Resource Allocation in 4G and 5G Networks: A Review

Lavanya-Nehan Degambur[1], Avinash Mungur[2], Sheeba Armoogum[3] and Sameerchand Pudaruth[4]

[1,2,3,4]ICT Department, Faculty of Information, Communication and Digital Technologies, University of Mauritius, Mauritius

Abstract**:** The advent of 4G and 5G broadband wireless networks brings several challenges with respect to resource allocation in the networks. In an interconnected network of wireless devices, users, and devices, all compete for scarce resources which further emphasizes the fair and efficient allocation of those resources for the proper functioning of the networks. The purpose of this study is to discover the different factors that are involved in resource allocation in 4G and 5G networks. The methodology used was an empirical study using qualitative techniques to perform literature reviews on the state of art in 4G and 5G networks, analyze their respective architectures and resource allocation mechanisms, discover parameters, criteria and provide recommendations. It was observed that resource allocation is primarily done with radio resources in 4G and 5G networks, owing to their wireless nature, and resource allocation is measured in terms of delay, fairness, packet loss ratio, spectral efficiency, and throughput. Minimal consideration is given to other resources along with the end-to-end 4G and 5G network architectures. This paper defines more types of resources, such as electrical energy usage, processor cycles, and memory space, including end-to-end architectures, whose allocation processes need to be emphasized owing to the inclusion of Software Defined Networking and Network Function Virtualization in 5G network architectures. Thus, more criteria, such as electrical energy usage, processor cycle, and memory to evaluate resource allocation have been proposed. Finally, ten recommendations have been made to enhance resource allocation besides the whole 5G network architecture.

Keywords**:** 4G; 5G; Resource Allocation; Radio Technology; Wireless Mobile Networks

1. Introduction

Wireless communication is data transfer between at least two devices that are not connected using electrical conductors, but most commonly using radio waves [1]. Fourth-generation mobile communication systems (4G) have complemented and are intended to replace third-generation wireless (3G) networks by using internet protocol (IP) as the common protocol, to shift application control and environment into the users' hands. Fourth Generation (4G) provides higher bandwidth, higher data rate, slicker handoff, and ensures seamless services across multiple wireless networks. Moreover, 4G inter-operates with second-generation wireless communication (2G), 3G, and digital broadcasting systems, by being an IP-based wireless internet in its entirety. Hence, 4G supports various multimedia applications, allows the sharing of resources between diverse users, and provides high data rates over wireless networks [2].

Fifth-Generation mobile network (5G) is the upcoming evolution of mobile digital wireless communication systems, which aims at providing features such as speeds up to 10 Gbps, virtually zero latency, and ubiquitous connectivity. 5G networks are expected to lay out new and wider frequency bands, allied to wider spectral bandwidth per frequency channel. Compared to the current 4G, 5G provides high system spectral efficiency and, hence, achieves larger volume of data per unit area, higher communications reliability, increased number of supported devices, lowered power consumption, and more concurrent and instantaneous connections between devices. 5G is modelled to be entirely based on internet protocol (IP). 5G's IP technology is designed to ensure enough control data for the routing of IP packets related to end-to-end connections according to user policies. This architecture allows 5G to be a unified global standard, enabling anywhere and anytime network availability. Additionally, using IPv6, mobile IP addresses can be assigned based on connected network and geographical positioning [3]. All networks have a common problem, which is an effective and fair allocation of resources between competing users. Bandwidth and buffers on routers and switches are examples of resources that are shared. Fairly allocating network resources is difficult since they are distributed across the network with a multitude of connections between devices. Failure to allocate resources results in many issues such as congestion whereby buffers in routers and switches are choked so that packets are dropped [4].

Resource Allocation is the process during which networking devices fulfill competing resource requirements, such as link bandwidth, and buffer space that applications have [5]. Resource allocation is a complex process because it is not limited to a single level of a protocol, but is partially implemented in routers, switches, network links, and transport protocols. Resource allocation is linked to flow control and congestion control, which are two different concepts. Flow control keeps a fast sender from overflowing a slow receiver with traffic, while congestion control keeps senders from transmitting too much data on a network because of resource limitations. Resource allocation methodologies can be characterized along three dimensions, namely, router-centric against host-centric, reservation-based compared to feedback-based, and windows-based to rate-based. Resource allocation mechanisms address the problem in two ways: firstly, within the network at the routers and switches and, secondly at the network edges in the hosts, or in the transport protocol [5].

