione ed inoltro della trasmissione, di attuare un procedimento di contention-based random access, il quale, soprattutto in presenta di molti utenti, risente di elevate latenze dovute a collisioni frequenti. NOMA si presenta, quindi, come tecnica alternativa al convenzonale OMA e si divide in due categorie: power-domain NOMA e code-domain NOMA. Prendendo come esempio quest'ultima modalità, a ciascun utente è assegnata una univoca codifica di canale (nel caso power-domain viene assegnato un certo livello di potenza del segnale), la quale permette alla stazione radio base di identificare utenti diversi, riducendo efficacemente l'interferenza multiutente, eliminando il procedimento di accesso random e

permettendo, quindi, la condivisione delle stesse risorse radio da molteplici UE. In tale modo è inoltre possibile abilitare trasmissioni Grant-free e migliorare, tramite adeguate codifiche, l'affidabilità della comunicazione.

In Figura 2 vengono mostrate congiuntamente le prestazioni di comunicazioni basate su OMA e NOMA in uno scenario in cui i dispositivi in gioco selezionano randomicamente una sottobanda per la loro trasmissione. L'ampiezza di banda totale viene presa pari a W = 100 MHz, la quale viene suddivisa in un numero di sottobande, denotato con Ns, di larghezza uniforme, ciascuna quindi di ampiezza pari a W/Ns. Come è possibile vedere, quando il numero di dispositivi è basso (<100), OMA è leggermente migliore di NOMA in termini di ritardo. Tale risultato non è sorprendente poiché la probabilità di avere collisioni in questo caso è bassa e quindi i vari dispositivi possono raggiungere efficienze spettrali maggiori dovute all'ortogonalità delle trasmissioni. Tuttavia, quando il numero di dispositivi aumenta, è chiaro che NOMA diventa largamente preferibile. OMA, infatti, in uno scenario in cui il traffico dati è molto elevato, è sovrastato dal numero di collisioni dovute all'accesso casuale, e conseguentemente ne risultano latenze inaccettabili. Il vantaggio maggiore di NOMA deriva dal fatto che non si rendono necessarie le procedure di Grant Acquisition e di Random Access, dunque, gli utenti sono in grado di trasmettere i loro dati ogni qualvolta ne hanno bisogno condividendo le stesse risorse radio.

