energies

Review

The Four-C Framework for High Capacity Ultra-Low Latency in 5G Networks: A Review

Anabi Hilary Kelechi [1,*], Mohammed H. Alsharif [2,*], Athirah Mohd Ramly [3], Nor Fadzilah Abdullah [3] and Rosdiadee Nordin [3]

[1] Department of Electrical Engineering and Information Engineering, College of Engineering, Covenant University, Canaanland, Ota P.M.B 1023, Ogun State, Nigeria
[2] Department of Electrical and Electronics Engineering, Faculty of Engineering and Architecture, Istanbul Gelisim University, İstanbul, Avcılar 34310, Turkey
[3] Centre of Advanced Electronic and Communication Engineering (PAKET), Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bangi 43600, Selangor, Malaysia
* Correspondence: hilary.anabi@covenantuniversity.edu.ng (A.H.K.); moh859@gmail.com (M.H.A.)

Received: 17 July 2019; Accepted: 28 August 2019; Published: 6 September 2019

Abstract: Network latency will be a critical performance metric for the Fifth Generation (5G) networks expected to be fully rolled out in 2020 through the IMT-2020 project. The multi-user multiple-input multiple-output (MU-MIMO) technology is a key enabler for the 5G massive connectivity criterion, especially from the massive densification perspective. Naturally, it appears that 5G MU-MIMO will face a daunting task to achieve an end-to-end 1 ms ultra-low latency budget if traditional network set-ups criteria are strictly adhered to. Moreover, 5G latency will have added dimensions of scalability and flexibility compared to prior existing deployed technologies. The scalability dimension caters for meeting rapid demand as new applications evolve. While flexibility complements the scalability dimension by investigating novel non-stacked protocol architecture. The goal of this review paper is to deploy ultra-low latency reduction framework for 5G communications considering flexibility and scalability. The Four (4) C framework consisting of cost, complexity, cross-layer and computing is hereby analyzed and discussed. The Four (4) C framework discusses several emerging new technologies of software defined network (SDN), network function virtualization (NFV) and fog networking. This review paper will contribute significantly towards the future implementation of flexible and high capacity ultra-low latency 5G communications.

Keywords: 5G; ultra-reliable communication; ultra-low latency; SDN; NFV; MU-MIMO; MM2M; caching; computing; virtualization

1. Introduction

Fifth Generation (5G) wireless communications will be driven by three use cases of enhanced mobile broadband (eMBB), massive machine-type communication (mMTC) and lastly, ultra-reliable low latency communication (URLLC). The eMBB is designed for high bandwidth internet access suitable for web browsing, video streaming, and virtual reality. The mMTC is responsible for establishing narrowband Internet applications such as narrowband IoT (NB-IoT). The URLLC facilitates certain delay-sensitive applications such as factory automation, remote surgery and autonomous driving [1]. Of all the above technologies, URLCC will be the most stringent to achieve based on the 1 ms end-to-end (E-2-E) latency, link reliability of 99.99999% and error rates that are lower than 1 packet loss in 10^5 packets as recommended by the ITU-R M.2410.0 [2]. New techniques are required to meet with the stringent latency and reliability requirements for URLLC as we migrate into the domain of haptic communications, tactile Internet, intelligent transport system (ITS) and industry 4.0 era revolution [3,4].

Studies have investigated URLCC generating a mixture of results [5,6]. In other to improve vehicle safety in vehicle-to-vehicle in the 5G user case, an application layer raptor code Q codes have been proposed with a target of end-to-end delay latency of 5ms [7]. Network latency is a very important metric in today's web mobile applications as there exists a relationship between network latency, online shopping and web surfing [8]. As an example, the e-commerce online trading platform Amazon has observed that there is a 1% decrease in sales on a 100 ms network latency. Likewise, Google has observed that in every 0.5 seconds increase in search latency, there is a corresponding 20% drop in network traffic. An integral part of 5G communications is the transmission of real-time touch perception communications (haptic Internet) lending support to low latency and ultra-low latency cellular communications. Among the various visions of 5G [9], high data rate, high capacity, and ultra-low latency are of uttermost importance with peculiar challenges. Studies have indicated there is a trade-off between the aforementioned metrics as the enhancement of one factor, deteriorates the others [2,9]. It is obvious that achieving ultra-low latency in the absence of a trade-off in link reliability, network coverage and capacity in 5G criteria is not feasible based on the current physical air interface limitations [10,11]. The scenario will be even more complex considering the evolving massive machine-to-machine communication (MM2MC) URLLC regime in which thousands of nodes are expected to transmit their payload in a real-time fashion. In this situation, a radio interface capable of sustaining low bandwidth orthogonal communication becomes crucial. The MU-MIMO technology is a possible candidate to drive the massive connectivity criterion in 5G as illustrated in Figure 1.

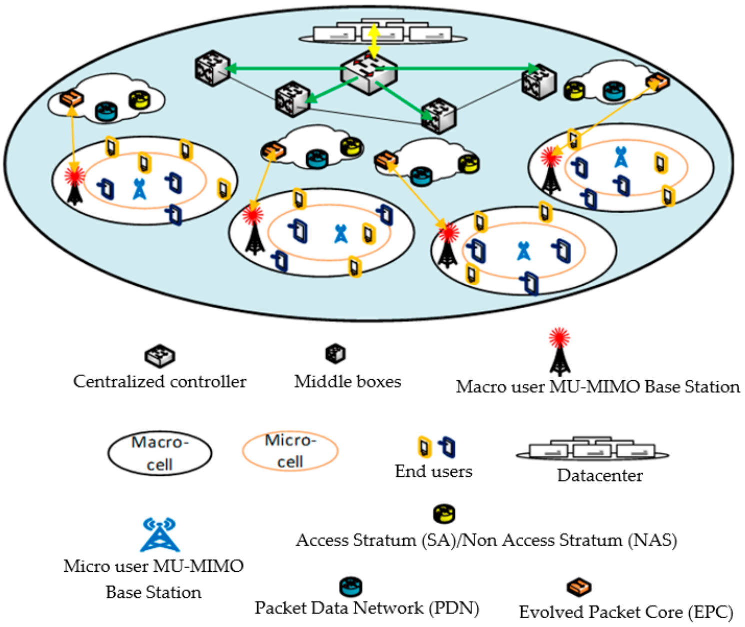
Centralized controller
Middle boxes
Macro user MU-MIMO Base Station
Macro-cell
Micro-cell
End users
Datacenter
Micro user MU-MIMO Base Station
Access Stratum (SA)/Non Access Stratum (NAS)
Packet Data Network (PDN)
Evolved Packet Core (EPC)

Figure 1. A potential high capacity 5G network architecture.

In some literature, MU-MIMO is often referred to as the massive-MIMO configuration [12,13]. For consistency sake, this work adopts MU-MIMO terminology throughout this study. MU-MIMO is already a consolidated technology in which a base station (BS) equipped with 100 antenna elements transmits concurrently to 10 mobile stations (MS) within the 1 ms budget [14,15] as highlighted in Table 1. The performance metrics of MU-MIMO lie towards the successful transmission of data streams while exploiting some spatial degrees of freedom (DoF). Inadvertently, there is no consideration for the latency issue in MU-MIMO as the technology was specifically designed for 4G communication. It has been reported that current WLAN and cellular systems which drive MU-MIMO are not capable of adhering to the 1 ms latency budget [14] and hence, new solutions are needed. Several studies have captured the role of MU-MIMO in 5G communications paradigm [16,17]. In light of the above, no analysis and discussions on the 5G low latency budget have been undertaken. Consequently, there is a need to design technologies that will support 5G E-2-E low latency criteria. Thereby, sustaining and enhancing the quality of service (QoS) and quality of experience (QoE) of 5G users. A high capacity 5G network topology comprised of: data centre, centralized network controller, middle boxes, packet data network (PDN), evolved-packet core (EPC), access stratum (AS), non-access stratum (NAS), MU-MIMO BS, micro and macro cell users respectively as depicted in Figure 1.

Table 1. MU-MIMO System Configurations.

Number of MIMO Processors	Number of Bit Processors	Max Number of Antennas
1	1	1–32
2	1	33–64
4	1	65–128

The 5G network architecture as illustrated in Figure 1 is an example of DenseNets, which has several limitations in the context of radio access network bottlenecks, control overhead issues and high operational costs [18]. LTE exhibits approximate 100 ms and 30 ms latency from the control plane and the user plane, respectively. The control plane latency is the signaling required to switch the user equipment (UE) from the idle mode to the active mode involving the radio resource control (RRC) connection and set up a dedicated mode. User-Plane latency is defined as one-way transmit time between a packet availability at the IP layer in the UE/Evolved-UMTS Terrestrial Radio Access Network (E-UTRAN)/edge node and the availability of this packet at the IP layer in the E-UTRAN/UE node. User-Plane latency is relevant to the performance of many applications. In addition, applications will suffer from service request delay. This delay is encountered by the nodes when trying to initialize cell search. In order to access the physical random access channel (PRACH) in the LTE, the user sends its information request and key into the primary control channel (PCCH) which is mapped to the paging channel (PCH). The PCCH and PCH are responsible for granting access to the physical downlink control channel (PDCCH) which oversees decoding of the downlink control information (DCI) signals. The 3GPP standard specifies the physical control format indicator channel (PCFICH) mandates the UE to communicate to the eNB to obtain the control format information (CFI) which contains information necessary to decode PDCCH information. To address this, software define networking (SDN) centralized solution as illustrated in Figure 2 has been proposed because it has a global view of the network. The SDN architecture integrates into: (i) data plane for traffic forwarding, (ii) control plane for instantiating network rules enforcement and; (iii) management plane for network-wide policies formulation such as: load balancing, QoS, and security [19,20].

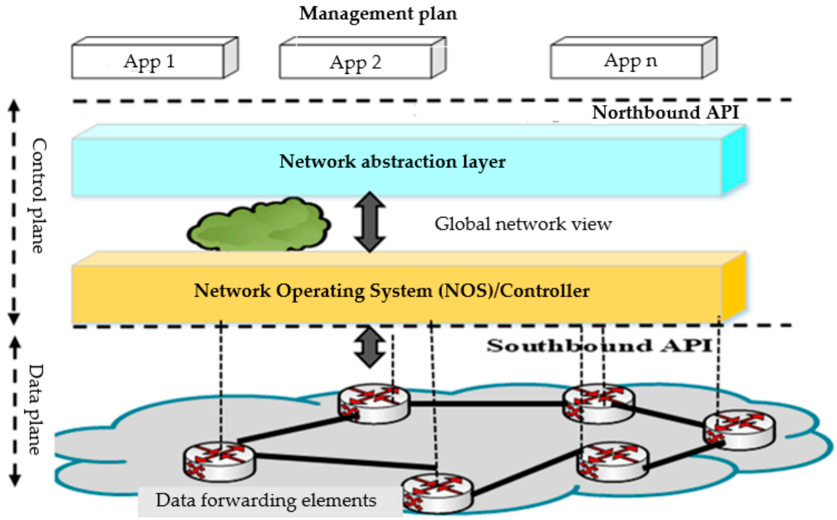

Figure 2. SDN 5G network architecture.

The SDN 5G architecture consists of network operating system (NOS)/controller. The NOS is a low-level language that expedites communication among the multicores threads. The multicores enable the implementation of distributed parallel computing topology in which many cores are distributed leading to lower task execution time. The notable platform is the OpenDaylight [21,22], drives the southbound OpenFlow application programming interface (API) towards the data plane [23]. The northbound API is currently implementing RESTful API [24] as the de-facto API to the management plane.

To address the low latency issue in MU-MIMO for future 5G, this paper highlighted Four (4) C framework consisting of: Computing, Cost, Complexity and Cross-Layer. Details for each of the C is summarized as below:

(1) *Computing*: Instead of sending sensor data directly to the cloud, the edge devices undertake the function of data processing, analysis and storage. Thus, minimizing overall network traffic and latency. Overall, fog computing is designed to enhance network efficiency, performance and minimize the quantity of data pushed to the cloud for handling, examination and ware-housing to achieve an URC with low-latency.

(2) *Cost*: Minimization of costs and maximization of resource utilization are crucial elements towards achieving ultra-reliable and low-latency. Virtualization is not a new technique, but only recently, with the advent of cloud computing (CC) and big data concept, has it become a staple in CC design. Virtualization techniques in CC facilitate the execution of multiple applications and operating systems on the same server, without interfering with any of the other services provided by the server or host platform, thereby enabling efficient resource utilization, cost reduction, and decrease in latency through increasing the speed of relocation process in the virtual server files.

(3) *Complexity*: Massive-MIMO hardware structures deploy a huge numbers of antennas and hence, this leads to vast complexity in order to detect the signal. Complexity is considered as a challenge to be solved in practical massive MIMO systems. The complexity is increased exponentially with the numbers of transmit antennas and thus, making the large-scale MIMO less practical. Efficient receiving schemes and precoding are much needed in massive MIMO systems to mitigate the computational complexity. To lessen the complexity as well as to strengthen the convergence rate, numerous methods and schemes have been suggested in detail in Section 5. It is suggested that henceforth, algorithms designed for 5G URLLC should be judged on their complexity O(N) indicator.

(4) *Cross-Layer*: The cross-layer design diverges from the traditional network design whereby each layer of the stack would be made to work individually. Cross-layer optimization is important to control the packet loss as well as the waiting period caused by transmission and queuing process. The correlation between cross-layer in massive MIMO with the low latency is discussed in depth in Section 6.

The contributions of our paper may be summarized as follows:

(1) To present an insight into the fundamental limitations of the MU-MIMO system from the context of low-latency requirements and potential solutions from the literature. For instance, a trade-off is necessary between fast convergence and high computational algorithms as highly computational and iterative algorithms might lead to signal processing delays. A case in hand is turbo encoders for line coding.

(2) To synthesize existing works on ultra-low latency 5G communications presenting their shortcomings and proposing a possible solution of enhancement. Some existing works have limited their work to millimeter-wave communication as the main drivers of URLLC. Thus, presenting a distorted view of the problem.

(3) To implement a framework for achieving a low latency 5G communications based on SDN, network function virtualization (NFV), edge computing, caching solutions. These tools can

enhance signal processing speeds by mapping some of these functions to the cloud in the case of high densification expected of 5G networks.

(4) We presented the challenges and progress towards the Four (4) C framework, which can serve as a future research direction in this regard.

This review paper is different from the other review paper related to URC and low-latency on 5G as published in [16,25,26] in terms of the following:

(1) All the above literature failed to address the ultra-low latency requirements of 5G from the MTC driven by the MU-MIMO technology.
(2) Work in [26] identified SDN/NFV as the core of the ultra-low latency 5G communications but failed to exploit the need to virtualize other network peripherals such as hard-disk, CPU, memory, etc.
(3) This study highlights the need for a soft integration between the fog networking, MU-MIMO and SDN/NFV as the driver towards attaining ultra-low latency MM2M 5G regime.

The rest of the paper is organized as follows: Section 2 presents a detailed and descriptive analysis 5G MU-MIMO latency problem and overview of the Four (4) C framework. Section 3 discusses how to attain low latency 5G communications considering computing, specifically fog computing as an intermediate layer between CC and end devices. Section 4 focuses on attaining low latency 5G communications from a cost perspective. Sections 5 and 6 address the low latency 5G communications issue from complexity and cross-layer. Section 7 discusses the challenges and opportunities of the four (4) Cs. Finally, Section 8 concludes the work.

2. Potential of MU-MIMO Based on Four C Framework

The MU-MIMO technology is analysed from the perspective of support for ultra-low latency 5G communications. Meanwhile, our Four (4) C model framework is presented in Section 2.2.

2.1. MU-MIMO Technology Candidature for Massive Network Densification

The MU-MIMO system comprises a transmit antenna concurrently beaming signals to a plethora of end-users as shown in Figure 3. 5G is equipped with the following capabilities: support for 1000-fold more data than the current aggregate data rate, 100-fold more than the current user data rate and a 100-fold increase in the number of concurrently connected devices [9]. The above is only realizable via a heterogeneous network architecture consisting of small cells, joint transmission coordinated multipoint (CoMP) topology and macro cells. MU-MIMO is an integral part of the wireless communications and has been widely adopted by: IEEE 802.11n, 802.11ac WLAN, 802.16e (Mobile WiMAX), 802.16m (WiMAX), 3GPP long-term evolution (LTE) and LTE-Advanced.