The router-centric mechanism makes each router responsible for deciding when and how packets are forwarded, and which packets can be dropped, while also notifying hosts about the number of packets they can send. In a host-centric design, the network conditions are observed by the end hosts, and accordingly adapt their respective behaviors for sending packets [5]. In reservation-based systems, an entity queries the network for an amount of capacity to be allocated for a flow, which is then fulfilled by each router by allocating resources such as buffers and the percentage of the link bandwidth. In feedback-based systems, data is transmitted by end hosts without reserving capacity, but then modulates sending rate based on the received feedback from the routers. There can be implicit or explicit feedback [5].

Resource allocation mechanisms require a method to communicate to the sender, the amount of data that can be transmitted. That can be done with a window or a rate. In protocols such as TCP, a sender gets an advertised window from a receiver, which includes the amount of buffer space the receiver has and limits the amount of data the sender can transmit. A similar window advertisement can be used to support resource allocation within a network to reserve buffer space. Rate-based control is used in multimedia applications that create data at an average rate and require a certain throughput threshold. A Rate-based approach is chosen in reservation-based systems so that different qualities of services can be achieved [5]. To evaluate if a resource allocation methodology is good, the criteria to be used are effective resource allocation and fair resource allocation. Effective resource allocation is about analyzing metrics throughput to delay ratio, while fair resource allocation deals with bandwidth allocation.

However, based on the findings of [6], when traffic flows go through a network and devices, apart from bandwidth and buffer space, resources such as processor, memory space and power, also need to be effectively and fairly allocated so that there is proper service provision and no Denial-of-Service issues. In edge networks, such as the internet or wireless networks, bandwidth seems to be the most critical resource, which should be fairly allocated among all the competing users. Buffers, in network devices, preserve throughput by temporarily storing packets, while a link is being used, so that packets are not discarded. The processor and memory usually work in tandem. In packet-switched networks, packet headers need to be analyzed, or even modified so that traffic flows are transmitted as required. This uses processor cycles for central processing units (CPU) and memory from dual in-line memory modules (DIMM), that is RAM. Thus, processor and memory also need to be allocated to flows and users. Finally, electronic devices use electricity to function and normally do not have an infinite lifetime, so that a fair allocation of electrical power between competing users is a crucial aspect, even more so in mobile networks, ad hoc networks, and sensor networks.

Resource allocation in a network has been observed to be evaluated using five criteria namely, delay, fairness, packet loss ratio, spectral efficiency, and throughput. This paper aims to conduct a systematic and comprehensive review of resource allocation in the two latest wireless network technologies 4G and 5G with the specific objective to identify criteria to enhance resource allocation in 5G networks.

This paper is structured as follows: Section II discusses 3G, 4G and 5G networks. Section III discusses the resource allocation mechanisms in 4G and 5G networks. Section IV describes the research methodology and the architecture analysis. Section V depicts the results obtained and recommendations made. Section VI concludes the paper.

2. Literature Review

In this section, we present a brief history of 1G to 3G cellular communication and an in-depth analysis of the 4G and 5G wireless communication networks. Moreover, this section highlights the detailed architectures of both 4G and 5G networks.

According to [5], in the 1980s, the Japanese firm Nippon Telegraph and Telephone (NTT) developed the 1st generation wireless cellular network (1G), using analog signals based on the Advanced Mobile Phone Service (AMPS). Frequency Division Multiple Access (FDMA) scheme has been used as the multiplexing technique.

Owing to capacity, technological and security issues with 1G, 2nd Generation (2G) was developed in the 1990s in Finland to replace 1G. 2G was based on the Global System for Mobile (GSM) communication, whereby digital radio signals were used so that the available spectrum was more efficiently utilized, security was improved, and text messages could be exchanged, all while using circuit switching. 2G later evolved to use general packet radio service (GPRS) providing internet access. As the need for increased data rate arose, 2G further evolved to use Enhanced Data GSM Environment (EDGE) so that data rates increased by four. Time Division Multiple Access (TDMA) combined with Frequency Division Multiple Access (FDMA) were used by 2G to allow many users to connect at a time on the frequency bands [7].