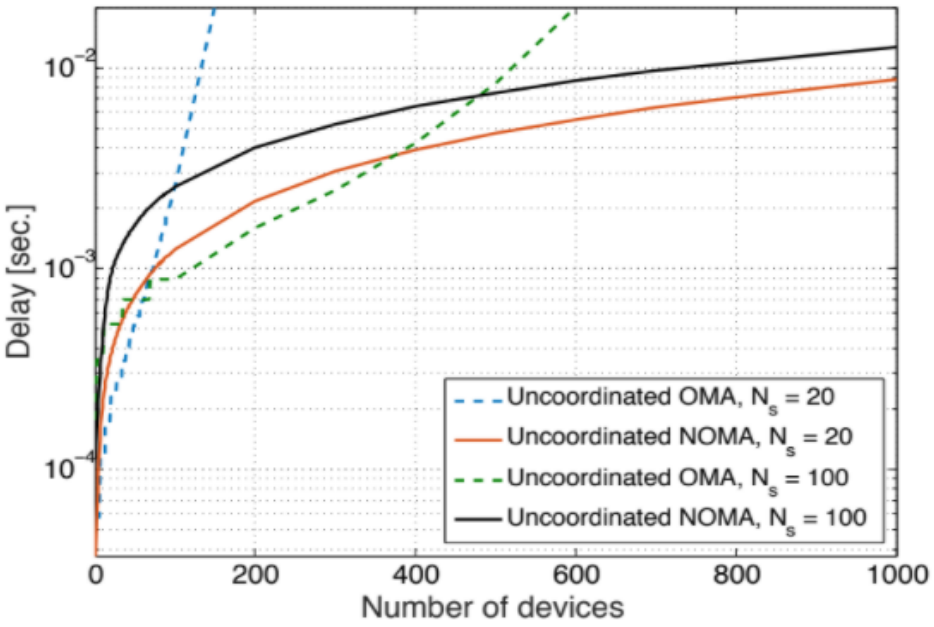


Figura 4: tempo di ritardo per NOMA e OMA in funzione del numero di dispositivi

a.2) *Resource Reservation via Resource Block Slicing*

Nella corrente rete LTE, la gestione dei Resource Blocks (RB) per non è differenziata in base al servizio ma viene eseguita congiuntamente. In questo modo le latenze di servizi distinti sono interdipendenti, per cui un sovraccarico di traffico generato da un particolare servizio apporta peggioramenti del ritardo di altri servizi. Un modo per aggirare questo problema è quello di riservare delle risorse radio per ogni servizio effettuando uno *slicing* dei Resource Blocks ed allocando differenti "fette" ai differenti servizi presenti nella richiesta di traffico del momento. In aggiunta, nel caso in cui non siano presenti richieste di trasmissione per un particolare servizio, la sezione di risorse destinata a tale servizio può essere riallocata per soddisfare altre richieste. Come dimostrato nella simulazione in [28], in tal modo è possibile ottenere una maggiore efficienza spettrale ed eliminare il problema della latenza causato dal sovraccarico di traffico derivante da un particolare servizio. In tale simulazione vengono prese due tipologie di servizi con differenti requisiti di latenza e data rate, ovvero Intelligent Transportation System con taglia media di pacchetto di 100 bytes e intervallo medio di pacchetto di 100 ms e Smart Grids con taglia media di pacchetto di 300 bytes e intervallo medio di pacchetto di 80 ms. I dispositivi trasmittenti per tali servizi vengono assunti distribuiti su un'area di 1 km quadro secondo un Processo Puntuale di Poisson (PPP) con media 400 e 600 rispettivamente per ITS e SG, i quali vengono serviti da quattro Base Station LTE operanti a 20MHz di larghezza di banda. In Figura 3 viene rappresentata congiuntamente la funzione di ripartizione (Cumulative Density Function (CDF)) per la latenza end-to-end di un pacchetto nel caso di un normale network LTE in cui tutti i RB disponibili sono condivisi equamente tra i vari servizi, e nel caso di funzionamento secondo RB slicing il quale isola il traffico dati tra ITS e SG. Come è possibile vedere, quando la domanda di traffico generato da tutti i dispositivi per ciascun servizio corrisponde alla proporzione di RB riservati, ovvero il 15% e 85%, le performance in latenza delle trasmissioni dei due servizi sono praticamente le stesse. Si noti il fatto che ITS genera traffico per 3.2 Mbps (400 dispositivi trasmettono 100 bytes ogni 100 ms) mentre SG genera 18 Mbps (600 dispositivi trasmettono 300 bytes ogni 80 ms). In particolare, rispetto il tradizionale funzionamento LTE, riservando RB per ciascun servizio viene diminuita la latenza da una media di 10 ms a 5 ms e 6 ms per ITS e SG rispettivamente. Si osserva inoltre che riservando porzioni di RB differenti, ad esempio 80% e 20%, le performance delle trasmissioni per ITS migliorano poiché un maggior numero di RB sono allocati a tale servizio mentre, al contrario le performance delle trasmissioni per SG peggiorano difatti le risorse ad esse allocate calano dall'85% al 20%. In ogni caso i valori

di latenza raggiunti sono comunque migliori di quelli di una rete LTE senza RB slicing comprovando così la validità di tale approccio.