From Figure 3, mobile users can only transmit and receive mobile data after the data has traversed from the packet data network (PDN) to the evolved packet core (EPC) assuming a reliable wireless link. If call set up time and joint transmission protocol overhead are neglected, the two sources of network latency in wireless communications are [25]: (i) sojourn time i.e., control plane (C-plane) and (ii) mean waiting time i.e., user plane (U-plane). The former is responsible for the establishment of the necessary network control information such as; scheduling algorithms, rate control techniques, bandwidth reservation strategies, call admission control policies, transmitter assignment and handover [27]. Additional considerations are the time interval data packets dwell on the equipment as it travels from the PDN network equipment to the BS, i.e., backhaul communications. The network equipment includes: load-balancers, middle-boxes, routers, switches, deep packet inspection, firewalls, intrusion detection system and traffic engineering boxes. The sojourn time denotes the time interval that UE switches from idle mode to active mode and successfully establish a link between the UE and BS. Precisely, it is responsible for the network end-to-end orchestration (EEO) between the UE and BS. Furthermore, reachability, on the fly network configuration and system policies updates are all linked with sojourn time. Meanwhile, ultra-low latency 5G postulates a built-in mechanism for network fault dynamics and load changes adaptive automatic reconfiguration. The traction of EEO is to provide zero-touch service instantiation request functionality in AS and NAS resulting in tight coupling between the constituent network elements. Traditional IP network will struggle to meet the fast EEO required by the NGMN considering operations, administrations and maintenance (OAM) metrics. 5G MU-MIMO will witness unprecedented ultra-low latency of 1 ms with no more than 10^{-9} packet loss applications as highlighted in Table 2. As could be seen from the table, the vision of Industry 4.0/Factory Automation and Made in China 2025/Internet Plus are the same in terms of expected latency and acceptable packet loss rate (PLR). However, both have been driven by diverse key enabler. Furthermore, 5G will witness an era of user-defined PLR in which the backbone technology will be crucial to the offered services such as gaming, robotics, smart grid, E-health and visual learning environment.

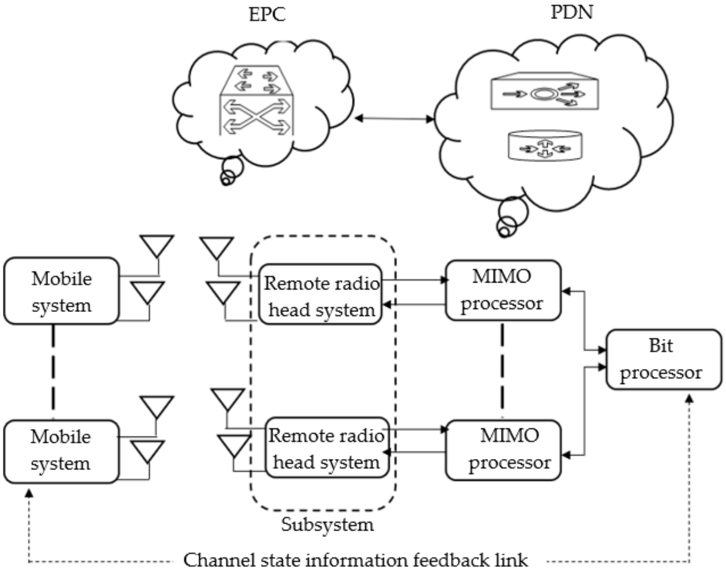

Figure 3. The system architecture of the current wireless network.

Table 2. Highlights of potential 5G applications and their requirements of ultra-low latency.

Application	Expected Latency	Acceptable Packet Loss Rate	Key Enabler
Industry 4.0/Factory Automation	0.5–5 ms [28–30]	10^{-9}	Network slicing and visual computing
Intelligent transport system (ITS)	100 ms	10^{-5} – 10^{-3}	Device-2-Device and AI
Made in China 2025/Internet Plus	0.5–5 ms [31]	10^{-9}	Motivated by Industry 4.0
Robotics	1 ms [16,32]	User defined	Haptic feedback
Virtual Learning Environment	5–10 ms [33]	User defined	Haptic communications
Tactile Internet	zero ms [32,34,35]	10^{-9}	Haptic communications
Virtual/Augmented Reality	1–4 ms	10^{-4}	Haptic environment
E-health	1–10 ms [16,36]	User defined	Tactile Internet and CODEC system

The expected latency for intelligent transport system (ITS) was stated at 100 ms in Table 2. ITS system architectures consist of four subsystems: M2M capillary, M2M access domain, M2M core and M2M application [37]. Using the LTE-A standard around which 5G technology is revolving, the four ITS subsystem has a total latency of 33.5–309 ms. In this system design, the M2M access domain, M2M core are denoted as having 13–129 ms and 12–150 ms latency, respectively. Hence, 100 ms is a reasonable estimate to be expected considering the number of devices involved. Device-2-device (D-2-D) will be a key enabler for smart grid technology considering the fact that smart grid technology is a long-distance communication system. By breaking the transmission distance into the shorter distance and deploying D-2-D communications, the latency can be reduced if amplify and forward (AF) technology is deployed. Furthermore, D-2-D will ostensibly reduce the need for retransmission in the face of delivery failure. It is believed that our Four C model can assist in the issue by deploying SDN, mMIMO and NFV. Ultimately, the MU-MIMO system faces two (2) key complexity challenges namely: implementation complexity and computational complexity [38]. The former focuses on signal overhead reduction strategies with the emphasis on the physical air interface and control communication between diverse network entities, while the latter defines the processing time of underlying algorithms. Next, we discuss these key challenges and possible solutions.

2.2. The Four (4) C Model for Ultra-Low Latency High Capacity 5G Networks.

A traditional network is constrained by latency issues which must be overcome by the proposed 5G requirements. Traditional networks are not programmable and hard to upgrade and hence making it difficult to deploy new architecture to meet the dynamic market demands [39]. To meet the demand for low latency 5G communications, there should be an interlinkage between core telecommunication networks, Internet and IT networking paradigm and recent advances in hardware and software technologies. Unlike the traditional network, the next generation of mobile technology is expected to meet customized capabilities as outlined below as:

- Elastic, scalable, network-wide capabilities.
- Network service automation, standardization and abstraction.
- Automated operations, administrations, maintenance and provisioning (OAM&P) capabilities.
- Dynamic traffic steering, hardware accelerators and service function chaining.

To achieve this goal, this paper is implemented a Four (4) C framework as shown in Figure 4. The Four (4) C model is an acronym of computing, complexity, cost and cross-layer. Details are provided in the following sections.

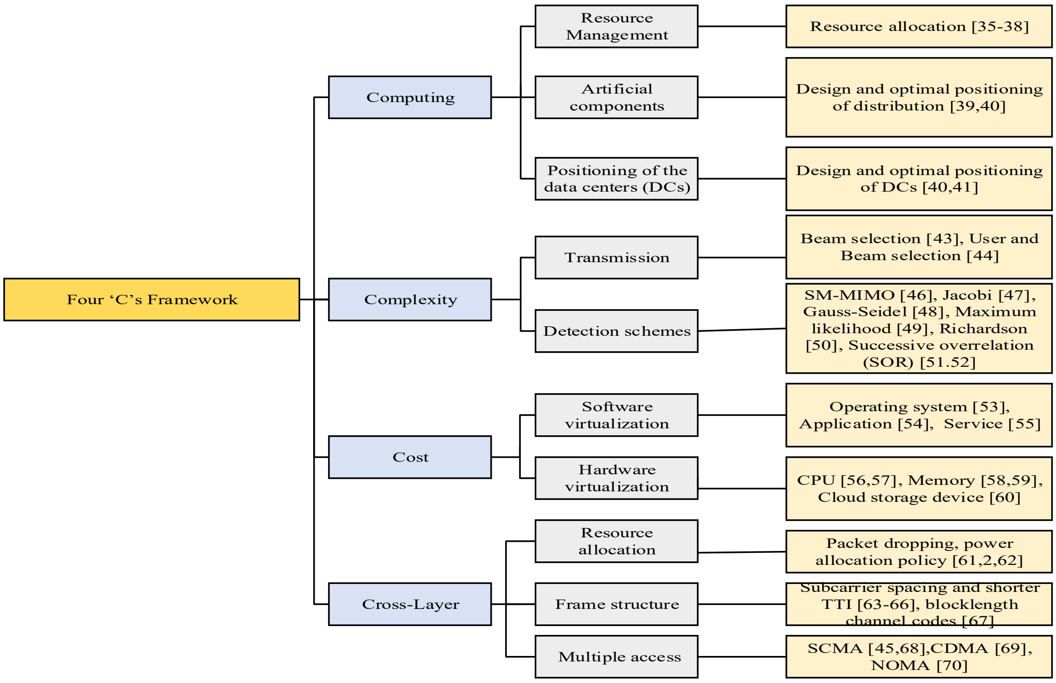
Four 'C's Framework
Computing
Complexity
Cost
Cross-Layer
Resource Management
Artificial components
Positioning of the data centers (DCs)
Transmission
Detection schemes
Software virtualization
Hardware virtualization
Resource allocation
Frame structure
Multiple access
Resource allocation [35-38]
Design and optimal positioning of distribution [39,40]
Design and optimal positioning of DCs [40,41]
Beam selection [43], User and Beam selection [44]
SM-MIMO [46], Jacobi [47], Gauss-Seidel [48], Maximum likelihood [49], Richardson [50], Successive overrelation (SOR) [51.52]
Operating system [53], Application [54], Service [55]
CPU [56,57], Memory [58,59], Cloud storage device [60]
Packet dropping, power allocation policy [61,2,62]
Subcarrier spacing and shorter TTI [63-66], blocklength channel codes [67]
SCMA [45,68],CDMA [69], NOMA [70]

Figure 4. The classification of Four (4) C Model for Ultra-Low latency high capacity 5G communications.

3. Computing

Figure 5 summarizes the major milestones in the evolution of cellular networks topology towards achieving E-2-E low latency. Besides, the literature provides a comprehensive overview about the various new cellular networks topology involving software defined networking (SDN) [40,41], network function virtualization (NFV) [39,42], and network slicing [43,44]. However, these efforts are considered insufficient to achieve latency close to zero with the Internet of Things (IoTs) limelight [45]. These dramatic increase in data traffic will lead to network congestion. Thus, making analysis, processing and storage of the data by the cloud data centers. Cloud data centres are normally characterized by having slow data rates, low bandwidth, and high latency, a very challenging task. Besides that, the real-time and medical data are an additional challenge due to the low latency (close to zero) and high-reliability requirements of data availability and processing at the core data centers [46,47]. These challenges are the enigma propelling a new computing infrastructure model called fog computing (FC), an element of the URC to achieve a low-latency. However, the FC is not intended to replace the existing CC system, but to enhance and support the existing system to significantly reduce the latency, especially with regards to real-time and medical data. Thus, the FC is an integral part of the current CC in the 5G environment, while ensuring compatibility with the new cellular networks topology [48]. Fog computing was introduced by Cisco in 2012 [49]. FC is defined as a decentralized computing infrastructure, but it is distributed over large geographical areas to handle billions of Internet-connected devices. Figure 5 summarizes the architecture of the FC, focusing on how it can be integrated with the CC and devices layer. As seen in the Figure 6, the network edge of the FC includes a set of intelligent devices such as gateways, access points, routers, and layer 3-switches that are making smart decisions to provide computation and routing functions by examining whether an application request requires communication with CC layer or not [50]. The philosophy of FC is to serve the requests of real-time, low-latency services by the FC devices, where data analytics are embedded directly within endpoints such as network nodes, data sinks, controllers, or even sensors themselves. Thus, this approach gives twofold benefits of the analytics of real-time data with low latency and energy consumption. Due to that, the fog nodes need to be as close as possible to the end users.

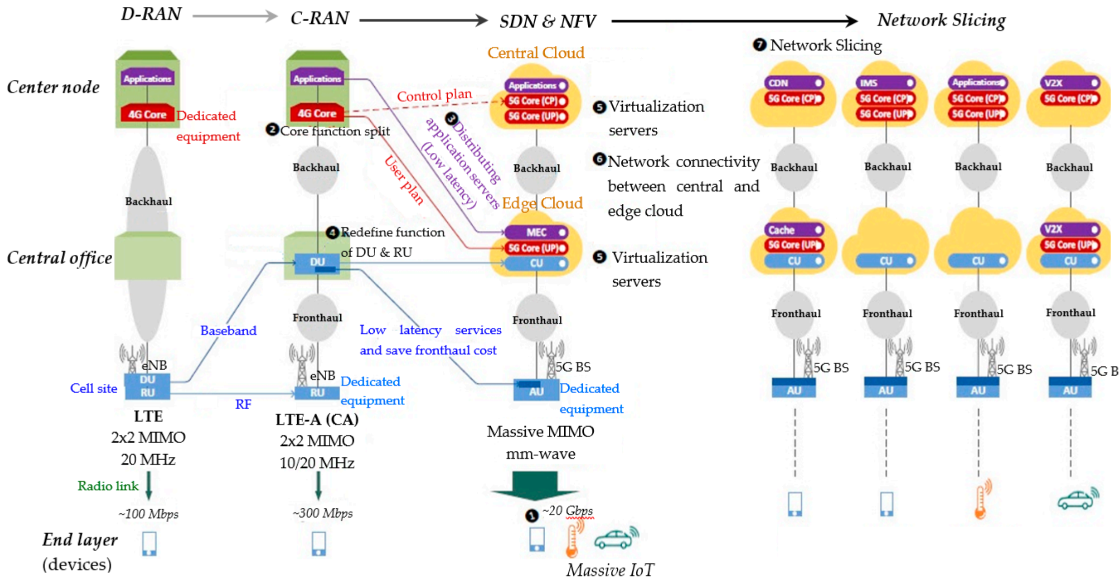

Figure 5. Major milestones of the evolution of cellular networks topology to achieve end-to-end low latency.

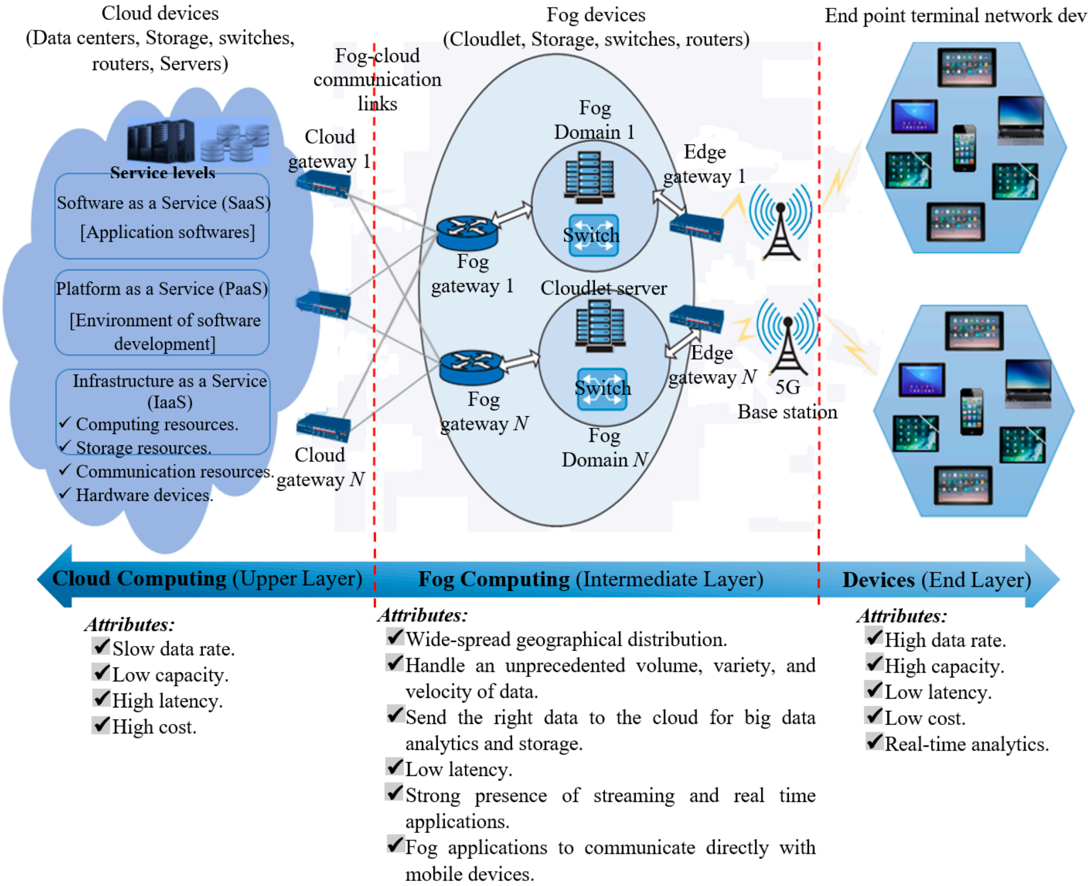
Cloud devices
(Data centers, Storage, switches, routers, Servers)
Fog-cloud communication links
Fog devices
(Cloudlet, Storage, switches, routers)
End point terminal network devices
Service levels
Software as a Service (SaaS)
[Application softwares]
Platform as a Service (PaaS)
[Environment of software development]
Infrastructure as a Service (IaaS)
✓ Computing resources.
✓ Storage resources.
✓ Communication resources.
✓ Hardware devices.
Cloud gateway 1
Cloud gateway N
Fog Domain 1
Fog gateway 1
Switch
Cloudlet server
Edge gateway 1
Switch
Fog gateway N
Fog Domain N
Edge gateway N
5G Base station
Cloud Computing (Upper Layer)
Fog Computing (Intermediate Layer)
Devices (End Layer)
Attributes:
✓Slow data rate.
✓Low capacity.
✓High latency.
✓High cost.
Attributes:
✓Wide-spread geographical distribution.
✓Handle an unprecedented volume, variety, and velocity of data.
✓Send the right data to the cloud for big data analytics and storage.
✓Low latency.
✓Strong presence of streaming and real time applications.
✓Fog applications to communicate directly with mobile devices.
Attributes:
✓High data rate.
✓High capacity.
✓Low latency.
✓Low cost.
✓Real-time analytics.

Figure 6. Fog computing and integration with the CC and devices layer.