As more users of mobile phones needed internet access, the faster and more robust 3G was developed at the start of the 2000s. It used Code Division Multiple Access (CDMA) and Wideband Code Division Multiple Access (WCDMA) standards and was backward compatible with 2G technologies. With CDMA, 3G operates in the typical frequency band of 1.6GHz to 2GHz. With WCDMA, a larger carrier frequency was used to cater to a greater number of users as compared to CDMA. 3G systems used both circuit switching and packet switching at the core network. Later, High-Speed Packet Access (HSPA) was developed to allow 3G networks to enhance data speeds. HSPA+ further upgraded 3G networks so that broadband speeds, as high as 42 Mbps, could be achieved by using Multiple Input Multiple Output (MIMO), which is an adaptation of multipath propagation. Using MIMO, on the receiver side, the same signals are received several times to decrease error probability and increase performance. With 3G, the hand-off process was introduced so that connected equipment is connected to two stations at the same time, which prevents calls from being dropped when handing off [7].

According to [8], as wireless network systems became increasingly user-centric, the 4th Generation wireless network (4G) was designed with key features such as user-friendliness, user personalization, terminal heterogeneity, network heterogeneity, evolutionary design, and personalization transfer as its basis. User-friendliness and personalization meant that terminals and users should interact naturally without hassle. Terminal heterogeneity should allow various terminal types and technologies to interact smoothly. Network heterogeneity refers to the growing amount of available network access technologies which, when used together, should provide a seamless service while still differing in terms of coverage, rate of loss, data rate, and latency. As devices were created using evolutionary design to be adaptable, personalized, and less specialized, 4G needed to provide a complete package, maximize the variety of services supported and exploit the variety of personalized terminals in use. Hence, 4G allows interoperability between heterogeneous network technologies to provide better coverage almost anywhere while respecting Quality of Service (QoS) requirements, bandwidth resource QoS utilization optimization, and better power consumption among all the available networks.

As per [9], 4G provides a fully IP-based converged telecommunication network with mobility and very high data rates. This requires establishing a new band on the wireless system to provide high data speeds, asymmetric uplink and downlink rates, uninterrupted coverage, enhanced QoS mechanisms, and massive multimedia information hosting. Also, multiple types of communications networks and media should be seamlessly hosted to allow users to roam between network environments freely. Thus, key technologies in 4G include Orthogonal Frequency Division Multiplexing (OFDM), Software Defined Radio (SDR), smart antennas, and IPv6.

OFDM is a multi-carrier modulation technique whereby a channel is divided into several orthogonal sub-channels, the high-speed data signal is converted into parallel low-speed sub-data streams, and each subchannel is modulated individually before all sub-channels are recombined for transmission. The advantages of OFDM include increased spectral efficiency, anti-fading ability, data transmission at high speeds, and inter-symbol interference (ISI) ability. SDR is the use of programmable software to control a variety of hardware platforms so that terminals adapt themselves to the network wireless interfaces. SDR’s advantages include flexible system structure, system interoperability, improved anti-interference, and hardware integrity. Smart antenna is the use of multi-beam antennas that are adaptive and provide multiband so that signal interference is decreased, signal to noise ratio is improved, spectrum resources overuse is limited, and communication system quality is improved. IPv6 is the core of 4G as the transmission of data streams is based on IP packets. IPv6 provides a huge address space, automatic control, differentiated and customizable QoS, and mobility, since a mobile device can have a fixed home address regardless of its position and whether its connection is active, or not. Based on the findings of [10], 4G features wireless download speeds of 100 Mbps, 50 times faster than 3G, worldwide roaming and mobility, interoperability between many terminals, optimally converged network services, low-cost user-friendly device interfaces, improved GPS, easy scalability and better crisis management. The low costing is because existing networks can be used without retooling or extra spectrum acquisition. Wireless mobile communication ystems can be set up in a matter of days so that a crisis can be managed easily.