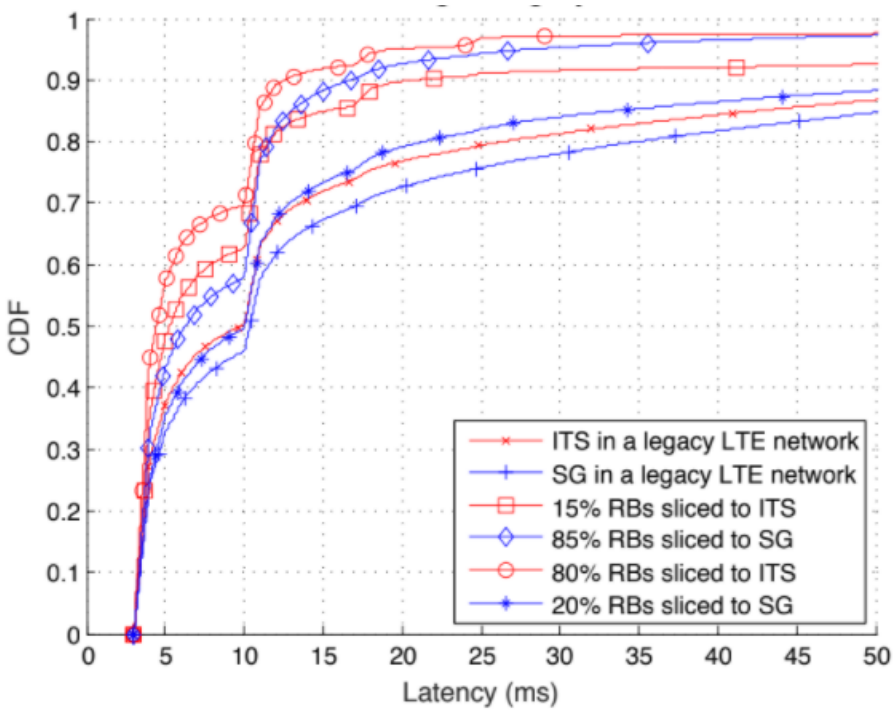


Figura 5: funzione di distribuzione della latenza nel caso di trasmissioni con e senza RB slicing

b) Offloading computazionale

Meccanismi di offload possono provvedere alla limitatezza delle risorse dei dispositivi al bordo della rete, aiutando a migliorare le prestazioni e a risparmiare batteria. Un possibile modello per risolvere situazioni di offload di calcolo in tale contesto sfrutta la teoria dei giochi [29]. L'approccio consiste nel formulare il problema della decisione della distribuzione dei compiti di calcolo come un gioco di offload computazionale multiutente. Quando diversi dispositivi effettuano simultaneamente l'offload sullo stesso canale wireless, viene eseguito solo quello che permette di ridurre effettivamente il tempo di esecuzione e l'energia consumata. In un ambiente con molteplici canali wireless, la decisione degli offload da eseguire dipenderà dalla performance totale. Nel caso di offload tra nodi Fog si può proporre un framework [30] che possa compiere l'offload di compiti anche parziali verso nodi vicini per ridurre, anche in questo caso, il consumo di tempo ed energia. La decisione dipenderà principalmente dal livello della potenza di calcolo e dell'energia dei nodi vicini e dalla probabilità che avvenga una connessione tra loro in futuro.

c) Latency management

L'obiettivo primario del latency management nel FC è quello garantire che il tempo di risposta di un servizio rimanga entro un limite accettabile. Tale limite è rappresentato dalla latenza massima tollerabile che comunque soddisfi la richiesta del servizio e i requisiti di Quality of Service (QoS). Una possibile soluzione della gestione della latenza [31] è quella di instaurare un meccanismo di cooperazione per il quale la risoluzione di incarichi computazionali possa essere eseguita collaborativamente da più nodi in modo da rispettare i requisiti di ritardo. Nel caso di studio della minimizzazione del tempo per il completamento di una richiesta di servizio [32], la latenza complessiva di trasmissione e calcolo di tutte le richieste può essere minimizzata tramite la distribuzione efficiente delle richieste di calcolo e il bilanciamento del carico di lavoro tra entrambi client e nodi Fog. In un altro lavoro [33] viene definito un modello matematico per il tracciamento dei tempi di ritardo di computazione e di comunicazione, e viene proposta un'architettura fog orientata alla gestione della latenza nella quale tale modello può essere utilizzato per guidare la selezione dei nodi del network Fog per ottenere in minimo tempo di ritardo nello svolgimento delle richieste di servizio.

d) Network function virtualization (NFV)

L'NFV si può forse dire essere la caratteristica chiave fondamentale per la realizzazione di una idea rete FC. L' idea basilare del NFV è che le funzionalità del network vengano separate dal livello fisico dell'hardware "virtualizzandole" ed astraendole dai dispositivi dedicati. Ciò significa che le risorse della rete (e.g. potenza computazionale e storage) possono essere ottimamente gestite e quindi condivise in modo pieno e flessibile, permettendo inoltre il rapido sviluppo e la rapida distribuzione di nuovi servizi.