Meanwhile, in requests that demand semi-permanent and permanent storage or require extensive analysis, the primary purpose of fog equipment is to liaise with routers or gateways and initiate the redirection of request to the core CC. Accordingly, FC is considered a natural extension of CC that aims primarily to relieve the performance bottlenecks of the network and minimize data analytics latency at the central servers of a cloud. In [50] the authors expounded the diverse system units of FC and highlighted some typical use cases scenarios. In addition, the authors emphasized the need for fog-cloud marriage and the duty of fog computing from the perspective of the Internet of Things. References [51–54] proposed several architectures for resource management of the IoT use case cutting across cloud and FC. The request, processing, and response time are considered as the performance metrics in these studies. Meanwhile, Bonomi et al. [55] studied the properties of the paradigm as it relates to latency, location knowledge, geographical spread, mobility, non-homogeneity, and the predominant access to wireless devices. Intharawijitr et al. [56] suggested a low-latency FC solution for latency management. Simplifying computing delay and communication delay, an analytical model was proposed as a guide towards the selection of entities in fog network that offers the barest delay. Xiao et al. [57] considered the issue on the strategy and optimal positioning of cloud data centres (DCs) to enhance the QoS focusing on system latency as well as cost-efficiency. It was observed that the study is positively correlated with the proficiency of the data-centric networks. Dastjerdi et al. [58], shed more light on FC focusing on principles, architectures, and applications on the perspective of Internet of Things. Meanwhile, Sarkar et al. [59] assessed the functionality of the recently suggested FC paradigm to cater for the demands of the latency-intolerant use cases in the context of IoT. Results indicate the FC outperforms the traditional CC system as the number of devices demanding real-time services increases. In scenarios with 50% applications driven by instant, real-time services, the total service latency for FC was observed to reduce by 50.09%. Additionally, references [60–62] studied the various scenarios of resource allocation as it relates to FC framework. The essence and use-cases of FC were explored by Yannuzzi et al. [63] at a shallow level. In [64], the authors analyzed divergent computing scenarios including CC and explored the desirability of having a robust and fault-resistant FC platform.

Meanwhile, Chen et al. [65] studied the issue of latency for video streaming services. The authors suggested the deployment of a unit DC under a single cyber-physical system (CPS). Nevertheless, the scenario can be obtained based on the fact that in real-world IoT deployment, a unit of DC under a unit global CPS might disrupt the macro service competence fueled by inadequate management and the lack of cloud storage. Hong et al. [66] proposed a programming model to wheel enormous IoT use-cases via mobile FC. The model embraces the service offering to geographically dispersed, latency-intolerant use cases.

In the same context, Zeng et al. [67] deployed convex programming to study service minimization completion duration of job image and job scheduling. By simply invoking distributed load balancing mechanism between the client and fog entities, the total computation and transmission latency can be reduced. In addition, the job accomplishment duration convex problem was re-modeled as a mixed-integer nonlinear mathematical programming problem and solution obtained via a low-complexity three-stage algorithm. Oueis et al. [68] suggested an excellent start-up strategy for nodes association in which the management roles are collaboratively performed by many entities within the latency limitations. However, these studies on FC are primarily limited on the fundamentals and the principles; and have not digressed into the technical domain resulting in implementation.

In addition, recent researches discussed several of the important aspects of fog computing such as: location awareness [69], security and privacy [70], geographical distribution, mobility [55], energy consumption and OPEX [59], heterogeneity, real-life applications, and the predominant access to wireless devices [58]. Herein, we will pay attention to the latency issue as the major contribution and a new approach to this study.

Obviously, in use cases driven by low-latency communications, FC results in high overhead when compared to traditional CC. Hence, studies extracted using a high amount of IoT latency-sensitive shows that FC exhibits superior performance to CC. A summary of the state-of-the-art is given in Table 3.

Table 3. Summary of the state-of-the-art for 5G computing paradigm.

References	Contribution	Results	Limitations
Xiao et al. [55]	Considered the issue of formation and optimal placement of DCs to enhance QoS in respect to service latency and cost minimization.	The placement of DCs is positively correlated by the data centric networks performance.	Complex optimization model where the locations of DCs strongly affect the efficiency of the data centric networks.
Chen et al. [65]	Targets the issue of video streaming services latency.	Suggested the concept of a unit DC controlled by a unit CSP. In this setting, the total computation and transmit delay needs could be reduced by parallel computing jobs and aligning the tasks on both client and fog nodes.	A unit DC controlled a unit global CSP may obstruct the total service performance as a result of inadequate management and lack of cloud storage.
Yangui et al. [51]	Proposed a stratum driven system topology for IoT application provisioning comprising of cloud and FC. Secondly, advanced design nomenclatures and interfaces were exposed, with E-2-E latency as a performance metric.	The E-2-E latency was reduced to 484 ms on the occasion of fog proximity to the IoT devices. Nevertheless, it is of interest to note that the worst result of 2033 ms is not when all the components are in the cloud.	The prototype needs more investigation on the further performance index, including fire detection delay and robot dispatching delay.
Agarwal et al. [52]	Proposed resource allocation topology comprising of an algorithm that allocates the tasks between the cloud and FC. The demands, processing, and response time are considered as performance metrics.	The proposed algorithm has shown that the response time was 309.53 ms. In addition, the response time of reconfigurable load balancing was 632.87 ms and the overall response time to optimizing response time was 630.11 ms.	Targets the fundamentals, core notions and the doctrines of fog computing; and not in the technical aspect from an implementation perspective.

Table 3. *Cont.*

References	Contribution	Results	Limitations
Krishnan et al. [53]	Proposed a solution comprising of the fog and CC, where latency is considered as a performance metric.	The E-2-E delay was reduced by 70% compared with traditional Cloud.	The prototype needs more investigation and deeper discussion, especially the technical aspect from an implementation perspective.
Hong et al. [66]	Based on a mathematical programming model, a solution was derived to support large-scale IoT use-cases via mobile FC.	The model supports the service provisioning to location dispersed, delay-sensitive use-cases.	The main issue is to formulate a parallel runtime network capable of migrating Mobile Fog processes across wide spectrum of devices.
Bonomi et al. [55]	Focused on the properties of the concept in terms of delay, geographical awareness, location dispersion, mobility, non-homogeneity, and wireless access to devices.	The FC interacted well with machine-to-machine (M2M) and decreased the latency of the real-time processing from milliseconds to sub seconds.	Fog devices considered localization, hence allow small latency and context awareness, the cloud provisioning and global centralization.
Zeng et al. [67]	Studied service requests completion duration reduction issue by focusing on job image positioning and job scheduling jointly; and the job completion duration reduction mechanism.	The total compute and transmit delay could be reduced by sharing computing jobs and balancing the task on both client and fog nodes.	Focused on the principles, basic notions, and the doctrines of fog computing; and not in the technical aspect from an implementation perspective.
Intharawijitr et al. [56]	Suggested a low-latency FC architecture for latency management. To simplify computing latency and communication latency, a mathematical model was defined.	The smallest delay policy offers appreciable performance based on the speed of resource availability. Additionally, the authors discovered there was an optimal value for the latency threshold	The proposed optimization model is preliminary and requires consideration of different applications running on the source to enable more accurate analysis of a real network environment.
Sarkar et al. [59]	Assessed the applicability of the novel suggested FC concept to service the needs of the latency-sensitive applications in the context of IoT.	Fog computing outperforms CC in the context of IoT, with a high number of latency-sensitive applications.	Deeper study needed concerning different applications running on the source to enable more accurate analysis of a real network environment.

4. Cost: Computing Resource Factors

Virtualization is not a new technique, but recently with the advent of CC and big data concept, it has become a staple in CC design [71]. Virtualization techniques in CC permit the execution of several applications and operating systems utilizing same server, without obstructing the optimal performance of similar applications running on the server or host machine, resulting in judicious resource utilization, costs minimization, and lowering the latency via speeding-up of the transfer procedure of the virtual server files by a software known as Hypervisor, as seen in Figure 7. Thus, the hypervisor is a fundamental part of the virtualization infrastructure. The hypervisor plays the role of a bridge connecting the hardware and the virtual environment and distributes the hardware utilities such as CPU usage, memory allotment between the various virtual platforms.

The hypervisor is equipped with small scale virtual server managing attributes, comprising higher virtual server's capability or booting it down [72]. The VM offers a wide spectrum of attributes for managing many hypervisors across physical platforms. By design, a hypervisor is restricted to one physical server and hence, only add virtual images of the underlying server. On the occasion that a virtualized platform is instantiated, the hypervisor configurations are then transported to the memory. AMD Processor and Intel VMX are main hardware that facilitates in the transfer of the VMs and hypervisor. With the creation of a VM, the CPU is equipped with the execution commands of the hypervisor, which subsequently acts on the information located in memory [58]. Kim et al. [73], analyzed various categories of existing hypervisors, such as Citrix Xenserver, VMware ESXi, KVM, and Hper-V hypervisor. The researchers proposed a system labelled as VM placement to address the problem of access latency recommended for specific VM configured on non-uniform memory access scheme. Hypervisors can be grouped according to different nomenclature of; (Para-Virtualization Hypervisor, Fully Virtualized Hypervisor, Hybrid Model Hypervisor, and Micro-Kernelized Design Hypervisor) [55]. Virtualization technique is deployed at the computing, storage, network, and application levels. Thus, virtualization can be classified into two main categories [74], which are software virtualization and hardware virtualization.

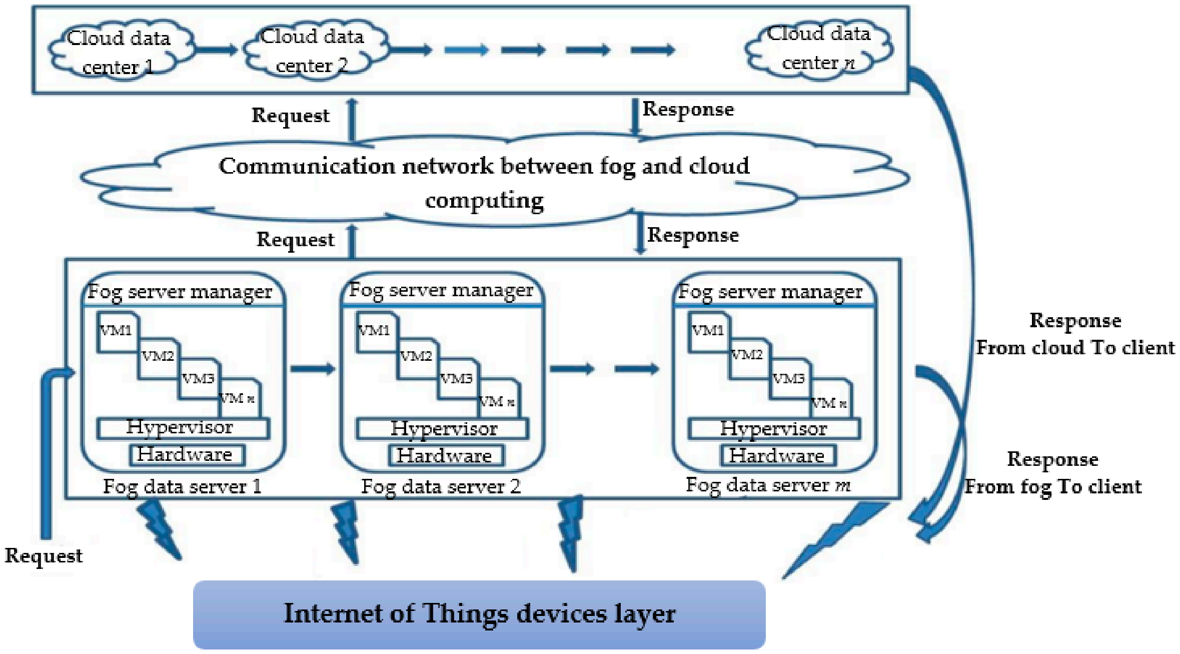

Figure 7. Scheme of the hypervisor mechanism process in a cloud.

4.1. *Software Virtualization*

Software virtualization embraces the concept of having of various virtual environments on the host machine; and can be classified into:

- Operating system virtualization (OSV): hosting multiple operating systems on the native operating system.
- Application virtualization (AV): endowed with the capability to host singular applications in a virtual environment separate from the native OS.
- Service virtualization (SV): relates to hosting service specific processes and services related to a particular application on a virtual platform.

4.2. *Hardware Virtualization*

Often referred to as hardware-assisted virtualization or server virtualization is executed on the paradigm that a singular autonomous section of hardware or a physical server, can be construed as various smaller hardware sections or servers. Thus, resulting in an amalgamation of various physical servers into virtual servers that drive a single primary physical server. Each minute server is capable of hosting a virtual machine (VM), however, the entire cluster of servers are seen as a solitary entity by any process associating with the hardware. It is the duty of hypervisor to execute resource allocation algorithm. The advantages of the aforementioned strategy are the higher processing power derived based on efficient hardware utilization and service uptime. Besides, hardware virtualization can be applied into CPU, cloud storage device, memory, and switches in order to speed up the response of CC [75]. Further details on Vcpu, Vram and vDisk is explained below:

4.2.1. VCPU

A virtual CPU (Vcpu) or virtual processor is a physical CPU core allocated to a virtual machine. It is the hypervisor's processing power made available to the virtual server. Even though vCPUs can be viewed as different cores, the amount of Vcpu per physical CPU (or core) must be calculated. Four to eight vCPUs can usually be allocated to each physical core to accommodate varying workloads. There can be more virtual processors assigned than actual physical cores available, permitting a single core to be accessed by the virtual machines [76,77]. Nishio et al. [78] in turn studied a similar case, however, they limited its study to only a mobile fog system. The topology consists of a fog stratum made up of mobile devices and a cloud accessed by a mobile cellular network. Consequently, the study focused on CPU efficiency, bandwidth, and facility allotment to service computing needs. Driven by the non-homogeneity resources, a translation was designed via time resources. Thus, making it possible to quantify them via a similar unit. The problem was formulated under two objectives: (i) sum maximization and; (ii) utility product function maximization problem. It was solved via convex optimization.

4.2.2. Memory

In an environment consisting of virtual devices, the norm is to allocate the physical memory to the virtual physical memory. Inspired by the need to allocate extra memory to a virtual server, the virtual memory management strategy is evoked. Virtual RAM (vRAM) denotes the quantity of RAM a hypervisor assigns to a virtual server. Be informed that not all assigned vRAM is physical RAM. The hypervisor may assign physical memory and disk space concurrently to satisfy the vRAM requirements. This implies a dual-stage translation process should be kept by the image OS and the virtual machine monitor (VMM), correspondingly: virtual memory to physical memory and physical memory to machine memory vice-versa. Additionally, a memory management unit (MMU) virtualization is encouraged and must not be made opaque to the guest OS. The guest OS never ceases to administer the translation of virtual addresses to the physical memory addresses of VMs. But the guest OS cannot directly access the actual machine memory. The VMM is saddled with the responsibility of translating the guest physical memory to the actual machine memory [79,80].

4.2.3. Cloud Storage Device

The cloud storage device (CSD) system signifies storage entities specifically formulated for the cloud-based network. A key issue that must be addressed in cloud storage is the security, integrity, confidentiality, as well as low latency during transfer the data [81]. The CSD performance monitor system is deployed to maintain pre-defined targets. The CSD performance monitor system is equipped to undertake additional functions, including routinely validating the present position of datasets compared to the pre-defined objectives. This is essential to ensure datasets always reside in a CSD that satisfies its requirements. In case of any abnormalities, an alert is transmitted to the CSD that is not in accordance with the cloud consumer's requirements. It is possible to virtualize these devices just as how physical servers can create virtual server images. These techniques can easily provision a fixed-increment storage allotment supporting the pay-per-use topology [82].

On the other hand, a virtual disk (vDisk) denotes a strategy deployed in the CC to enhance the frequency of relocating the virtual server files via decomposing vDisk into reduced bits that signify the virtual server's hard disk. A vDisk is the amalgamation of hard drives assigned to a virtual server before or after its creation. The various vDisk is kept in a solitary file, or multiple of files, via formats executable by the hypervisor. If the Microsoft Hyper-V is used, the vDisks will be stored in a .VHD file format. While, if the VMware ESX(i) is used, the hard disks will be stored in a .VMDK file format [83]. Figure 8 shows three virtual servers (VM A, VM B, and VM C). VM A has two vDisks arranged as a solitary file; VM B has two hard disks, one is kept as a solitary file and the other as a split disk, while VM C has two hard disks stored as split disks.