To deploy 4G, in addition to SDR and OFDM, Multiple Input and Multiple Output (MIMO) and Universal Mobile Telecommunications Systems (UMTS) standards are used. MIMO uses many antennas at the transmitter and receiver sides to enhance performance. UMTS has been standardized in the Third Generation Partnership Project (3GPP). 3GPP has also standardized Long Term Evolution (LTE) as the 4G standard because most telecommunications operators in the world are members of Long-Term Evolution/System Architecture Evolution (LTE/SAE). LTE is advantageous as it provides download rates of 100 Mbps, upload rates of 50 Mbps, and it provides low latency, which enhances real-time interaction on mobile networks.

A 4G end-to-end network architecture as depicted in Figure 1 consists of four parts, namely, the LTE User Equipment (LTE UE), the evolved UMTS Terrestrial Radio Access Network (eUTRAN), the Evolved Packet Core (EPC), and the IP services part as shown in Figure 1. The IP Services part deals with the internet, IMS, and cloud applications [11]. The LTE UE is a mobile equipment that is built to process 4G technology and contains a 4G enabled UMTS SIM card. The UE handles all communication functions, terminates data streams, and identifies a user on the network.

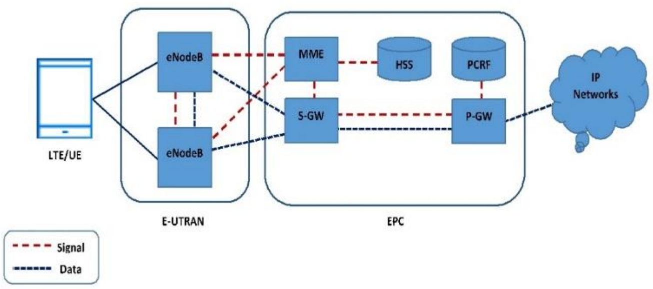

Figure 1. 4G Network Architecture

The eUTRAN is made of base stations called evolved nodeBs (eNodeB), which perform management of radio resources, radio bearer control, connection mobility control, dynamic allocation of resources to UEs in both uplink and downlink, management, and reporting of configuration for mobility and scheduling. The eNodeBs are also involved in compressing IP headers, data streams encryption, data routing from user plane to the serving gateway, scheduling, and transmission of paging messages, and broadcast information from Mobility Management Entity (MME), admission control, congestion control, and data buffering during handover. The eNodeBs also select MME for UE attachment when routes to MME cannot be found from information about UE [11].

The EPC consists of five main components named MME, Serving Gateway(S-GW), Packet Data Network (PDN) Gateway (P-GW), Home Subscriber Server (HSS), and Policy and Charging Rules Function (PCRF). EPC provides an all-IP core network architecture to serve various purposes. Access control is enforced, user authentication is done, and certain application services are implemented at the EPC using user plane and control plane mechanisms. The MME manages UE mobility and keeps UE context when in idle mode, performs Non-Access Stratum (NAS) signaling and security, and manages the bearers of the UE session. S-GW is the mobility anchor for the data bearer, and buffers downlink data when the UE is in idle mode. P-GW provides IP addresses to mobile UEs, filters users' IP packets in downlink direction based on different QoS-based bearers, and performs traffic steering and enforcement of policies. HSS is used to charge users by taking appropriate enforcement decisions and to perform authentication, user identification, and credentials management. Finally, PCRF controls the different planes based on set policies [11]. eNodeBs are connected using X2 interfaces, to MME using S1-MME interface, to S-GW using S1 user plane external interface (S1-U), and to UEs using UMTS air interface or Uu interfaces [11].

Mobile communication networks have evolved from voice exchange communications to an integrated system that can support data transmission of billions of heterogeneous devices [12]. This has made the current 4G/LTE networks reach their limits in terms of capacity and speed when compared to the upcoming needs. Complex networks of today are comprised of several types and standards of services while coordinating multiple technologies. Also, both real-time and non-real-time service requirements are becoming increasingly in demand, which further increases the need for flexibility in network function orchestration, and shortened periods to deploy services.