e) Software defined networking (SDN)

SDN è un paradigma computazionale e di networking che nasce come metodo implementativo della virtualizzazione del network. Esso descrive una architettura che consente una gestione della rete completamente basata su software. A questo scopo il piano di controllo (Control Plane) e il piano d'utente (User Plane o Data Plane) vengono separati. In altre parole, il concetto di SDN è sinonimo di separazione tra l'infrastruttura e la sua configurazione, ovvero la logica di controllo (CP) implementata di default nei componenti hardware è astratta ed indipendente dall'hardware. In questo modo l'attività del CP, responsabile del corretto traffico di dati nell'architettura, poiché staccato dall'hardware e implementato in un software centralizzato, risulta sostanzialmente di più facile programmazione nel Software Defined Networking e quindi molto più flessibile e scalabile nell'amministrazione della rete. Con SDN, attraverso una

interfaccia standardizzata, è perciò possibile eliminare le differenze tra dispositivi eterogenei sottostanti alla rete (e.g., routers, switches, firewalls).

Nel FC, SDN può quindi aiutare a gestire più efficientemente network Fog eterogenee, ottimizzando la distribuzione e l'utilizzo delle risorse, quindi aiutando a ridurre la latenza.

f) Mobile Caching per Content Delivery

Schemi di caching intelligente possono essere una efficacie soluzione nel migliorare le performance di bassa latenza per applicazioni ad alta quantità di dati (e.g. multimedia, realtà aumentata). I nodi Fog, infatti, possono cercare di predire le domande dell'utente e selezionare in modo proattivo i contenuti più richiesti da memorizzare in cache permettendo così il loro riutilizzo e apportando una drastica diminuzione del ritardo e ad un miglioramento dell'efficienza della rete di backhaul. Una cache di questo può essere adoperata da ogni base station, facendo così pieno utilizzo delle risorse di storage, così che ogni qualvolta un dispositivo mobile richiede un contenuto pre-cached, la base station può intercettare la richiesta e ritornare direttamente il contenuto senza la necessità di contattare un data center remoto. La sfida principale in tale tecnologia consiste nel fatto che le capacità di storage dei nodi Fog sono limitate mentre il numero di possibili contenuti visualizzabili sono enormi. Per tale motivo alla base della predizione vanno utilizzati criteri di popolarità del file e pattern di correlazione tra file e utenti.

5. Conclusioni

In questa tesi si è mostrata la storia del settore delle telecomunicazioni attraverso le tappe fondamentali dello sviluppo delle tecnologie e dei sistemi di comunicazione che si sono resi necessari a fronte della costante crescita della domanda di servizi a banda larga e dell'utenza Internet in generale. Poiché, come si è visto, questo trend non è destinato a diminuire ma, al contrario a crescere, sono state presentate le caratteristiche delle nuove reti 5G, le quali si assumono l'onere di supportare il futuro ecosistema digitale e i relativi servizi. Sono stati dunque esposti alcuni scenari tecnologici ritenuti fondamentali e di grande rilievo nella ridefinizione dell'industria e della società. Tali casi d'uso sono caratterizzati dall'utilizzo della nuova tipologia di comunicazione URLLC introdotta dalle reti di quinta generazione. A seguito di ciò sono state indagate e quantificate le maggiori fonti di latenza presenti negli attuali network di quarta generazione distinguendo tra quelle derivanti dall'accesso wireless e quelle derivanti dalla Core Network. Una volta introdotti, quindi, gli obiettivi tecnologici, i requisiti necessari al loro soddisfacimento e le insufficienti prestazioni delle reti attuali, sono state introdotte delle possibili soluzioni tecnologiche che permettano l'abilitazione delle comunicazioni URLLC. Primo tra queste è stato descritto il paradigma del Fog Computing con relativi obiettivi, caratteristiche e vantaggi. In seguito alla definizione del modello architetturale di riferimento sono stati poi riportati alcuni risultati numerici che verificano la bontà di tale approccio e la sua capacità di ridurre ampiamente i tempi di ritardo caratteristici degli attuali network basati sul Cloud Computing. In aggiunta si sono riportate, insieme a delle tecnologie caratterizzanti del FC come l'offload computazionale e la virtualizzazione delle funzionalità di rete, alcune altre tecniche, in particolare nel campo della gestione delle risorse radio, che possono essere implementate a livello protocollare. In particolare, è stata presentata la tecnica di accesso al mezzo NOMA e i risultati di una simulazione che ne provano la validità attraverso il confronto con l'attuale utilizzo di OMA, in particolare in condizioni di elevata domanda di traffico dati; ed è stata presentata la tecnica di slicing dei Resource Blocks con relativa simulazione e risultati numerici che si propone efficacemente di ovviare al problema dell'interdipendenza della latenza derivante da servizi differenti.