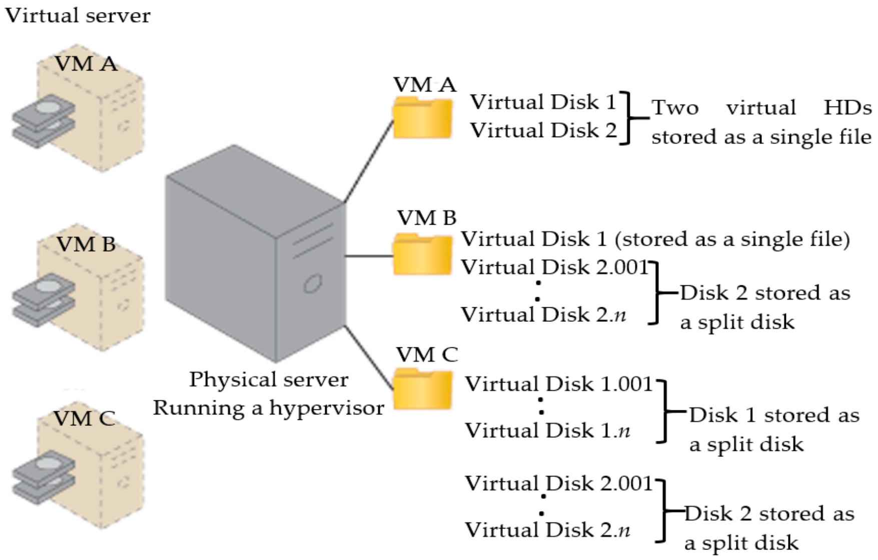

Figure 8. Virtual disk technique.

5. Complexity

Low complexity algorithms have fast convergence rates, thus expediting the execution of fast resource allocation algorithms. There is a tradeoff between high complexity algorithms, low latency algorithms and most probably, algorithm accuracy. It is widely acknowledged that turbo encoder algorithms used for channel line coding exhibit superior performance at the expense of high computational complexity and low convergence. User cases with high interference tolerance capability such as LTE applications are assigned low complexity and fast convergence algorithms. Hence, otherwise. From the algorithmic perspective, 5G system designers might need to rely on the big O($\mathbb{N}$) complexity indicator for choosing algorithms. The term $\mathbb{N}$ indicates arrays cardinality. Using the big O($\mathbb{N}$) complexity indicator, 5G system designers can infer if the algorithm complexity is linear, logarithmic, quadratic or constant. It is widely known that constant complexity algorithms have lower complexity and faster execution time than quadratic algorithms.

Massive MIMO system is a significant technology in 5G wireless communication systems where they offer many potential solutions in achieving high data rates as well as robustness to mitigate fading, hardware failures and interferences [84]. Massive MIMO consists of unprecedentedly large antenna arrays that potentially provide high spectral efficiency and high channel capacity, however, this could also lead to hardware and computational complexity [85]. The installation of huge numbers of antennas at the base station could lead to a high cost of implementation and increase the amount of power consumption due to a huge number of radio frequency chains, analogue-to-digital converters, digital-to-analogue converters, power amplifiers and numerous transceivers. In addition, the number of antennas in the detection scheme will raise the computational complexity and thus, increase the latency of the systems. Hence, to mitigate the complexity issue, one of the solution is by using RF chains to reduce the complexity and energy consumption by employing a Hybrid Analogue-Digital transceiver [86]. The phase shifter is also one of the solutions by introducing successive interference cancelation (SIC) hybrid precoding with sub-connected architecture [87].

To resolve the issue of the RF chain hardware limitations [88], a variable phase shifter with high-dimensional phase only radio frequency processing is fully utilized in order to control the phases of the up-converted radio frequency signal [89,90]. In achieving the optimum performance in massive MIMO systems, several methods can be deployed such as hybrid precoding scheme, zero-forcing (ZF) and minimum mean square error (MMSE) methods [91]. Moreover, in order to obtain practically ideal performance of a singular value decomposition, hybrid methods are used in [92,93] where the concept of orthogonal matching pursuit (OMP) is decomposed of optimal precoder and combiner. Method to compensate for the overall performance and complexity is by introducing a linear minimum mean square error (MMSE) signal detection. Table 4 shows a summary of previous works on the low-complexity.

Table 4. Summary of previous works on the low complexity method.

References	Contributions	Results
Valduga et al. [94]	Proposed beam selection scheme by exploiting the geometric sparsity of the multi-user massive MIMO to mitigate complexity in [94].	Proposed scheme improved the overall sparsity channel performance.
Yang et al. [95]	Introduced low-complexity Spatial modulation (SM) SM-MIMO schemes.	Proposed schemes can be capably quantized and this strategy is suitable for limited feedback systems.
Jiang et al. [96]	Researchers formulated a study on reducing overall complexity by proposing a matrix-vector product initialization and iteration steps.	The total computational complexity system bit error rate is significantly reduced.
Qin et al. [97]	Jacobi method is studied in [61] to accelerate the convergence rate while maintaining the low complexity.	Proposed method outperformed Neumann Series, Richardson method and conjugate gradient-based methods.
Dai et al. [98]	Approach exploited Gauss-Seidel (GS) method and which is iteratively realized via the minimum mean square error (MMSE) algorithm.	It achieved the near optimal performance of the conventional MMSE algorithm with small numbers of iterations. It also outperformed the Neumann series approximation algorithm.
Jiang et al. [99]	Proposed fast processing algorithms by transforming the large-scale matrix inverse into linear equations. Also, the properties of the block matrix are utilized. Lastly, individually updated the small size block.	At low latency and low complexity, the overall results showed a good system performance.
Vikalo et al. [100]	To significantly simplified the computation, a sphere decoding is presented in [63] to mitigate the increasing computational complexity of maximum likelihood (ML)	Significant higher performance gains are achieved as compared to the heuristic method in [101].
Gao et al. [101]	Based on the Richardson method, a MMSE method is proposed.	Outperformed Neumann series approximation algorithm as well as achieving a near optimal performance of the conventional MMSE.
Gao et al. [102]	Presented an algorithm based on successive overrelaxation (SOR) where the results showed an almost similar to the MMSE method in achieving low complexity signal detection.	Achieving the near optimal performance of the conventional MMSE algorithm with small numbers of iterations. It also outperformed the Neumann series approximation algorithm.
Xie et al. [103]	Proposed a systemic successive overrelaxation (SSOR) based precoding scheme.	The result showed similar to the results based on the ZF method
Wu et al. [104]	Two-stage user scheduling scheme (i.e., user classification at the first stage; User and beam selection at the second stage) is proposed by considering the correlation between the channel energy and inter-user channel.	The results validated the proposed scheme has reduced computational complexity as well as achieving a higher sum rate.

Table 4. *Cont.*

References	Contributions	Results
Liu et al. [105]	The proposed system able to decrease the intra-cell interference with simple signal processing and it is due to the finer spatial resolution that is achieved by a large number of antennas at the BSs.	The schemes provide very close to favorable performance while the computational complexity is greatly reduced.
Kim et al. [106]	Proposed a weighted minimum distance (wMD) decoding by reviewing the multi-user multiple input multiple output detection problem into an equivalent coding problem.	Almost achieve the favorable performance of the wMD decoding with a much lower complexity of the decoding.
Sabeti et al. [107]	Proposed a low-complexity carrier frequency offset compensation technique.	Interference matrix can be calculated diagonalized due to the circulant property and thus the inverse matrix can be obtained straightforward.
Wu et al. [108]	Proposed a low-complexity and hardware efficient signal detection algorithm as well as a VLSI architecture. The architecture is said to be scalable and easily reconfigurable according to the increasing the numbers of the antenna.	The reduction of processing latency per iteration is obtained.
Ahmed et al. [109]	The authors proposed a hierarchical codebook search algorithm.	Provides good trade-off between complexity reduction and performance as well as robustness against feedback channels errors.
Minango et al. [110]	Based on the first order Neumann Series (NS) expansion, MMSE detectors are introduced in order to lower down the complexity in the approximate inverse matrix.	The proposed results showed the equivalent of performance that better than the others approach in the literature reviews.
Minango et al. [111]	Proposed a Damped Jacobi method for the reduction of complexity MMSE detector algorithm in their study.	Without performance loss, the proposed method was able to reduce the classical MMSE detector complexity by one order of magnitude.
Haghighatshoar et al. [112]	The proposed algorithm is reminiscent of Multiple Measurement Vectors which also has low complexity and able to track the sharp transition in the channel statistics. It is also proved to be similar to the Approximate Maximum-Likelihood (AML) algorithm.	Improvements in subspace estimation.
Shikida et al. [113]	The throughput performance of zero-forcing and MMSE based coordinated beamforming is evaluated in parallel to the changing of the number of spatially multiplexed users.	The results showed the throughput performance of the MMSE based scheme is 23% higher than the zero-forcing scheme.
Zhang et al. [114]	The authors adopted the Landweber method algorithm.	The results show better performance as compare to the existing algorithm.
Tang et al. [115]	The authors developed a low latency and complexity of massive spatial modulation MIMO detection scheme with zero-forcing (ZF) and Maximum Ratio Combining Zero Forcing (MRC -ZF).	The results showed significantly improved performance in terms of latency and SINR as compare to the existing methods that been reviewed in the paper and it is validated by using Software Defined Radio (SDR) platform.
Lei et al. [116]	The researcher introduced the Sparse Code Multiple Access (SCMA) to enhanced low complexity in receiver design. The test design is simulated and verified in real-time prototyping.	The results showed that it is three times outperformed the overall throughput system while sustaining the low latency transmission as homogenous to flexible orthogonal.

However, the method presented in [90] causes high complexity particularly in a huge number of users due to the full matrix inversion method. MIMO pilot contamination approach has been proposed as a possible technique to reduce URLLC in 5G communications [117].

6. Cross-Layer

Another technique capable of resulting in low latency 5G era is a cross-layer approach. As stated by Srivastava et al. [118], cross-layer is referred to as designing a protocol by manipulating the dependence between protocol layers to achieve the optimum performance gain. The cross-layer approach offers the system designer a comprehensive overview of the network topology leading to a compact and interoperable architecture. Cross-layer designs can be seen from two perspectives: tight coupling and loose coupling. In tight coupling, all the system layers are highly interconnected to each other, leading to faster EEO and OAMP. Although this approach on the surface is ideal, it comes with some limitations. The most notable limitation is that it creates an entry barrier to more innovative solutions from other vendors. Conversely, the loose coupling seems attractive to the 5G low latency regime and the open system nomenclature has gained traction for this. The open system both from the hardware and software perspective has created an avenue to maximize the plug and play syndrome and code reusability with attendance resulting in a faster system deployment scenario.

One way of achieving information sharing between all the cross-layer designs is to leverage the interdependencies and interactions between different layers of the networking layer. In cross-layer design, each layer is assigned one or more roles in the architecture of wireless communication. Each layer is given a set of roles to function similar to the TCP/IP protocol stack and the OSI layer. To enhance the performance, the dependence found across different layers of roles in the cross-layer design is utilized [118]. High service throughput can be achieved by massive MIMO technique where the technique deploys large scale of the antenna at the BS and it provides full-dimensional spatial multiplexing as well as a high order of beamforming gain [119,120]. In achieving low latency, short orthogonal frequency division modulation (OFDM) symbol length and wider bandwidth are retained based on the OFDM-based waveform [121]. Some of the related previous works proposed methods such as channel vector correlation coefficient [122–125], real-time sum capacity [126–128], an approximation of polynomial [129] and asymptotic capacity [130].

As a side note, delays caused by transmission and signaling as mentioned in [131] can be reduced by using short frame structure as stated in [132]. Delays caused by transmission and signaling also can be reduced by using a polar codes coding scheme [133]. The performance of tactile internet can be investigated based on statistical queueing requirements, an effective bandwidth [134], as well as effective capacity is used [135]. In [136], a practical packet dropping along with finite power mechanism is proposed to enable the quality of service (QoS). Similarly, a framework for cross-layer is established by assuming a frequency-flat fading channel model, optimization of the power allocation as well as the packet dropping mechanism [137].

Transmission queueing delay violation probability, proactive packet dropping probability and error probability are considered in the reliability of the system in [138]. Moreover, it is also validated by simulation and numerical results where the results are near-optimal performance. In [139], a multiple resource block (RB) are implemented simultaneously, the introduction of power-based weight in RB scheduling, and the low complexity calculation method are applied at the per-layer which is based on the Hermitian Matrix Blockwise Inversion Dilemma. Other researchers in [140] proposed a low latency radio interface by using 180 kHz and 360 kHz OFDM subcarrier spacing with 256 and 128 points fast Fourier transform (FFT). Other findings in [141] in regards to the low latency in ultra-dense small cell networks proposed a frame structure based on 312.5 kHz OFDM subcarrier spacing. Subsequently, to reduce latency in TDD air interface by exploiting the OFDM subcarrier spacing and the frame length should be reduced to 0.25 ms [135]. Ref. [142] has proposed an algorithm that considered a new weight function for adapting handover margin level over contiguous carrier aggregation deployment scenarios in LTE-Advanced system. Related works of low latency in cross-layer are summarized in Table 5.

Table 5. Summary of previous works on the cross-layer design.

References	Contributions	TCP/IP	Results
She et al. [5]	The authors proposed a framework for cross-layer optimization where they optimize the packet dropping policy, power allocation policy and bandwidth policy.	Not mentioned	The numerical results showed that minor power loss in packet dropping probability, transmission error probability and queueing delay violation probability and when all the parameters are having the same order of magnitude.
Song et al. [143]	This paper introduced a balanced design for uplink and downlink control channels by making a proper selection in modulation, time resources and as well as the diversity scheme. Also, a radio frame structure is proposed.	Packet switching protocol	The link-level simulation showed that the proposed control channel console with the block-error rate of 10^9 under Rayleigh fading condition at SINR compared to the current 4G system.
Au et al. [144]	The console of massive connectivity is deploying an uplink contention-based sparse-code multiple access (SCMA) scheme.	PHY/MAC protocol	The simulation results showed a 2.8 times gain using contention-based sparse-code multiple access (SCMA) over a contention-based OFDMA.
Kela et al. [145]	Proposed a novel frame structure for the radio access interface to support multi-user spatial multiplexing, low latency on the radio access interface.	Asynchronous HARQ for LTE protocol	The results lead to a better performance of the spectral efficiency by a factor of 2.4. In addition to that, the average latency is equal to the factor of 5. Meaning, it has a shorter time than 1 ms.
Pedersen et al. [146]	The authors presented a configurable 5G time division duplex. In wide area scenarios, they also introduced the flexible scheduling framework.	Full Duplex protocol	The result presented in this paper showed that the achievement of different targets in flexible scheduling depends on the user's condition.
Wirth et al. [147]	The authors presented a concept of the frame structure which is supported by different levels of ultra-low latency by combining the subcarriers spacing and shorter transmission time interval (sTTI) frame structure in a clean state design.	Internet control message protocol	The proposed concept achieved delays below 1ms or 2 ms for the round trip time (RTT) results.
Liu et al. [148]	This paper presented a Cross Sliding Window (CSW) scheduling method alongside with Cross Parallel Window (CPW) scheduling method in achieving less memory capacity & lower latency.	Not mentioned	Proposed CSW method has shown a high throughput and low latency with higher efficiency.
Jang et al. [149]	This paper presented an Inverse Fast Fourier Transform (IFFT) design method by reordering of IFFT input data from the resource element mapper towards input signal of IFFT.	Medium access network, packet data convergence protocol, radio link protocol	The proposed algorithm results has shown that the Inverse Fast Fourier Transform (IFFT) is reduced in terms of the memory depth and output data latency.
Ohseki et al. [150]	The researchers proposed a fast outer loop link adaptation scheme based on the reception of acknowledgment and negative acknowledgment for the initial transmission of hybrid ACK.	Not mentioned	The results showed that high throughput is achieved right after user equipment deployed the data transmission as compared to the conventional methods.
Moreira et al. [151]	The authors introduced a software defined wireless network (SDWN)-enabled fast cross authentication scheme combined with a non-cryptographic and cryptographic algorithms.	Authentication and key agreement protocol	The proposed scheme has fulfilled and verified by simulation for the 5G security requirements.
She et al. [152]	A short frame structure as such, a transmission error with finite block-length channel codes should be considered, a packet dropping mechanism, optimization of a queue state information and as well as channel state information (CSI) dependent transmission policy are proposed.	User-plane and control-plane protocol	The results showed that the optimization magnitude is in the same order for the three probabilities which are the packet error probability, queueing delay violation probability and packet dropping probability.

Table 5. *Cont.*

References	Contributions	TCP/IP	Results
Mathur et al. [153]	The authors proposed an underlying CDMA transmission technique and the IoT will not wait for access of small size uplink transmissions.	MAC/PHY protocol	This paper aims to achieve significantly lower control plan signaling and low latency by designing the transmission technique.
Wang et al. [154]	This paper presented non-orthogonal multiple access (NOMA) that translate the physical layer of NOMA to improve the QoE in the upper layer.	MAC/PHY protocol	The results depicted that the QoE-aware NOMA is able to fit the diverse demands of users and hence, it provides better service quality for the users that comes with higher quality preferences.
Miyim et al. [155]	The authors proposed an intelligent vertical handover algorithm in HetNets and it is said to take into cognizance velocity, the current received signal strength (RSS) and predicted the RSS of candidate networks.	Not mentioned	The proposed algorithm saved time (low latency) and it identifies the best candidate network and hence, the technique showed a good ground to minimize ping-pong in HetNets.
Dombrowski et al. [156]	The researchers presented an EchoRing, or also known as wireless token-passing in MAC protocols.	Token passing MAC protocol	The proposed protocol showed better performance by several orders of magnitude in terms of reliability for latencies (below 10ms) compare to the other schemes.
Pocovi et al. [157]	The researchers extended research on the different MAC layer enhancement in supporting the ultra-reliable low latency communication (uRLLC) by introducing low pass filtered interference information at CQI report. In addition to that, a short transmission time interval (TTI) and faster processing at the user equipment to ensure the reduction of delay during hybrid automatic repeat request (HARQ) retransmissions.	Asynchronous HARQ with chase combining protocol	The results shown that 99% of latency and reliability are achieved at the low load situation and vice-versa.