According to [13, 14], to tackle those challenges and, since the mobile consumer has been given the ultimate priority, the International Telecommunication Union (ITU) and 3GPP have classified three service scenarios that 5G networks need to provide. These are Enhanced Mobile Broadband (eMBB), Ultra-Reliable and Low Latency Communications (uRLLC), and Massive Machine Type Communications (mMTC). The eMBB caters to high data rates and traffic to meet humanity's urge for a digital lifestyle such as video streaming, virtual reality, and cloud gaming. The uRLLC caters for ultra-reliable low latency and low error rates so that industries' strict demands, such as GPS driving, remote management, and automation are met. The mMTC looks after the tremendous number of connections needed to sustain the world's smart digital and optimized lifestyle such as smart agriculture and smart cities and saves energy while doing it. This drives the service-driven nature of 5G networks and the adoption of cloud technologies for its implementation.

5G networks are Software Defined Networks (SDN) and use Network Function Virtualization (NFV) to virtualize the network hardware, and the physical network infrastructure consists of multi-mode sites forming macro, micro, and pico base stations to implement the Radio Access Network (RAN), and of three-layered Data Centers (DC) which provide pools of computing and storage resources to maximize resource utilization [13, 14]. Transport networks connect the three layers of DCs. Using NFV, networks create service-oriented network slices, which are sets of network functions so that service functions are customized and independent. 5G uses CloudRAN, which is built around Mobile Cloud Engines (MCE) to provide network efficiency and multi-connectivity and fuse the network capabilities of several Radio Access Terminals (RAT), frequency bands, and heterogeneous sites.

The flexibility of CloudRAN is also ground proof against uncertainties. 5G networks separate the control plane and user plane by defining the control plane as component-based, and the user plane as programmable. This separation, allied to unified databases for each plane, simplifies signaling interconnections and promotes the use of distributed gateways. This allows the creation of even more customized network slices and satisfies flexible customer needs. Service-Oriented Network Auto Creation (SONAC) is used as the methodology for network automation using techniques such as Software Defined Topology (SDT), Software Defined Protocols (SDP) and Software Defined Resource Allocation (SDRA) to automatically deploy services, schedule and allocate resources, discover, and resolve faults and analyze network data in depth. Moreover, 5G networks have an entirely IP-based core network focused on mobility so that an IPv6 address is assigned to a device according to its connected network and location while supporting many wireless mobile devices. IPv6 removes the need for Network Address Translation (NAT) as it has a near-limitless address space [13, 14].

Since 5G changes the model of Data Centers (DC) with SDN and NFV, their infrastructures evolve as well with the increasing number of virtual machines created to service various virtualization needs. More intelligent DC management systems and procedures are needed to fulfill both the service and the sustainability requirements of the DCs in terms of energy consumption efficiency, running cost, and latency [15]. To achieve the thousand-fold increase in speed from 4G to 5G, millimeter-wave (mm-wave) spectrum, in smaller cells, having wavelengths in the millimeter order and 3-300 GHz range as carrier frequency, are used with occasional traffic offloading done onto unlicensed 5 GHz Wi-Fi spectrum. 5G also uses the sub-6GHz frequency bands used by previous standards for larger cells. 5G also uses OFDM as it combines advantages of Quadrature Amplitude Modulation (QAM) and Frequency Division Multiplexing (FDM) to provide a higher data rate wireless network, even when applied to ultra-dense wireless networks [16-18].

As shown in Figure 2, 5G end-to-end network architecture is made up of the cloud RAN defined as Next Generation RAN (NG-RAN), Multi-Access Edge Computing (MEC) in the Virtual Evolved Packet Core (vEPC), Data Network (DN), and cloud service, which is provided by the network [19]. NS, NFV, NFV Management and Orchestration (MANO), and SDN are the pillars of 5G architecture implementation. SDN decouples control from traffic forwarding and processing provides logically centralized control and allows the programming of network devices [20].