Le tecnologie esposte e il loro utilizzo, nello sviluppo delle reti 5G, è già stato in parte previsto o implementato, e rappresentano una risorsa essenziale nella riduzione della latenza e nel raggiungimento dei requisiti caratterizzanti le comunicazioni URLLC.

BIBLIOGRAFIA

[1] M. A. Lema, A. Laya, T. Mahmoodi, M. Cuevas, J. Sachs, J. Markendahl, M. Dohler, “Business Case and Technology Analysis for 5G Low Latency Applications” IEEE Access, May 2017, pp. 5919-1922

[2] G. J. Sutton, J. Zeng, R. P. Liu, W. Ni, D. N. Nguyen, B. A. Jayawickrama, X. Huang, M. Abolhasan, Z. Zhang, E. Dutkiewicz, T. Lv “Enabling Technologies for Ultra-Reliable and Low Latency Communications: From PHY and MAC Layer Perspectives” IEEE Communications Surveys & Tutorials, Feb. 2019

[3] Third Generation Partnership Project (3GPP), “Study on Scenarios and Requirements for Next Generation Access Technologies (Release 15)”, document TR 38.913 V15.0.0, 2018

[4] International Telecommunication Union (ITU), “Minimum requirements related to technical performance for IMT-2020 radio interface(s)” Report M.2410-0, July 2017

[5] M. Galante, P. Marchese, G. Romano “Le prospettive del 5G e dell’Internet delle cose” AEIT, Nov. 2018, pp. 10-22

[6] ETSI TC-SMG “Digital cellular telecommunications system (Phase 2+); Radio transmission and reception” GSM 05.05 Version 5.0.0, Mar. 1996

[7] L. Badia, M. Levorato, F. Librino, M. Zorzi, “Cooperation techniques for wireless systems from a networking perspective,” IEEE Wireless Communications Magazine, vol. 17, no. 2, pp. 89-96, April 2010

[8] 3GPP “General Packet Radio Service (GPRS); Overall description of the GPRS radio interface; Stage 2” Release 4, Oct. 1999

[9] C. Bettstetter, H.-J. Vögel, and J. Eberspächer, “GSM phase 2+ general packet radio service GPRS: Architecture, protocols, and air interface,” IEEE Commun. Surveys, Vol. 2 (3), 1999

[10] Ericsson white paper, “The evolution of EDGE” 285 23-3107 Uen Rev. © Ericsson AB 2007

[11] 3GPP “UTRAN overall specifications (Release 11)” Technical Specification Group Radio Access Network, Dec. 2012

[12] 3GPP (Online page) J. Wannstrom “HSPA”

[13] LTE-Advanced (3GPP Release 10 and beyond), RF aspects,Takaharu Nakamura, 3GPP TSG-RAN-WG4 Chairman, December 2009.

[14] F. Alecu “The WiMAX Technology” Oeconomics of Knowledge, Volume 2, Issue 2, 2Q 2010

[15] M. Maresca, O. Trevisan, “Diffusione della banda larga attraverso la tecnologia di accesso WiMAX: il caso del Veneto” 2009 pp. 46-49

[16] Cisco “Cisco Vision: 5G Thriving Indoors”. 2017

[17] L. Badia, A. Erta, L. Lenzini, M. Zorzi, “A general interference-aware framework for joint routing and link scheduling in wireless mesh networks,” IEEE Network, 22(1), 32-38, 2008