7. Challenges and Opportunities in Ultra-Low Latency Massive MIMO

Following are challenges related to ultra-low latency massive MIMO based on the 4Cs framework.

7.1. Computing

Despite the advantages provided by the computing center such as increased organizational agility and improved scalability, there are some key challenges to consider. This section presents the key challenges and opportunities for computing services, which are categorized into four areas: (1) Security and privacy; (2) Resource management; (3) Scalability, and (4) Complexity [158].

7.1.1. Security and Privacy

Data security and privacy are always high on any list of priorities. Predominantly, there are four types of data services in FC: data storage, data sharing, data query and data computation. Consistent with the aforementioned four data services, they uniquely need a divergent specific data security and privacy requirements [159]. Table 6 summarized the security and privacy requirements for the main data services. These security and privacy processes are necessary; however, they may also increase the latency.

Table 6. Security and privacy requirements for the data services.

Data Services	Security and Privacy Requirements
Storage	• Integrity confirmation. • Reduced overhead. • Public inspecting. • Dynamics provisioning
Sharing	• Authorization cancellation. • Access competence.
Query	• Security tight search ability. • Dynamics provisioning. • Refined results.

7.1.2. Resource Management

Resource management in CC is aligned with random tasks, presenting a significant issue to the elasticity of CC. The fluctuation incident can be observed in two ways. One is a premeditated spike, and another is an inadvertent spike in tasks. The former, occurs in scenarios that are predictable resulting in advanced resource allocation. While the latter, resources needs to be assigned when there is a need and reassigned when required. The general framework essential in cloud resource management are (i) Admission control: is solely responsible if can admit a job/request to be executed in the cloud, (ii) Resource allocation: maps VMs onto Physical Machines (PMs) and tasks onto VMs, (iii) Quality of Service (QoS): connotes to metrics such as response duration, running cost, throughput, minimization of loss and so on, (iv) Tasks balancing: task balancing of schedules based on the resources so as to enhance its efficiency, and (v) Energy Management: connotes to optimal usage energy in the DC. Resource allocation in the cloud can be divided into two types [160]:

(a) Mapping of VMs onto PMs

Cloud based resources comprise of the software and hardware needed to implement user tasks such as memory, CPU, bandwidth, storage and network. Resource allocation denotes the act of assigning optimal utilities to the demands required by the user, these schedules are executed optimally. Conversely, resource allocation in the CC context connotes assigning a Virtual Machine that meets the system specifications requested by the user.

(b) Mapping of Workloads onto VMs

The cloud consists of a class of prevailing VMs and a built environment with predefined memory, CPU and bandwidth. The users tender their demands which may be fluctuating, and deadline driven. These jobs need to be assigned with optimal resources such that the workloads are processed efficiently. This type of assignment is known as mapping of workloads onto VMs.

7.1.3. Scalability

The number of IoT-driven devices is estimated to be in the billions of folds. Inadvertently, such an enormous volume of data will require a huge volume of resources such as processing power and storage. Consequently, fog servers are expected to provide connectivity for all these devices with adequate resources. The main challenge will be the ability to attain to the swift progress in IoT devices and use cases [161].

7.1.4. Complexity

The spectrum of IoT and wireless technology are dominated by many manufacturers who operate on different design specifications. It is plausible that this will result in difficulties in selecting an optimal component as different vendors utilize different software and hardware configurations to meet their personal requirements. Furthermore, in some scenarios, user-cases with high-security demands require specific hardware and protocols to operate. Thus, escalating the already existing operational and latency issues [112].

7.2. Cost of Management the Cloud Resources

This section presents various issues pertaining to the cost and management of cloud resources in order to ultra-reliable and low-latency.

7.2.1. Costs Reduction and Maximization of Resource Allocation Utility

The two notable challenges that must be satisfied in executing a resource allocation algorithm are the reduction in total operational cost and maximization of resource utility. A fault tolerant computing system is expected to provision services to its subscribers with the notion of continuity in-terms of providing the same service. Motivated by this, the service providers are expected to offer the subscribers services at a reduced cost. To achieve this, an excellent mechanism to administer resource usage and minimize the cost for the subscribers should be implemented. Nevertheless, there is a need to implement judicious utilization of the service provider resources.

7.2.2. Predictable and Unpredictable Workloads

Computing data centers consists of mainly virtualization resources such as; CPU, storage and network. These virtualized resources expedite low power in comparison to traditional data centers. Virtual Machines (VMs) are the class of virtualized platform offered to the subscribers. These VMs are allocated to big random jobs. The demands may fluctuate rapidly and escalate the need for an application intensifies. These class of jobs is denoted as predictable and unpredictable jobs.

7.2.3. Homogeneous and Heterogeneous Workloads

Two classes of jobs exist in the cloud broadly denoted as homogeneous and heterogeneous jobs. In homogeneous workloads system configuration, the CPU, RAM, storage and time are equally allocated. Cloud systems must be designed considering both types of workloads to be allocated.

7.2.4. VM Migration

VM migration is a subset of the techniques available for handling insufficient resources in the cloud. To increase resources availability, VMs can be migrated between hosts.

7.2.5. Parallel Task Scheduling

Parallel computing can increase the make span of the jobs. Two categories of jobs exist; independent and dependent jobs. Independent jobs can be programmed to be implemented via various virtual machines in parallel. Although, dependent jobs have inherent communication challenges, consequently, extra caution is required during design.

7.2.6. Elasticity

Elasticity in the cloud denotes the degree of flexibility in handling the unpredictability in resource demands. The demand for resources may grow asymptotically over time. The onus is on the cloud to automatically detect the demand and act accordingly. Optimal resource management in the cloud should have been equipped with such functionality.

7.3. Complexity

The ultra-reliable low latency massive MIMO complexity challenges have been presented by Shen et al. [162]. The authors studied the capacity-based user selection and Fronbenius norm-based algorithms. Both algorithms are combined with block diagonalization precoding for MU-MIMO for maximising the total throughput in the selected user set. Therefore, below are the challenges in uRLLC:

(a) Capacity-based algorithms still have high computational complexity due to their frequent use of SVD in the channel matrices.

(b) In vehicle-to-infrastructure (V2I), Jin et al. [163] has mentioned that the challenges of deploying low complexity ultra-low latency massive MIMO can meet the requirement through dynamic vehicular channels since the number of vehicles in services are significantly high.

(c) Linear Zero Forcing (ZF) or minimum mean square error (MMSE) receivers are more favorable in a real-world implementation. However, those schemes are computationally costly. However, these fast processing time and low complexity in vehicular to infrastructure make it appropriate for the applications designed for the modern vehicular using massive MIMO.

(d) In addition to that, Sabeti et al. [107] have stated several challenges in the area of computational complexity, such as an algorithm for joint carrier frequency offset compensation and multiuser detection for multiuser MIMO systems. This approach is not feasible for the huge scales of MIMO systems.

As for the opportunities in complexity, there are a few suggestions that being suggested by other researchers, such as:

(a) A suitable scheme such as the MMSE scheme is favourable for a coordinated beamforming scheme for massive MIMO as mentioned by Shikida et al. [113].

(b) Additionally, Jin et al. [163] have suggested that the volume-based algorithm is needed because its computational complexity is greatly lesser than capacity-based algorithms.

(c) Haghighatshoar et al. [112] mentioned that there is an opportunity in tracking mode such that the subspace estimate is updated upon arrival in low complexity algorithm.

(d) A minimal complexity 2D-AOA estimator based on unitary estimation of the signal parameters through rotational invariance techniques (2D-UESPRIT) algorithm is introduced in [117] due to the conventional 2D-AOA estimation method will produced high computational complexity.

7.4. Cross-Layer

In order to support uRLLC, issues in radio resource allocation are among the key challenges. This is due to the mandatory transmission and queuing delay is shorter than the coherence channel time [164]. The aftermath of the issues arise are:

(a) Baligh et al. [165] in their study stated the main challenges in cross-layer of ultra-reliable low latency massive-MIMO are due to the progressively larger numbers of access nodes in 5G systems, making it challenging to optimize the cross-layer system utility maximization problem as it is huge.

(b) Meanwhile, in mMTC, the effective connectivity as well as scalability for massive numbers of devices is not done adequately in cellular systems. Considerations of low cost and energy efficiency thus enabling wide area coverage and deep indoor penetration have been mentioned by Bockelmann et al. [166]. Other challenges or limitations and their causes in the cross-layer for low-latency massive MIMO are also tabulated and discussed in Table 7.

Nonetheless, in every challenge lies an opportunity. For instance, the opportunity that can be gained in [166] stated that a promising candidate is the flexible waveform design enabling in-band massive machine type communication channels. Another opportunity mentioned by Moreira et al. [151] in his findings stated that by attempting to build an efficient and reliable radio zone, a machine learning techniques would be focused to properly classify the numerous RSS profile of HetNets. However, the challenges of delay, as well as weak security, needs to be tackled by developing Java API in order to calculate the complexity of the systems and later compare it to the non-cryptographic and cryptographic methods.

Table 7. Limitations and its causes in cross-layer for low-latency massive MIMO.

Limitations	Causes
Reliability cannot be improved by using the ARQ mechanism	• Re-transmitting packets in will only introduce addition latency but when it stayed in fading, it barely enhanced the successful transmission. • In deep fading, unbounded transmit power is resulted due to the average latency approaching the coherence time channel. Therefore, with finite power, it is uncertain to balance the ultra-low latency and ultra-high reliability.
Block-length of channel code is finite	• Discovering the most optimal resource allocation for uRLLC is becoming more challenging due to the achievable maximal rate is convex.
Using effective bandwidth for URLLC is not recommended	• The effective bandwidth can only be used when the delay violation probability is small whereas the delay bound is huge. Thus, it is not recommended.

8. Current Progress in 4Cs in Ultra-Low Latency Massive MIMO

Following are the current progress that related to ultra-low latency massive MIMO based on the 4Cs framework:

8.1. Computing

The direct consequences of the deployment of wide spectrum of distributed end devices are the generation of a large chunk of data. Consequently, the processing demands attention. Cloud and FC are crucial in the sustenance of these devices. Herein, we summarize the various aspects of current progress in computing for formulating next generation machines and applications over computing environment to achieve low latency:

(1) Partitioning of tasks or services: this issue classified into multiple stages—(i) estimate the fog devices resources, (ii) task segregation driven by fog devices resource availability and the probable response duration to achieve task completion, (iii) estimate overheads for task segregation and relocation, (iv) estimate the overheads for outcome gathering for different sub-tasks, (v) optimal positioning of sub-tasks at designated fog device and to the cloud [167]. While task positioning and task assignment issues are deeply established in the literature for various scenarios, for instance, under dispersed scenarios [168], for parallel computing architecture [169], and lately, for mobile and CC environments [170], the situation is not the same in FC environment. The cost-advantage limitations are also crucial. The goal is to divest the jobs from a resource surplus environment (cloud) to a resource limited platform (fog) primarily for the purpose of better response time and greater privacy acceptance. Hence, the job of relocation algorithms is critical and faultless with a high degree of accuracy. The aforementioned performance criteria are a major issue in developing a fog computing system, and thorough research is needed so as to develop the necessary frameworks.

(2) Semantics in fog computing: the fog computing platform encompassing various heterogeneous sensors, actuators, edge devices and cloud servers. Fog infrastructure is mainly distributed from the IoT perspective. In these largely distributed platforms, comprehending service-oriented computing is a big challenge. Therefore, embedding semantics for deliveries is problematic. For instance, various applications may deploy equivalent actuator service and result in conflicting actions. It is plausible that the application-based actuation service can be dissimilar from the actuation service of another application. A strategy is required for managing the work-flow of dissimilar facilities of an application. An application work-flow requires proper management with semantically correct services so as to attain the end-goals. The main issue in this regard is to apply appropriate connotation to the activities in order to execute the application. Several

works have been proposed previously for administering semantics in a distributed computing platform [171]. Nevertheless, these mechanisms cannot be directly apply to fog computing, because: (i) fog is a partly distributed framework, in which clouds roles are rigidly defined, (ii) the role of the cloud in an application relies on many scenarios, hence, the formulation of a generic framework is not easy, (iii) the resource availabilities at each fog device are not static, and (iv) for a SOA based architecture, the dependency among micro services may be complex with a mixture of dependable services. These concepts need further investigation with regards to fog computing scenarios [172].

(3) Multi-domain orchestration: the fog platform is a continuum of utilities starting from the network edge to the cloud data centers, offering various components to be overseen by various entities. E-2-E service delivery entails synchronization among various divergent domains with likely heterogeneous control rules. Standardization of interfaces has become imperative so that services can be provisioned, while satisfying E-2-E performance limitations. Global knowledge of network topology across the various domains has become essential. Before now, some works in CC domain on multi-domain orchestration exists [173,174], yet, buoyed by the distributed nature of fog entities, sustaining resource assignment considering a stratum of non-homogeneous systems with divergent rules are difficult and requires peculiar attention.

(4) Interaction among fog devices: Ensuring real-time service provisioning in a distributed environment under a heterogeneous system (with different resource availabilities under various policy domains) is difficult. To enable a fog node to generate a response within a much smaller amount of time, the dependency on the other fog nodes should be taken care of. The edge devices interact among themselves for different service calling and data sending can represent the dependencies among multiple services as a dependency tree (or graph, based on the application). Parsing such dependency tree or dependency graph of services for an application under a distributed environment requires complex interaction among various fog nodes. The research challenge in this case, is to make these interactions very fast, so that the overall system generates the output within a predefined time threshold to ensure real-time service execution is guaranteed and preserved [48,175,176].

In addition, with the current progress made in computing via the augmented and virtual reality (AR/VR) users' view can be enhanced by using augmented reality. Virtual reality, on the other hand, offers more immersive experience that in practice is typical. In these innovations, network latency can be addressed by using hybrid cloud processing where it able to reduce the processing and bandwidth requirements. Moreover, content distribution can be achieved by introducing mobile edge computing (MEC) at the edge of the cellular network [177]. On the other hand, Verizon engineers are also currently working on the MEC platform software by using MEC equipment to reduce the network latency in half on a live 5G network. The engineers have relied heavily on the carrier's Intelligent Edge Network architecture where the equipment and software reduced the physical distance data needed to travel between the computer infrastructure and the wireless devices [178].

Researchers' from King's College London are also making progress in futuristic applications for 5G. It is called as "Internet of Skills" where it able to transfer physical skills across the network by using technology [179]. Being developed by NeuroDigital, haptic gloves that are equipped with motion and pressure sensors are used by surgeons [180]. The haptic gloves are made of wearable "Sarotis" technology that being fabricated from soft fabric that wraps around the skin and acts as a "second skin". This "second skin" allows the users to feel the same things the actors feels such as handshakes, catching a ball and many more [180]. During the Qualcomm Snapdragon Tech Summit, they have showcased the world first telemedicine example through a demo in collaboration with Verizon and Columbia University [181]. Virtual reality (VR) telemedicine has become one of the most revolutionary and potentially life-saving applications of 5G.

8.2. Cost of Management the Multiple Operating Environments and Applications

There is a need to run multiple operating environments and applications on a single fog device. Besides, the resource allocation as well as various services need to be coordinated for proper orchestration of application services. Herein, we present different aspects of current progress to provide multiple operating environments:

(1) Service separation and encapsulation: The services or tasks from two different applications or users may need to have a separate environment. Therefore, service separation and encapsulation are important, when both the services or tasks are run on a single fog device [182].
(2) Application fairness: To ensure application fairness, resource reservation and provisioning to the services over a single fog device need to be monitored, and the management algorithms need to take care of the fairness aspect [183].
(3) Data privacy: When services from different applications or users run on a single fog node, data privacy for individual application or user needs to be ensured [184].
(4) Fault tolerance: This is an important aspect. When a fog device fails, the services running on that device need to be migrated to a different device, while maintaining application and user transparency. Seamless migration of services is an essential requirement to ensure high availability of resources over the fog computing environment. The environment should support various types of fault tolerance, like crash faults, network faults and byzantine faults [185].

8.3. Complexity

Wireless communications are expected to be embedded with seamless connectivity functionality to an enormous number of devices or users with the combination of ultra-reliable low latency with extremely high reliability, security and availability. This is also being called as Tactile Internet as mentioned in [86] that introduces a revolution of development in society, economics and culture. However, to implement such extreme requirements in tactile internet is not a trivial task and formidable challenges exist in hardware implementation. This is due to the exponential increase in computational complexity at the signal detection. One way to mitigate the computation complexity at signal detection. The most practical solution is to introduce iterative approaches. Among the method that can be implemented are Richardson [101], Gauss-Seidel [98], Jacobi [97], Successive Overrelation (SOR) [102] as well as Symmetric Successive Overrelaxation (SSOR) [103] methods. Current progress on the detection scheme in mitigating computational complexity is by proposing rate-compatible puncturing polar (RCPP) codes in which two different puncturing methods are identified—quasi uniform puncturing and shortening [186]. Also, another detection scheme for polar codes that can reduce complexity is blind detection based on the cyclic redundancy check (CRC) [187]. The emergence of 5G also introduces the MIMO-non orthogonal multiple access (NOMA) where the low-complexity receiver used LMMSE multi-user detector that operates iteratively with single user message passing decoder [186] for the latest current progress in achieving low-complexity low-latency in 5G.