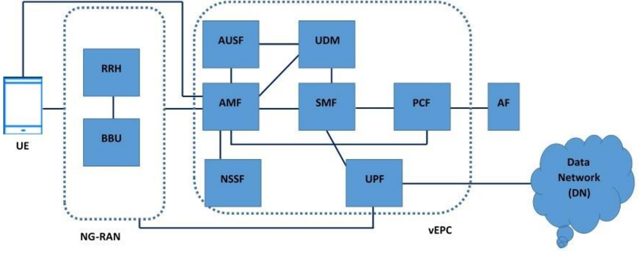

Figure 2. 5G Network Architecture

In NG-RAN, the base station is divided into virtualized Baseband Units (BBU) and Remote Radio Heads (RRH). Since 5G uses a high spectrum, which means high attenuation, the base station is divided into two parts for improved control. MEC reduces latency by acting as a lightweight vEPC, as some functions have been offloaded onto it. The vEPC is a virtualized LTE EPC, which breaks down the LTE EPC functions into new parts and new functions. SDN isolates and provides flexibility in NS lifecycles. NFV is the technique used to virtualize a vEPC's network function. NFV MANO is the management platform for NFVs [19].

NG-RAN is made up of two connected fronthaul networks, RRHs, and pools of BBUs. RRHs tame wireless signals from mobile devices and are sent to the BBU pools using the fronthaul networks. BBU pools, which are virtualized and centrally controlled, can manage many base stations simultaneously and allocate spectrum as per dynamic traffic needs [19]. The vEPC network functions are based on LTE EPC functions that have been virtualized into VNFs, namely, Access and Mobility Management Function (AMF), Session Management Function (SMF), User Plane Function (UPF), Policy Control Function (PCF), Authentication Server Function (AUSF), Unified Data Management (UDM), Network Exposure Function (NEF), Application Function (AF) and Network Slice Selection Function (NSSF) [19]. AMF is concerned with authorization, registration, authentication, connection, access, context, and mobility. SMF steers traffic, enforces policies, allocated IP addresses, and manages sessions. UPF manages packet forwarding and routing and is crucial in connecting to the RAN and the DN. PCF handles the policy rules that manage the plane and handles the policy framework. AUSF is the authentication server. UDM handles authentication credentials, subscriber data, and user identification. NEF is the interface for information exchange between internal and external networks. AF gets access to the NEF and communicates with PCF. NSSF provides the user's required network slice instances and AMF [19].

As NG-RAN and vEPC have been virtualized into VNFs, a software-defined method is needed to implement new services easily, decrease deployment costs, manage running services, and handle resources. This is done by NFV MANO. The general NFV MANO architecture is composed of NFV Infrastructure (NFVI), VNF, Operations Support System / Business Support System (OSS/BSS), Virtual Infrastructure Manager (VIM), VNF Manager (VNFM), and Network Function Virtualization Orchestrator (NFVO). NFVI is the infrastructure that virtualizes computing, storage, and network technologies. VNF is the virtual network function hosted by the NFVI. OSS/BSS caters to telecom services such as user data and billing. VIM manages the NFVI and allocates virtual resources. VNFM configures and manages VNFs, NFVO coordinates the VNFM, and the VIM based on OSS/BSS requirements to orchestrate services [19].

3. Resource Allocations in 4G and 5G Networks

In 4G LTE networks, Radio Resource Allocation (RRA) is done on the Downlink (DL) channel from eNodeBs to UEs [21]. The eNodeB informs the scheduler about the channel conditions by signaling the Channel Status Information (CSI), and about the applications services required based on QoS specifications. Resource Allocation (RA) is done using 10ms of radio frames in a time-frequency grid. Each frame is broken down into 1ms subframes, which are again divided into slots of 0.5ms. In each slot, a user request is mapped to 7 OFDM symbols, and a Resource Block (RB) is made of 12 subcarriers of 180 kHz. Radio Resource Allocation (RRA) algorithms can be categorized as channel aware, whereby channel conditions are considered, or QoS aware, whereby QoS specifications by applications are considered. The main RRA algorithms used in 4G LTE are [21]:

1. Maximum Throughput (MT) which uses the user’s channel condition. It provides the best possible spectral efficiency by serving the user having the highest data rate. It does not care about the QoS requirements of users and is unfair towards users with bad channel conditions.
2. Proportional Fair (PF) adds a fairness factor to MT to guarantee RA to users with bad channel conditions. It is channel-aware and uses a user’s history. As with MT, PF is not QoS aware and thus cannot serve real-time flows.
3. Modified Largest Weighted Delay First (M-LWDF) is a channel-aware iteration of the LWDF algorithm, which uses the probability of packet drop to handle both real-time and non-real-time flows.
4. Exponential PF (EXP-PF) uses the properties of an exponential function applied to past average throughput, which comprises the head of a line delay, the number of users, and packet drop probability. It is a QoS-aware iteration of PF.
5. Exponential Rule (EXP-Rule) is a form of EXP-PF that uses user resource block (RB) and spectral efficiency measured in bits/sec/Hz, instead of average throughput.
6. LOG-Rule is also a modified version of EXP-PF whereby a logarithmic function, that uses the probability of packet drop and head of line delay, is applied to past average throughput.
7. Frame Level Scheduler (FLS) defines two levels of scheduling in LTE DLs. At a high level, FLS controls the amount of data that each real-time flow can transmit to avoid violating delay bounds. At the low level, the scheduler uses MT to maximize throughput by assigning RBs every Transmission Time Interval (TTI).

Apart from MT and PF, all algorithms mentioned above can perform resource allocation in real-time and non-real-time flows on 4G networks. However, MT and PF can be used alongside the other algorithms within their constraints to schedule real-time flows. Also, other algorithms have been created for specific purposes such as VoIP flows, video flows, gaming flows, Device-to-Device (D2D) communications, and Heterogeneous Networks (HetNets). 3GPP has yet to standardize any specific RRA algorithm for 4G [21].

An energy-efficient RA scheme has been provided for D2D communications in 4G networks since D2D communications can increase interference between cellular devices and increase energy consumption to fit QoS [22]. In the RA scheme, the base station is the main decision-maker, as it performs RA on uplink resources optimally for each D2D link by first checking if QoS is not violated by admitting a D2D pair as an underlay of the cellular network, and then by determining the minimum power level needed for transmission. Finally, a cellular user partner is found for which the network’s overall power usage is minimized. This scheme increases the capacity of a fully loaded 4G network and minimizes total power consumed in uplinks while maintaining QoS.

Research in [23] proposes an efficient RA scheme adapted to 4G networks in China, which uses an all-IP Differentiated Service platform (DiffServ) and an adaptive buffer sharing scheme to provide assured QoS for real-time multimedia traffic flows. The DiffServ architecture connects HetNets to the internet, allows global roaming, guarantees QoS for multimedia traffic, and provides broadband services. In the DiffServ architecture, a DiffServ domain is created by grouping nearby RANs that use the same air interface. All domains are then connected to the DiffServ backbone so that internet services are provided to mobile stations. The buffering scheme provides packet loss and delays guarantee by provisioning UDP layer-coded multimedia traffic. Techniques, such as packet loss analysis, statistical multiplexing, and adaptive optimal buffer configuration are used to provide maximum utilization of resources. Also, to enhance TCP performance on the air interface, TCP and link layers are coupled and cross-layer optimization is performed to provide efficient RA. The same architecture can thus be leveraged for other 4G networks outside China.

RA is the main pillar of 5G wireless networks, and its energy usage and spectrum allocation are key to efficient and fair RA schemes. Since a new spectrum is used, spectrum sharing is at the forefront of RA in 5G networks. There are two main spectrum sharing mechanisms currently in use: distributed and centralized [24]. Using the distributed methodology, systems coordinate between themselves equally to share the spectrum. This allows the management and regulation of transmissions that cause interference between systems and makes resource allocation more efficient in a local framework. Centralized mechanisms allow a central device or unit to coordinate with each connected system discretely, without systems communicating between them. It enables systems that need granularity in spectrum sharing to be effectively serviced while providing reliable control. However, it does not cater to interference between systems.