[18] R. Garelli, B. Melis, J. Ramos “New Radio Interfaces Beyond 4G” Apr. 2018

[19] L. Badia, M. Miozzo, M. Rossi, M. Zorzi, “Routing schemes in heterogeneous wireless networks based on access advertisement and backward utilities for QoS support,” IEEE Communications Magazine, vol. 45, no. 2, pp. 67-73, February 2007

[20] 5GPPP Association, “5G automotive vision” 5GPPP, White Paper, Oct. 2015

[21] M. Valenti, G. Graditi, “Le Smart Grid per un futuro energetico sostenibile e sicuro” Focus ENEA, Feb 2020, pp.105-107

[22] 3rd Generation Partnership Project (3GPP), “Study on Latency Reduction Techniques for LTE (Release 14)”, document TR 36.881, 2016

[23] L. M. Vaquero, "Finding your Way in the Fog: Towards a Comprehensive Definition of Fog Computing," Hewlett-Packard Labs, Bristol, U.K., 2014

[24] S. Yi, Z. Hao, Z. Qin, Q. Li “Fog Computing: Platform and Applications”, Third IEEE Workshop on Hot Topics in Web Systems and Technologies, Nov. 2015, pp. 75

[25] H. Atlam, R. Walters, G. Wills, “Fog Computing and the Internet of Things: A Review” MDPI Big Data Cogn. Comput. 2018, pp. 4

[26] S. Sarkar, S. Misra, “Theoretical modelling of fog computing: a green computing paradigm to support iot applications” IET Netw. 5 (2), 2016, pp. 23–29

[27] P. Hua, S. Dhelima, H. Ninga, T. Qiu, “Survey on fog computing: architecture, key technologies, applications and open issues” Journal of Network and Computer Applications 98, 2017, pp. 27-42

[28] H. Chen, R. Abbas, P. Cheng, M. Shirvanimoghaddam, W. Hardjawana, W. Bao, Y. Li, B. Vucetic, “Ultra-Reliable Low Latency Cellular Networks: Use Cases, Challenges and Approaches” IEEE Communications Magazine, Dec. 2018, pp. 119-124

[29] X.Chen, L.Jiao, W. Li, X. Fu, “Efficient multi-user computation offloading for mobile- edge cloud computing” IEEE/ACM Trans. Netw. 24 (4), 2015, pp. 974–983

[30] L. Badia, “A Markov analysis of selective repeat ARQ with variable round trip time,” IEEE Communication Lett., 17(11), 2013, pp. 2184–2187

[31] L. Prospero, R. Costa, and L. Badia, “Resource sharing in the Internet of Things and selfish behaviors of the agents,” IEEE Trans. on Circuits and Systems II-Express Briefs, vol. 68, no. 12, 2021, pp. 3488–3492

[32] J. Oueis, E.C. Strinati, S. Sardellitti, S. Barbarossa, “Small cell clustering for efficient distributed fog computing: A multi-user case” Proceedings of the IEEE 82nd Vehicular Technology Conference (VTC Fall), 2015, pp. 1–5

[33] L. Badia, M. Levorato, M. Zorzi, “A channel representation method for the study of hybrid retransmission-based error control,” IEEE Transaction Communication, 57(7), 2009, pp. 1959-1971

[34] W.Gao, “Opportunistic peer-to-peer mobile cloud computing at the tactical edge” Proceedings of the IEEE Military Communications Conference, 2014, pp. 1614–1620

[35] L. Badia, A. Baiocchi, S. Merlin, S. Pupolin, A. Todini, A. Zanella, M. Zorzi, “On the impact of physical layer awareness on scheduling and resource allocation in broadband multicellular IEEE 802.16 systems,” IEEE Wireless Commun., vol. 14, no. 1, pp. 36-43, February 2007

[36] D. Zeng, L. Gu, S. Guo, Z. Cheng, S. Yu, “Joint optimization of task scheduling and image placement in fog computing supported software-defined embedded system” IEEE Trans. Comput. 65 (12), 2016, pp. 3702–3712

[37] K. Intharawijitr, K. Iida, H. Koga, “Analysis of fog model considering computing and communication latency in 5G cellular networks” Proceedings of the IEEE International Conference on Pervasive Computing and Communication Workshops 2016, pp. 1–4