On a side note, AT&T along with Nokia is set to produce a platform based on open source software that can be aligned with Open-RAN target architecture by allowing development on it. SDN is heavily reliant upon by the carrier to power the network. This includes a network cloud platform that supports all the carrier's application and SDN services [188].

8.4. Cross-Layer

As mentioned above, cross-layer is defined as the protocol design that disrupts the architecture layers in optimizing the overall performance of the network topology. To guarantee the extreme low E-2-E of the tactile internet, promising solutions will be introduced herein. In current OFDM, it faces a challenge for opportunistic and dynamic spectrum access [185] which is due to the high out-of-band (OOB) emission [186]. In addition to that, there are various types of modulation schemes such as Filter Bank Multicarrier (FMBC) [16], Universal Filtered Multicarrier (UFMC) [187] and Bi-Orthogonal

OFDM (BFDM) [188] has been discussed for 5G, however, the most promising solution of modulation scheme for PHY layer for 5G is Generalized Frequency Division Multiplexing (GFDM) [187]. As discussed in [146], a short transmission time interval in frame structure could be one of the practical solutions in achieving 1ms E2E TTI latency. Moreover, achieving low latency in cross-layer for 5G can be performed by using forward error correction (FEC) channel coding with iterative decoding [186]. Finally, [188] provided the implementation of low-density parity check (LDPC) convolutional codes to validate the stringent requirements of the tactile Internet application. In addressing the ultra-reliable low latency strict requirement, an emerging transport technology called Flexible-Ethernet (FlexE) is currently developed by ZTE where it forwards and secures the isolated network slicing at layer 2 and layer 3 technologies that are unable to match. A new protocol is built by the Common Public Radio Interface (CPRI) consortium in order to address the fronthaul transport needs [188].

9. Conclusions

This paper presents the state-of-the-art on the Four-C framework for high capacity ultra-reliable and low latency 5G networks with regards to Computing, Cost, Complexity and Cross-Layer requirements. The requirements and potential applications related to ultra-reliable and low latency are initially presented, followed by further definition and establishment of the Four-C Framework. A classification based on previous works under each of the 'C' elements is also presented, followed by a review on the emerging technologies such as SDN, NFV, fog networking, Hypervisor in virtual machines (vCPU, vRAM, vDisk, etc.), and complex MU-MIMO and Massive-MIMO schemes. The new uRLLC network architecture framework needs to consider factors such as security and privacy, scalability, resource management and utilization, efficiency as well as complexity. Finally, challenges and opportunities to achieve ultra-low latency communication have also been discussed.

Funding: This work is partly supported by research grant FRGS/1/2018/ICT03/UKM/02/3 under Malaysian Ministry of Education and Universiti Kebangsaan Malaysia.

Conflicts of Interest: The author declare that they have no competing interest.

References

1. Boccardi, F.; Heath, R.W.; Lozano, A.; Marzetta, T.L.; Popovski, P. Five disruptive technology directions for 5G. *IEEE Commun. Mag.* **2014**, *52*, 74–80. [CrossRef]
2. Pedersen, K.I.; Frederiksen, F.; Berardinelli, G.; Mogensen, P.E. The Coverage-Latency-Capacity Dilemma for TDD Wide Area Operation and Related 5G Solutions. In Proceedings of the 2016 IEEE 83rd Vehicular Technology Conference: VTC Spring, Nanjing, China, 15–18 May 2016; pp. 1–5.
3. Popovski, P. Ultra-reliable communication in 5G wireless systems. In Proceedings of the 2014 1st International Conference on 5G for Ubiquitous Connectivity (5GU), Levi, Finland, 26–27 November 2014; pp. 146–151.
4. Johansson, N.A.; Wang, Y.-P.E.; Eriksson, E.; Hessler, M. Radio access for ultra-reliable and low-latency 5G communications. In Proceedings of the 2015 IEEE International Conference on Communication Workshop (ICCW), London, UK, 8–12 June 2015; pp. 1184–1189.
5. She, C.; Yang, C.; Quek, T.Q. Radio resource management for ultra-reliable and low-latency communications. *IEEE Commun. Mag.* **2017**, *55*, 72–78. [CrossRef]
6. Hou, Z.; She, C.; Li, Y.; Quek, T.Q.; Vucetic, B. Burstiness-Aware Bandwidth Reservation for Ultra-Reliable and Low-Latency Communications in Tactile Internet. *IEEE J. Sel. Areas Commun.* **2018**, *36*, 2401–2410. [CrossRef]
7. Thota, J.; Abdullah, N.F.; Doufexi, A.; Armour, S. V2V for vehicular safety applications. *IEEE Trans. Intell. Transp. Syst.* **2019**, 1–15. [CrossRef]
8. Shalom, N. The Tera-Scale Effect. Available online: https://natishalom.typepad.com/nati_shaloms_blog/2010/11/the-tera-scale-effect-part-i.html (accessed on 5 July 2019).
9. Wen, T.; Zhu, P. *5G: A Technology Vision*; Huawei: Beijing, China, 2013.
10. Li, Q.; Li, G.; Lee, W.; Lee, M.-I.; Mazzarese, D.; Clerckx, B.; Li, Z. MIMO techniques in WiMAX and LTE: a feature overview. *IEEE Commun. Mag.* **2010**, *48*, 86–92. [CrossRef]

11. Lahetkangas, E.; Pajukoski, K.; Vihriala, J.; Berardinelli, G.; Lauridsen, M.; Tiirola, E.; Mogensen, P. Achieving low latency and energy consumption by 5G TDD mode optimization. In Proceedings of the 2014 IEEE International Conference on Communications Workshops (ICC), Sydney, Australia, 10–14 June 2014; pp. 1–6.
12. Lee, G.; Sung, Y. A new approach to user scheduling in massive multi-user MIMO broadcast channels. *IEEE Trans. Commun.* **2018**, *66*, 1481–1495. [CrossRef]
13. Björnson, E.; Sanguinetti, L.; Hoydis, J.; Debbah, M. Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer? *IEEE Trans. Wirel. Commun.* **2015**, *14*, 3059–3075. [CrossRef]
14. Bjornson, E.; Kountouris, M.; Debbah, M. Massive MIMO and small cells: Improving energy efficiency by optimal soft-cell coordination. In Proceedings of the 2013 20th International Conference on Telecommunications (ICT), Casablanca, MA, USA, 6–8 May 2013; pp. 1–5.
15. Alsharif, M.H.; Nordin, R.; Shakir, M.M.; Ramly, A.M. Small Cells Integration with the Macro-Cell Under LTE Cellular Networks and Potential Extension for 5G. *J. Electr. Eng. Technol.* **2019**, 1–11. [CrossRef]
16. Simsek, M.; Aijaz, A.; Dohler, M.; Sachs, J.; Fettweis, G. 5G-enabled tactile internet. *IEEE J. Sel. Areas Commun.* **2016**, *34*, 460–473. [CrossRef]
17. Jungnickel, V.; Manolakis, K.; Zirwas, W.; Panzner, B.; Braun, V.; Lossow, M.; Svensson, T. The role of small cells, coordinated multipoint, and massive MIMO in 5G. *IEEE Commun. Mag.* **2014**, *52*, 44–51. [CrossRef]
18. Farhang, A.; Marchetti, N.; Figueiredo, F.; Miranda, J.P. Massive MIMO and waveform design for 5th generation wireless communication systems. In Proceedings of the 1st International Conference on 5G for Ubiquitous Connectivity, Akaslompolo, Finland, 26–28 November 2014; pp. 70–75.
19. Ali-Ahmad, H.; Cicconetti, C.; de la Oliva, A.; Dräxler, M.; Gupta, R.; Mancuso, V.; Sciancalepore, V. CROWD: An SDN approach for DenseNet. In Proceedings of the 2013 Second European Workshop on Software Defined Networks, Berlin, Germany, 10–11 October 2013.
20. Kreutz, D.; Ramos, F.M.; Verissimo, P.E.; Rothenberg, C.E.; Azodolmolky, S.; Uhlig, S. Software-defined networking: A comprehensive survey. *Proc. IEEE* **2015**, *103*, 14–76. [CrossRef]
21. McKeown, N.; Anderson, T.; Balakrishnan, H.; Parulkar, G.; Peterson, L.; Rexford, J.; Turner, J. OpenFlow: enabling innovation in campus networks. *ACM SIGCOMM Comput. Commun. Rev.* **2008**, *38*, 69–74. [CrossRef]
22. OpenDaylight. Linux Foundation Collaborative Project. Available online: http://www.opendaylight.Org (accessed on 5 July 2019).
23. Truong, N.B.; Lee, G.M.; Ghamri-Doudane, Y. Software defined networking-based vehicular adhoc network with fog computing. In Proceedings of the 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM), Ottawa, ON, Canada, 11–15 May 2015; pp. 1202–1207.
24. Richardson, L.; Ruby, S. *RESTful Web Services*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2008.
25. Parvez, I.; Rahmati, A.; Guvenc, I.; Sarwat, A.I.; Dai, H. A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions. *arXiv* **2017**, arXiv:1708.02562. [CrossRef]
26. Bonfim, M.S.; Dias, K.L.; Fernandes, S.F. Integrated NFV/SDN Architectures: A Systematic Literature Review. *arXiv* **2018**, arXiv:1801.01516. [CrossRef]
27. Agiwal, M.; Roy, A.; Saxena, N. Next generation 5G wireless networks: A comprehensive survey. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 1617–1655. [CrossRef]
28. Schulz, P.; Matthe, M.; Klessig, H.; Simsek, M.; Fettweis, G.; Ansari, J.; Puschmann, A. Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture. *IEEE Commun. Mag.* **2017**, *55*, 70–78. [CrossRef]
29. Yilmaz, O.N.; Wang, Y.-P.E.; Johansson, N.A.; Brahmi, N.; Ashraf, S.A.; Sachs, J. Analysis of ultra-reliable and low-latency 5G communication for a factory automation use case. In Proceedings of the 2015 IEEE International Conference on Communication Workshop (ICCW), London, UK, 8–12 June 2015; pp. 1190–1195.
30. Kalør, A.E.; Guillaume, R.; Nielsen, J.J.; Mueller, A.; Popovski, P. Network slicing for ultra-reliable low latency communication in Industry 4.0 scenarios. *arXiv* **2017**, arXiv:1708.09132.
31. MIIT. Explaining Made in China 2025. Available online: http://www.miit.gov.cn/n11293472/n11293877/n16553775/n16553822/16633916.html (accessed on 1 May 2015).
32. Fettweis, G.P. The tactile internet: Applications and challenges. *IEEE Veh. Technol. Mag.* **2014**, *9*, 64–70. [CrossRef]
33. ITU-Report, The Tactile Internet, ITU-T Technology Watch. August 2014. Available online: https://www.itu.int/dms_pub/itu-t/opb/gen/T-GEN-TWATCH-2014-1-PDF-E.pdf (accessed on 5 July 2019).

34. Aijaz, A.; Dohler, M.; Aghvami, A.H.; Friderikos, V.; Frodigh, M. Realizing the tactile internet: Haptic communications over next generation 5G cellular networks. *IEEE Wirel. Commun.* **2017**, *24*, 82–89. [CrossRef]
35. Szabo, D.; Gulyas, A.; Fitzek, F.H.; Lucani, D.E. Towards the tactile internet: Decreasing communication latency with network coding and software defined networking. In Proceedings of the 21th European Wireless Conference; Proceedings of European Wireless 2015, Budapest, Hungary, 20–22 May 2015; pp. 1–6.
36. Arata, J.; Takahashi, H.; Pitakwatchara, P.; Warisawa, S.I.; Tanoue, K.; Konishi, K.; Fujino, Y. A remote surgery experiment between Japan and Thailand over Internet using a low latency CODEC system. In Proceedings of the 2007 IEEE International Conference on Robotics and Automation, Roma, Italy, 10–14 April 2007; pp. 953–959.
37. Knopp, R. Latency Requirements in M2M Application Scenarios. In Proceedings of the 2013 European Telecommunications Standards Institute (ETSI) M2M Workshop, Sophia Antipolos, France, 27 June 2013.
38. Castañeda, E.; Silva, A.; Gameiro, A.; Kountouris, M. An overview on resource allocation techniques for multi-user MIMO systems. *IEEE Commun. Surv. Tutor.* **2017**, *19*, 239–284. [CrossRef]
39. Li, Y.; Chen, M. Software-defined network function virtualization: A survey. *IEEE Access* **2015**, *3*, 2542–2553.
40. Hu, F.; Hao, Q.; Bao, K. A survey on software-defined network and openflow: From concept to implementation. *IEEE Commun. Surv. Tutor.* **2014**, *16*, 2181–2206. [CrossRef]
41. Xia, W.; Wen, Y.; Foh, C.H.; Niyato, D.; Xie, H. A survey on software-defined networking. *IEEE Commun. Surv. Tutor.* **2015**, *17*, 27–51. [CrossRef]
42. Mijumbi, R.; Serrat, J.; Gorricho, J.-L.; Bouten, N.; de Turck, F.; Boutaba, R. Network function virtualization: State-of-the-art and research challenges. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 236–262. [CrossRef]
43. Foukas, X.; Patounas, G.; Elmokashfi, A.; Marina, M.K. Network slicing in 5g: Survey and challenges. *IEEE Commun. Mag.* **2017**, *55*, 94–100. [CrossRef]
44. Afolabi, I.; Taleb, T.; Samdanis, K.; Ksentini, A.; Flinck, H. Network slicing and softwarization: A survey on principles, enabling technologies and solutions. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 2429–2453. [CrossRef]
45. Al-Fuqaha, A.; Guizani, M.; Mohammadi, M.; Aledhari, M.; Ayyash, M. Internet of things: A survey on enabling technologies, protocols, and applications. *IEEE Commun. Surv. Tutor.* **2015**, *17*, 2347–2376. [CrossRef]
46. Wang, T.; Su, Z.; Xia, Y.; Hamdi, M. Rethinking the data center networking: Architecture, network protocols, and resource sharing. *IEEE Access* **2014**, *2*, 1481–1496. [CrossRef]
47. Wang, B.; Qi, Z.; Ma, R.; Guan, H.; Vasilakos, A.V. A survey on data center networking for cloud computing. *Comput. Netw.* **2015**, *91*, 528–547. [CrossRef]
48. Mahmud, R.; Kotagiri, R.; Buyya, R. Fog computing: A taxonomy, survey and future directions. In *Internet of Everything*; Springer: Singapore, 2018; pp. 103–130.
49. Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*; MIT Press: Cambridge, MA, USA, 2009.
50. Bonomi, F.; Milito, R.; Natarajan, P.; Zhu, J. Fog computing: A platform for internet of things and analytics. In *Big Data and Internet of Things: A Roadmap for Smart Environments*; Springer: Basel, Switzerland, 2014; pp. 169–186.
51. Yangui, S.; Ravindran, P.; Bibani, O.; Glitho, R.H.; Hadj-Alouane, N.B.; Morrow, M.J.; Polakos, P.A. A platform as-a-service for hybrid cloud/fog environments. In Proceedings of the 2016 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN), Beijing, China, 22–24 April 2016; pp. 1–7.
52. Agarwal, S.; Yadav, S.; Yadav, A.K. An efficient architecture and algorithm for resource provisioning in fog computing. *Int. J. Inf. Eng. Electron. Bus.* **2016**, *8*, 48. [CrossRef]
53. Kapsalis, A.; Kasnesis, P.; Venieris, I.S.; Kaklamani, D.I.; Patrikakis, C.Z. A cooperative fog approach for effective workload balancing. *IEEE Cloud Comput.* **2017**, *4*, 36–45. [CrossRef]
54. Krishnan, Y.N.; Bhagwat, C.N.; Utpat, A.P. Fog computing—Network based cloud computing. In Proceedings of the 2015 2nd International Conference on Electronics and Communication Systems (ICECS), Coimbatore, India, 26–27 February 2015; pp. 250–251.
55. Bonomi, F.; Milito, R.; Zhu, J.; Addepalli, S. Fog computing and its role in the internet of things. In Proceedings of the 2012 First Edition of the MCC Workshop on Mobile Cloud Computing, Helsinki, Finland, 13–17 August 2012; pp. 13–16.

56. Intharawijitr, K.; Iida, K.; Koga, H. Analysis of fog model considering computing and communication latency in 5G cellular networks. In Proceedings of the 2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), Sydney, Australia, 14–18 March 2016; pp. 1–4.
57. Xiao, J.; Wen, H.; Wu, B.; Jiang, X.; Ho, P.-H.; Zhang, L. Joint design on DCN placement and survivable cloud service provision over all-optical mesh networks. *IEEE Trans. Commun.* **2014**, *62*, 235–245. [CrossRef]
58. Dastjerdi, A.V.; Gupta, H.; Calheiros, R.N.; Ghosh, S.K.; Buyya, R. Fog computing: Principles, architectures, and applications. In *Internet of Things*; Elsevier: Cambridge, MA, USA, 2016; pp. 61–75.
59. Sarkar, S.; Chatterjee, S.; Misra, S. Assessment of the Suitability of Fog Computing in the Context of Internet of Things. *IEEE Trans. Cloud Comput.* **2015**, *6*, 46–59. [CrossRef]
60. Do, C.T.; Tran, N.H.; Pham, C.; Alam, M.G.R.; Son, J.H.; Hong, C.S. A proximal algorithm for joint resource allocation and minimizing carbon footprint in geo-distributed fog computing. In Proceedings of the 2015 International Conference on Information Networking (ICOIN), Kota Kinabalu, Malaysia, 13–15 January 2015; pp. 324–329.
61. Aazam, M.; Huh, E.-N. Fog computing micro datacenter based dynamic resource estimation and pricing model for IoT. In Proceedings of the 2015 IEEE 29th International Conference on Advanced Information Networking and Applications (AINA), Gwangju, Korea, 25–27 March 2015; pp. 687–694.
62. Aazam, M.; Huh, E.-N. Dynamic resource provisioning through Fog micro datacenter. In Proceedings of the 2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), Louis, MO, USA, 23–27 March 2015; pp. 105–110.
63. Yannuzzi, M.; Milito, R.; Serral-Gracià, R.; Montero, D.; Nemirovsky, M. Key ingredients in an IoT recipe: Fog Computing, Cloud computing, and more Fog Computing. In Proceedings of the 2014 IEEE 19th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), Athens, Greece, 1–3 December 2014; pp. 325–329.
64. Madsen, H.; Burtschy, B.; Albeanu, G.; Popentiu-Vladicescu, F. Reliability in the utility computing era: Towards reliable fog computing. In Proceedings of the 2013 20th International Conference on Systems, Signals and Image Processing (IWSSIP), Bucharest, Romania, 7–9 July 2013; pp. 43–46.
65. Chen, W.; Cao, J.; Wan, Y. QoS-aware virtual machine scheduling for video streaming services in multi-cloud. *Tsinghua Sci. Technol.* **2013**, *18*, 308–317. [CrossRef]
66. Hong, K.; Lillethun, D.; Ramachandran, U.; Ottenwälder, B.; Koldehofe, B. Mobile fog: A programming model for large-scale applications on the internet of things. In Proceedings of the Second ACM SIGCOMM Workshop on Mobile Cloud Computing, Hong Kong, China, 12–16 August 2013; pp. 15–20.
67. Zeng, D.; Gu, L.; Guo, S.; Cheng, Z.; Yu, S. Joint optimization of task scheduling and image placement in fog computing supported software-defined embedded system. *IEEE Trans. Comput.* **2016**, *65*, 3702–3712. [CrossRef]
68. Oueis, J.; Strinati, E.C.; Sardellitti, S.; Barbarossa, S. Small cell clustering for efficient distributed fog computing: A multi-user case. In Proceedings of the 2015 IEEE 82nd Vehicular Technology Conference (VTC Fall), Boston, MA, USA, 6–9 September 2015; pp. 1–5.
69. Luan, T.H.; Gao, L.; Li, Z.; Xiang, Y.; Wei, G.; Sun, L. Fog computing: Focusing on mobile users at the edge. *arXiv* **2015**, arXiv:1502.01815.
70. Chaudhary, R.; Kumar, N.; Zeadally, S. Network Service Chaining in Fog and Cloud Computing for the 5G Environment: Data Management and Security Challenges. *IEEE Commun. Mag.* **2017**, *55*, 114–122. [CrossRef]
71. Riddle, A.R.; Chung, S.M. A survey on the security of hypervisors in cloud computing. In Proceedings of the 2015 IEEE 35th International Conference on Distributed Computing Systems Workshops (ICDCSW), Columbus, OH, USA, 29 June–2 July 2015; pp. 100–104.
72. Ahmad, R.W.; Gani, A.; Hamid, S.H.A.; Shiraz, M.; Yousafzai, A.; Xia, F. A survey on virtual machine migration and server consolidation frameworks for cloud data centers. *J. Netw. Comput. Appl.* **2015**, *52*, 11–25. [CrossRef]
73. Kim, C.; Park, K.H. Credit-based runtime placement of virtual machines on a single NUMA system for QoS of data access performance. *IEEE Trans. Comput.* **2014**, *64*, 1633–1646.
74. Rodríguez-Haro, F.; Freitag, F.; Navarro, L.; Hernánchez-sánchez, E.; Farías-Mendoza, N.; Guerrero-Ibáñez, J.A.; González-Potes, A. A summary of virtualization techniques. *Procedia Technol.* **2012**, *3*, 267–272. [CrossRef]

75. Nurmi, D.; Wolski, R.; Grzegorczyk, C.; Obertelli, G.; Soman, S.; Youseff, L.; Zagorodnov, D. The eucalyptus open-source cloud-computing system. In Proceedings of the 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, 2009 CCGRID'09, Shanghai, China, 18–21 May 2009; pp. 124–131.
76. Armbrust, M.; Fox, A.; Griffith, R.; Joseph, A.D.; Katz, R.; Konwinski, A.; Gunho, L.E.E.; PAttERSon, D.; RABKin, A. A view of cloud computing. *Commun. ACM* **2010**, *53*, 50–58. [CrossRef]
77. Berl, A.; Gelenbe, E.; di Girolamo, M.; Giuliani, G.; de Meer, H.; Dang, M.Q.; Pentikousis, K. Energy-efficient cloud computing. *Comput. J.* **2010**, *53*, 1045–1051. [CrossRef]
78. Nishio, T.; Shinkuma, R.; Takahashi, T.; Mandayam, N.B. Service-oriented heterogeneous resource sharing for optimizing service latency in mobile cloud. In Proceedings of the First International Workshop on Mobile Cloud Computing & Networking, Bangalore, India, 29 July 2013; pp. 19–26.
79. Armbrust, M.; Fox, A.; Griffith, R.; Joseph, A.D.; Katz, R.H.; Konwinski, A.; Lee, G.; Stoica, I. *Above the Clouds: A Berkeley View of Cloud Computing*; Technical Report UCB/EECS-2009-28; EECS Department, University of California: Berkeley, CA, USA, 2009.
80. Rimal, B.P.; Choi, E.; Lumb, I. A taxonomy and survey of cloud computing systems. In Proceedings of the 2009 Fifth International Joint Conference on INC, IMS and IDC, NCM'09, Seoul, Korea, 25–27 August 2009; pp. 44–51.
81. Dunham, G.M. Locally Connected Cloud Storage Device. US Patent App. 12/975,678, 23 December 2009.
82. Wu, J.; Ping, L.; Ge, X.; Wang, Y.; Fu, J. Cloud storage as the infrastructure of cloud computing. In Proceedings of the 2010 International Conference on Intelligent Computing and Cognitive Informatics (ICICCI), Kuala Lumpur, Malaysia, 22–23 June 2010; pp. 380–383.
83. Lagar-Cavilla, H.A.; Whitney, J.A.; Scannell, A.M.; Patchin, P.; Rumble, S.M.; de Lara, E.; Satyanarayanan, M. SnowFlock: Rapid virtual machine cloning for cloud computing. In Proceedings of the 4th ACM European conference on Computer systems, Scotland, UK, 1–4 April 2009; pp. 1–12.
84. Edfors, O.; Tufvesson, F. Massive MIMO for next generation wireless systems. *IEEE Commun. Mag.* **2014**, *52*, 186–195.
85. Chen, Y.; Boussakta, S.; Tsimenidis, C.; Chambers, J.; Jin, S. Low complexity hybrid precoding in finite dimensional channel for massive MIMO systems. In Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO), Kos Island, Greece, 28 August–2 September 2017; pp. 883–887.
86. Tsinos, C.G.; Maleki, S.; Chatzinotas, S.; Ottersten, B. On the energy-efficiency of hybrid analog–digital transceivers for single-and multi-carrier large antenna array systems. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 1980–1995. [CrossRef]
87. Hussein, J.; Ikki, S.S.; Boussakta, S.; Tsimenidis, C.C. Performance Analysis of Opportunistic Scheduling in Dual-Hop Multiuser Underlay Cognitive Network in the Presence of Cochannel Interference. *IEEE Trans. Veh. Technol.* **2016**, *65*, 8163–8176. [CrossRef]
88. Zhang, X.; Molisch, A.F.; Kung, S.-Y. Variable-phase-shift-based RF-baseband codesign for MIMO antenna selection. *IEEE Trans. Signal Process.* **2005**, *53*, 4091–4103. [CrossRef]
89. Nsenga, J.; Bourdoux, A.; Horlin, F. Mixed analog/digital beamforming for 60 GHz MIMO frequency selective channels. In Proceedings of the 2010 IEEE International Conference on Communications (ICC), Cape Town, South Africa, 23–27 May 2010; pp. 1–6.
90. Rusek, F.; Persson, D.; Lau, B.K.; Larsson, E.G.; Marzetta, T.L.; Edfors, O.; Tufvesson, F. Scaling up MIMO: Opportunities and challenges with very large arrays. *IEEE Signal Process. Mag.* **2013**, *30*, 40–60. [CrossRef]
91. El Ayach, O.; Heath, R.W.; Abu-Surra, S.; Rajagopal, S.; Pi, Z. Low complexity precoding for large millimeter wave MIMO systems. In Proceedings of the 2012 IEEE International Conference on Communications (ICC), Ottawa, ON, Canada, 10–15 June 2012; pp. 3724–3729.
92. El Ayach, O.; Rajagopal, S.; Abu-Surra, S.; Pi, Z.; Heath, R.W. Spatially sparse precoding in millimeter wave MIMO systems. *IEEE Trans. Wirel. Commun.* **2014**, *13*, 1499–1513. [CrossRef]
93. Méndez-Rial, R.; Rusu, C.; González-Prelcic, N.; Heath, R.W. Dictionary-free hybrid precoders and combiners for mmWave MIMO systems. In Proceedings of the 2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Stockholm, Sweden, 28 June–1 July 2015; pp. 151–155.
94. Valduga, S.T.; Deneire, L.; de Almeida, A.L.; Maciel, T.F.; Aparicio-Pardo, R. Low complexity beam selection for sparse massive MIMO systems. In Proceedings of the 2017 International Symposium on Wireless Communication Systems (ISWCS), Bologna, Italy, 28–31 August 2017; pp. 414–419.

95. Yang, P.; Xiao, Y.; Guan, Y.L.; Hari, K.; Chockalingam, A.; Sugiura, S.; Haas, H.; Di Renzo, M.; Masouros, C.; Liu, Z.; et al. Single-carrier SM-MIMO: A promising design for broadband large-scale antenna systems. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 1687–1716. [CrossRef]
96. Jiang, F.; Li, C.; Gong, Z. A low complexity soft-output data detection scheme based on Jacobi method for massive MIMO uplink transmission. In Proceedings of the 2017 IEEE International Conference on Communications (ICC), Paris, France, 21–25 May 2017; pp. 1–5.
97. Qin, X.; Yan, Z.; He, G. A near-optimal detection scheme based on joint steepest descent and Jacobi method for uplink massive MIMO systems. *IEEE Commun. Lett.* **2016**, *20*, 276–279. [CrossRef]
98. Dai, L.; Gao, X.; Su, X.; Han, S.; Chih-Lin, I.; Wang, Z. Low-complexity soft-output signal detection based on Gauss–Seidel method for uplink multiuser large-scale MIMO systems. *IEEE Trans. Veh. Technol.* **2015**, *64*, 4839–4845. [CrossRef]
99. Jiang, F.; Li, C.; Gong, Z. Low complexity and fast processing algorithms for V2I massive MIMO Uplink Detection. *IEEE Trans. Veh. Technol.* **2018**, *67*, 5054–5068. [CrossRef]
100. Vikalo, H.; Hassibi, B. Maximum-likelihood sequence detection of multiple antenna systems over dispersive channels via sphere decoding. *EURASIP J. Appl. Signal Process.* **2002**, *2002*, 525–531. [CrossRef]
101. Gao, X.; Dai, L.; Yuen, C.; Zhang, Y. Low-complexity MMSE signal detection based on Richardson method for large-scale MIMO systems. In Proceedings of the 2014 IEEE 80th Vehicular Technology Conference (VTC Fall), Vancouver, BC, Canada, 14–17 September 2014; pp. 1–5.
102. Gao, X.; Dai, L.; Hu, Y.; Wang, Z.; Wang, Z. Matrix inversion-less signal detection using SOR method for uplink large-scale MIMO systems. In Proceedings of the 2014 IEEE Global Communications Conference (GLOBECOM), Toronto, ON, Canada, 27 April–2 May 2014; pp. 3291–3295.
103. Xie, T.; Dai, L.; Gao, X.; Dai, X.; Zhao, Y. Low-complexity SSOR-based precoding for massive MIMO systems. *IEEE Commun. Lett.* **2016**, *20*, 744–747. [CrossRef]
104. Wu, Z.; Xue, Y.; You, X.; Zhang, C. Hardware efficient detection for massive MIMO uplink with parallel Gauss-Seidel method. In Proceedings of the 2017 22nd International Conference on Digital Signal Processing (DSP), London, UK, 23–25 August 2017; pp. 1–5.
105. Liu, F.; Xu, Y.; Bai, X. Low-complexity receiver for uplink massive MIMO systems. In Proceedings of the 2017 3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu, China, 13–16 December 2017; pp. 952–956.
106. Kim, S.; Lee, N.; Hong, S.-N. Uplink Massive MIMO Systems with One-Bit ADCs: A Low-Complexity Weighted Minimum Distance Decoding. In Proceedings of the GLOBECOM 2017–2017 IEEE Global Communications Conference, Singapore, 4–8 December 2017; pp. 1–6.
107. Sabeti, P.; Farhang, A.; Marchetti, N.; Doyle, L. Low-Complexity CFO Compensation for OFDM-Based Massive MIMO Systems. In Proceedings of the 2017 IEEE Globecom Workshops (GC Wkshps), Singapore, 4–8 December 2017; pp. 1–6.
108. Wu, H.; Liu, D.; Wu, W.; Na, C.; Liu, M. A low complexity two-stage user scheduling scheme for MmWave massive MIMO hybrid beamforming systems. In Proceedings of the 2017 3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu, China, 13–16 December 2017; pp. 945–951.
109. Ahmed, Y.N.; Fahmy, Y. On the complexity reduction of codebook search in FDD massive MIMO using hierarchical search. In Proceedings of the 2018 International Conference on Innovative Trends in Computer Engineering (ITCE), Aswan, Egypt, 19–21 February 2018; pp. 175–179.
110. Minango, J.; de Almeida, C. Low-complexity MMSE detector based on the first-order Neumann series expansion for massive MIMO systems. In Proceedings of the 2017 IEEE 9th Latin-American Conference on Communications (LATINCOM), Guatemala City, Guatemala, 8–10 November 2017; pp. 1–5.
111. Minango, J.; de Almeida, C.; Altamirano, C.D. Low-complexity MMSE detector for massive MIMO systems based on Damped Jacobi method. In Proceedings of the 2017 IEEE 28th Annual International Symposium on, Personal, Indoor, and Mobile Radio Communications (PIMRC), Montreal, QC, Canada, 8–13 October 2017; pp. 1–5.
112. Haghighatshoar, S.; Caire, G. Low-complexity massive MIMO subspace tracking from low-dimensional projections. In Proceedings of the 2017 IEEE International Conference on Communications (ICC), Paris, France, 21–25 March 2017; pp. 1–7.

113. Shikida, J.; Ishii, N. Performance Analysis of Low Complexity Coordinated Beamforming for Massive MIMO System. In Proceedings of the 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall), Toronto, ON, Canada, 24–27 September 2017; pp. 1–5.
114. Zhang, W.; Bao, X.; Dai, J. Low-complexity detection based on Landweber Method in the uplink of massive MIMO systems. In Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO), Greek Island, Greece, 28 August–2 September 2017; pp. 873–877.
115. Tang, H.; Zhang, W.; Hardjawana, W.; Vucetic, B. Improving latency and reliability in 5G Internet-of-Things networks. In Proceedings of the 2016 IEEE International Conference on Smart Grid Communications (SmartGridComm), Sydney, Australia, 6–9 November 2016; pp. 509–513.
116. Lei, L.; Yan, C.; Wenting, G.; Huilian, Y.; Yiqun, W.; Shuangshuang, X. Prototype for 5G new air interface technology SCMA and performance evaluation. *China Commun.* **2015**, *12*, 38–48.
117. Ali, E.; Ismail, M.; Nordin, R.; Abdullah, N.F. Beamforming with 2D-AOA estimation for pilot contamination reduction in massive MIMO. *Telecommun. Syst.* **2018**, *71*, 541–552. [CrossRef]
118. Srivastava, V.; Motani, M. Cross-layer design: A survey and the road ahead. *IEEE Commun. Mag.* **2005**, *43*, 112–119. [CrossRef]
119. Luo, C.; Yu, F.R.; Ji, H.; Leung, V.C. Cross-layer design for TCP performance improvement in cognitive radio networks. *IEEE Trans. Veh. Technol.* **2010**, *59*, 2485–2495.
120. Hanke, M. Accelerated Landweber iterations for the solution of ill-posed equations. *Numer. Math.* **1991**, *60*, 341–373. [CrossRef]
121. Araújo, D.C.; Maksymyuk, T.; de Almeida, A.L.; Maciel, T.; Mota, J.C.; Jo, M. Massive MIMO: survey and future research topics. *IET Commun.* **2016**, *10*, 1938–1946. [CrossRef]
122. Feng, W.; Wang, Y.; Ge, N.; Lu, J.; Zhang, J. Virtual MIMO in multi-cell distributed antenna systems: coordinated transmissions with large-scale CSIT. *IEEE J. Sel. Areas Commun.* **2013**, *31*, 2067–2081. [CrossRef]
123. 3GPP Study on New Radio (NR) Access Technology: Physical Layer Aspects. Available online: https://www.etsi.org/deliver/etsi_tr/138900_138999/138912/14.00.00_60/tr_138912v140000p.pdf (accessed on 3 September 2019).
124. Destounis, A.; Maso, M. Adaptive clustering and CSI acquisition for FDD massive MIMO systems with two-level precoding. In Proceedings of the 2016 IEEE Wireless Communications and Networking Conference (WCNC), Doha, Qatar, 3–6 April 2016; pp. 1–6.
125. Alkhaled, M.; Alsusa, E.; Pramudito, W. Adaptive user grouping algorithm for the downlink massive MIMO systems. In Proceedings of the 2016 IEEE Wireless Communications and Networking Conference (WCNC), Doha, Qatar, 3–6 April 2016; pp. 1–6.
126. Lee, G.; Sung, Y. Asymptotically optimal simple user scheduling for massive MIMO downlink with two-stage beamforming. In Proceedings of the 2014 IEEE 15th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Toronto, ON, Canada, 22–25 June 2014; pp. 60–64.
127. Benmimoune, M.; Driouch, E.; Ajib, W.; Massicotte, D. Joint transmit antenna selection and user scheduling for massive MIMO systems. In Proceedings of the 2015 IEEE Wireless Communications and Networking Conference (WCNC), New Orleans, LA, USA, 9–12 March 2015; pp. 381–386.
128. Liu, J.; She, X.; Chen, L. A low complexity capacity-greedy user selection scheme for zero-forcing beamforming. In Proceedings of the VTC Spring 2009 IEEE 69th Vehicular Technology Conference, Barcelona, Spain, 26–29 April 2009; pp. 1–5.
129. Bogale, T.E.; Le, L.B.; Haghighat, A. User scheduling for massive MIMO OFDMA systems with hybrid analog-digital beamforming. In Proceedings of the 2015 IEEE International Conference on Communications (ICC), London, UK, 8–12 June 2015; pp. 1757–1762.
130. Han, B.; Zhao, S.; Yang, B.; Zhang, H.; Chen, P.; Yang, F. Historical PMI based multi-user scheduling for FDD massive MIMO systems. In Proceedings of the VTC Spring 2016 IEEE 83rd Vehicular Technology Conference, Nanjing, China, 15–18 May 2016; pp. 1–5.
131. Huang, S.; Yin, H.; Wu, J.; Leung, V.C. User selection for multiuser MIMO downlink with zero-forcing beamforming. *IEEE Trans. Veh. Technol.* **2013**, *62*, 3084–3097. [CrossRef]
132. Shikida, J.; Ishii, N.; Kakura, Y. Performance analysis of low complexity multi-user MIMO scheduling schemes for massive MIMO system. In Proceedings of the 2016 22nd Asia-Pacific Conference on Communications (APCC), Yogyakarta, Indonesia, 25–27 August 2016; pp. 180–184.

133. Rosen, J.B. Existence and uniqueness of equilibrium points for concave n-person games. *Econom. J. Econom. Soc.* **1965**, *33*, 520–534. [CrossRef]
134. Poor, H.V. *An Introduction to Signal Detection and Estimation*; Springer Science & Business Media: Berlin, Germany, 1994.
135. Singh, N.; Vives, X. Price and quantity competition in a differentiated duopoly. *RAND J. Econ.* **1984**, *15*, 546–554. [CrossRef]
136. Niu, K.; Chen, K.; Lin, J.; Zhang, Q. TPolar Codes: Primary Concepts and Practical Decoding Algorithms. *IEEE Commun. Mag.* **2014**, *52*, 192–203. [CrossRef]
137. Zhong, W.; Xu, Y.; Wang, J.; Li, D.; Tianfield, H. Adaptive mechanism design and game theoretic analysis of auction-driven dynamic spectrum access in cognitive radio networks. *EURASIP J. Wirel. Commun. Netw.* **2014**, *2014*, 1–14. [CrossRef]
138. Federal Communications Commission. *Unlicensed Operations in the TV Broadcast Bands, Second Memorandum Opinion and Order*; FCC 10-174; Federal Communications Commission: Washington, DC, USA, 2010.
139. S. 802.22. *Information Technology—Telecommunications and Information Exchange between Systems—Wireless Regional Area Networks (WRAN)—Sepcific Requirements—Part 22: Cognitive Wireless RAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Policies and Procedures for Operation in the TV Bands*; IEEE: Piscataway, NJ, USA, 2011.
140. S. P802.11af™/D1.02. *Draft Standard for Information Technology—Telecommunications and Information Exchange between Systems—Local and Metropolitan Area Networks—Specific Requirements—Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: TV White Spaces Operation U.S.*; IEEE: Piscataway, NJ, USA, 2011.
141. Niyato, D.; Hossain, E. Competitive pricing in heterogeneous wireless access networks: Issues and approaches. *Netw. IEEE* **2008**, *22*, 4–11. [CrossRef]
142. Shayea, I.; Ismail, M.; Nordin, R.; Ergen, M.; Ahmad, N.; Abdullah, N.F. New Weight Function for Adapting Handover Margin Level over Contiguous Carrier Aggregation Deployment Scenarios in LTE-Advanced System. Wireless Personal Communications.
143. Song, Z.; Shangguan, L.; Jamieson, K. Wi-fi goes to town: Rapid picocell switching for wireless transit networks. In Proceedings of the Conference of the ACM Special Interest Group on Data Communication, Los Angeles, CA, USA, 21–25 August 2017; pp. 322–334.
144. Au, K.; Zhang, L.; Nikopour, H.; Yi, E.; Bayesteh, A.; Vilaipornsawai, U.; Ma, J.; Zhu, P. Uplink Contention Based SCMA for 5G Radio Access. *Globecom 5G workshop* **2014**, *22*, 4–11.
145. Kela, P.; Costa, M.; Salmi, J.; Leppanen, K.; Turkka, J.; Hiltunen, T.; Hronec, M. A Novel Radio Frame Structure for 5G Dense Outdoor Radio Access Networks. In *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*; IEEE: Glasgow, UK, 2015; pp. 1–6.
146. Pedersen, K.I.; Berardinelli, G.; Frederiksen, F.; Mogensen, P.; Szufarska, A. A flexible 5G frame structure design for frequency-division duplex cases. *IEEE Commun. Mag.* **2016**, *54*, 53–59. [CrossRef]
147. Wirth, T.; Mehlhose, M.; Pilz, J.; Holfeld, B.; Wieruch, D. 5G new radio and ultra low latency applications: A PHY implementation perspective. In Proceedings of the 2016 50th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 6–9 November 2016; pp. 1409–1413.
148. Liu, D.; Zuo, C.; Wu, Z. Benefit and cost of cross sliding window scheduling for low latency 5G Turbo decoding. In Proceedings of the 2015 IEEE/CIC International Conference on Communications in China (ICCC), Shenzhen, China, 2–4 November 2015; pp. 1–4.
149. Jang, I.; Jo, G. Study on the latency efficient IFFT design method for low latency communication systems, , Phuket, 2016, pp. –4. In Proceedings of the 2016 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), Phuket, Thailand, 24–27 October 2016; pp. 1–4.
150. Moreira, C.M.; Kaddoum, G.; Bou-Harb, E. Cross-Layer Authentication Protocol Design for Ultra-Dense 5G HetNets. In Proceedings of the IEEE ICC 2018 Communication and Information Systems Security Symposium, Kansas City, MO, USA, 20–24 May 2018.
151. She, C.; Yang, C.; Quek, T.Q. Cross-Layer Transmission Design for Tactile Internet. In Proceedings of the 2016 IEEE Global Communications Conference, Washington, DC, USA, 4–8 December 2016.
152. Mathur, S.; Saha, D.; Raychaudhuri, D. Cross-layer MAC/PHY protocol to support IoT traffic in 5G. In Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking, New York City, NY, USA, 3–7 October 2016.

153. Wang, W.; Liu, Y.; Luo, Z.; Jiang, T.; Zhang, Q.; Nallanathan, A. Toward Cross-Layer Design for Non-Orthogonal Multiple Access: A Quality-of-Experience Perspective. *IEEE Wirel. Commun.* **2018**, *25*, 118–124. [CrossRef]
154. Miyim, A.M.; Ismail, M.; Nordin, R.; Ismail, M.T. Technique for Cross-layer Vertical Handover Prediction in 4G Wireless Networks. *Procedia Technol.* **2013**, *11*, 114–121. [CrossRef]
155. Bertsekas, D.P. Nonlinear programming. *J. Oper. Res. Soc.* **1997**, *48*, 334. [CrossRef]
156. Qian, L.; Li, X.; Wei, S. Anomaly spectrum usage detection in multihop cognitive radio networks: A cross-layer approach. *J. Commun.* **2013**, *8*, 259–266. [CrossRef]
157. Tosh, D.K.; Sengupta, S. Self-Coexistence in Cognitive Radio Networks using Multi-Stage Perception Learning. In Proceedings of the 2013 IEEE 78th Vehicular Technology Conference (VTC Fall), Las Vegas, NV, USA, 2–5 September 2013; pp. 1–5.
158. David, H.A.; Nagaraja, H.N. *Order Statistics*; Wiley Online Library: London, UK, 1970.
159. Ramiro, J.; Hamied, K. *Self-Organizing Networks (SON): Self-Planning, Self-Optimization and Self-Healing for GSM, UMTS and LTE*; John Wiley & Sons: West Sussex, UK, 2011.
160. Kalil, I.M. *Cognitive Radio The IEEE 802.22 standard*; IEEE: Piscataway, NJ, USA, 2011.
161. Zukang, S.; Runhua, C.; Andrews, J.G.; Heath, R.W.; Evans, B.L. Low complexity user selection algorithms for multiuser MIMO systems with block diagonalization. *IEEE Trans. Signal Process.* **2006**, *54*, 3658–3663. [CrossRef]
162. Jin, L.; Hu, Z.; Gu, X. Low-complexity scheduling strategy for wireless multiuser multiple-input multiple-output downlink system. *IET Commun.* **2011**, *5*, 990–995. [CrossRef]
163. Abas, A.R. Adaptive competitive learning neural networks. *Egypt. Inform. J.* **2013**, *14*, 183–194. [CrossRef]
164. Baligh, H.; Hong, M.; Liao, W.-C.; Luo, Z.-Q.; Razaviyayn, M.; Sanjabi, M.; Sun, R. Cross-Layer Provision of Future Cellular Networks: A WMMSE-based approach. *IEEE Signal Process. Mag.* **2014**, *31*, 56–68. [CrossRef]
165. Bockelmann, C.; Pratas, N.; Nikopour, H.; Au, K.; Svensson, T.; Stefanovic, C.; Dekorsy, A. Massive machine-type communications in 5G: Physical and MAC-layer solutions. *IEEE Commun. Mag.* **2016**, *54*, 59–65. [CrossRef]
166. Li, G.; Wu, J.; Li, J.; Wang, K.; Ye, T. Service popularity-based smart resources partitioning for fog computing-enabled industrial Internet of Things. *IEEE Trans. Ind. Inform.* **2018**, *14*, 4702–4711. [CrossRef]
167. Ranganathan, K.; Foster, I. Decoupling computation and data scheduling in distributed data-intensive applications. In Proceedings of the 11th IEEE International Symposium on High Performance Distributed Computing, Edinburgh, UK, 24–26 July 2002; p. 352.
168. Minkenberg, C.J.; Prisacari, B.; Herrera, G.R. All-to-All Message Exchange in Parallel Computing Systems. US Patent US10042683, 7 August 2018.
169. Gkatzikis, L.; Koutsopoulos, I. Migrate or not? exploiting dynamic task migration in mobile cloud computing systems. *IEEE Wirel. Commun.* **2013**, *20*, 24–32. [CrossRef]
170. Rahman, M.A.; Hossain, M.S.; Hassanain, E.; Muhammad, G. Semantic Multimedia Fog Computing and IoT Environment: Sustainability Perspective. *IEEE Commun. Mag.* **2018**, *56*, 80–87. [CrossRef]
171. Aazam, M.; Zeadally, S.; Harras, K.A. Fog Computing Architecture, Evaluation, and Future Research Directions. *IEEE Commun. Mag.* **2018**, *56*, 46–52. [CrossRef]
172. Tusa, F.; Clayman, S.; Valocci, D.; Galis, A. Multi-Domain Orchestration for the Deployment and Management of Services on a Slice Enabled NFVI. In Proceedings of the IEEE Conference on Network Function Virtualization and Software Defined Networks, Verona, Italy, 27–29 November 2018.
173. Baranda, J.; Mangues-Bafalluy, J.; Pascual, I.; Nunez-Martinez, J.; de la Cruz, J.L.; Casellas, R.; Turyagyenda, C. Orchestration of end-to-end network services in the 5G-Crosshaul multi-domain multi-technology transport network. *IEEE Commun. Mag.* **2018**, *56*, 184–191. [CrossRef]
174. Mahmud, R.; Buyya, R. Modelling and simulation of fog and edge computing environments using iFogSim toolkit. *Fog Edge Comput. Princ. Paradig.* **2019**, *30*, 1–35.
175. Celesti, A.; Mulfari, D.; Fazio, M.; Villari, M.; Puliafito, A. Exploring container virtualization in IoT clouds. In Proceedings of the 2016 IEEE International Conference on, Smart Computing (SMARTCOMP), St Louis, MO, USA, 18–20 May 2016; pp. 1–6.
176. Hassebo, A.; Obaidat, M.; Ali, M. Commercial 4G LTE cellular networks for supporting emerging IoT applications. In Proceedings of the 2018 Advances in Science and Engineering Technology International Conferences (ASET), Dubai, UAE, 26–28 March 2018; pp. 1–6.

177. Du, J.; Jiang, C.; Gelenbe, E.; Xu, L.; Li, J.; Ren, Y. Distributed Data Privacy Preservation in IoT Applications. *IEEE Wirel. Commun.* **2018**, *25*, 68–76. [CrossRef]
178. Boudaa, A.; Belouadah, H. Fault-Tolerant Communication for IoT Networks. In Proceedings of the 2018 International Conference Europe Middle East & North Africa Information Systems and Technologies to Support Learning, Fez, Morocco, 25–27 October 2018; pp. 245–255.
179. Wang, X.; Jia, D.; Sun, C.; Huang, J.; Fei, Z. Low-complexity decoding architecture for rate-compatible puncturing polar codes. In Proceedings of the 2017 IEEE 17th International Conference on Communication Technology (ICCT), Chengdu, China, 25–27 October 2017; pp. 97–101.
180. Wang, X.; Qin, K.; Zhu, Z.; Zhang, Z. Low Complexity Blind Detection Scheme for Polar Codes: A Segmented CRC Approach. In Proceedings of the 2018 10th International Conference on Wireless Communications and Signal Processing (WCSP), Hangzhou, China, 18–20 October 2018; pp. 1–5.
181. Chi, Y.; Liu, L.; Song, G.; Yuen, C.; Guan, Y.L.; Li, Y. Practical MIMO-NOMA: Low complexity and capacity-approaching solution. *IEEE Trans. Wirel. Commun.* **2018**, *17*, 6251–6264. [CrossRef]
182. Hossain, E.; Niyato, D.; Han, Z. *Dynamic Spectrum Access and Management in Cognitive Radio Networks*; Cambridge University Press: New York, NY, USA, 2009.
183. Farhang-Boroujeny, B. OFDM versus filter bank multicarrier. *IEEE Signal Process. Mag.* **2011**, *28*, 92–112. [CrossRef]
184. Vakilian, V.; Wild, T.; Schaich, F.; Brink, S.T.; Frigon, J.-F. Universal-filtered multi-carrier technique for wireless systems beyond LTE. In Proceedings of the 2013 IEEE Globecom Workshops (GC Wkshps), Atlanta, GA, USA, 9–13 December 2013; pp. 223–228.
185. Fettweis, G.; Krondorf, M.; Bittner, S. GFDM-generalized frequency division multiplexing. In Proceedings of the VTC Spring 2009 IEEE 69th Vehicular Technology Conference, Barcelona, Spain, 26–29 April 2009; pp. 1–4.
186. Ayadi, R.; Siala, M.; Kammoun, I. Transmit/receive pulse-shaping design in BFDM systems over time-frequency dispersive AWGN channel. In Proceedings of the 2007 IEEE International Conference on, Signal Processing and Communications, ICSPC 2007, Dubai, UAE, 24–27 November 2007; pp. 772–775.
187. Hassan, N.U.; Lentmaier, M.; Fettweis, G.P. Comparison of LDPC block and LDPC convolutional codes based on their decoding latency. In Proceedings of the 2012 7th International Symposium on Turbo Codes and Iterative Information Processing (ISTC), Gothenburg, Sweden, 27–31 August 2012; pp. 225–229.
188. Alsharif, M.H.; Nordin, R.; Abdullah, N.F.; Kelechi, A.H. How to make key 5G wireless technologies environmental friendly: A review. *Trans. Emerg. Telecommun. Technol.* **2018**, *29*, 1–32.