Nevertheless, 5G networks require a reconsideration from traditional RA mechanisms, since intelligent mobile UEs coupled with IoT require increased compute resource, location-awareness, and ultra-low latency from the underlying network, giving rise to Multi-Access Edge Computing (MEC), which expands the cloud-driven architecture of 5G Network [25]. MEC has pushed some data offloading and computing tasks further to the network edge so that Base Stations (BS) will compute more and use more energy. Hence an RA procedure called Energy-Aware Adaptive Management (ENAAM) has been created that involves dynamically enabling the triggering off/on/sleep modes of BSs, soft scaling of VMs, transmission driver tuning for real-time data flows based on traffic loads and energy consumption, and harvest forecast. To minimize energy consumption, a Long, Short-Term Memory (LTSM) neural network is used to forecast traffic load and energy harvest, based on Limited Look-ahead Control (LLC) policies, so that QoS is ensured while also optimizing power usage.

For improved channel allocation, quality of transmission and spectrum usage must both be considered [26]. In this hybrid approach, in macro base stations communications, a centralized RA mechanism is used, while a distributed RA mechanism is used in micro base stations communications. The spectrum is broken into two sets with one set being allocated to a specific cell, while the other set is shareable between handoff calls. High QoS is maintained by keeping resources at the edge of cells, and lower QoS requirements are fulfilled by locating other resources at the center of these cells. Mobile station queues buffer their corresponding call requests and, when it is time for scheduling, resources are allocated to corresponding users according to information on channel availability. An RA matrix contains information about neighboring network devices and their respective allocated resources. In the RA matrix, the entry for an allocated resource is 1, while for an unallocated one it is 0. Hence, the hybrid approach improves reliability and transmission power.

The multitude of eNodeBs in macrocells of 5G ultra-dense cellular networks decreases performance considerably if the traffic intensity of the eNodeBs varies too much. To optimally allocate resources in 5G ultra-dense networks, a harmony-in-gradation RA mechanism has been proposed [27]. The harmony-in-gradation method uses an innovative formula called the harmony-in-gradation index, and the harmony-in-gradation coefficient that makes the bond between the RA fairness index and the level of network performance. The harmony-in-gradation index is calculated as the ratio of eNodeB traffic and eNodeB backhaul throughput. The calculated index is used with eNodeB of the small cells to send traffic flows from eNodeB backhauls to the macro cell gateway. The harmony-in-gradation coefficient labels the network availability levels. This fairly allocates throughput at the macro cell gateway.

Since 5G networks heavily involve D2D communication, the algorithms, as follows, have been proposed for RA; a heuristic algorithm for light load scenario, heuristic algorithm for medium load scenario, implementation of distributed information-theoretic link scheduling, inverse popularity pairing order algorithm, allocation of resource block and transmission power using message passing, coalition formation algorithm for the spectrum sharing problem, and iterative resource allocation algorithm [27].

Some novel methods based on economics concepts have been defined in [28] to be used as distributed RA mechanisms, namely, stable matching, factor-graph-based message passing, and distributed auction. A stable matching is based on matching theory and assigns resources to transmitters on the network based on their preferences. Factor-graph-based message passing nodes exchange messages among themselves and form a graphical structure to solve RA problems involving computational loads. Distributed auction involves transmitters that modulate the cost of resources and bid for resources, which are then allocated to the highest bidder.

4. Methodology and Architectural Analysis

For this paper, the methodology used for the research was an empirical study that used qualitative techniques, and the research instruments were systematic reviews based on technical analysis of documents. First, technical documents, such as research papers from scientific journals, reviews, and conferences, technical white papers, technical guides, technology websites, and service provider guidelines were collected and reviewed.

Materials on 1G to 5G were studied to understand the history of these technologies. The end-to-end network architectures and resource allocation processes in 4G and 5G networks were analyzed. Following this background study, systematic reviews were performed on different factors affecting resource allocation in 4G and 5G networks to critically appraise the various relevant research and observe discrepancies and missing links, following which some recommendations have been made.

In 4G networks, resource allocation mechanisms focus on radio spectrum allocation and bandwidth allocation. Also, most resource allocation mechanisms focus on Uu interfaces between User Equipment (UE) and evolved UMTS Terrestrial Radio Access Network (E-UTRAN) and, S1 interface between E-UTRAN and Evolved Packet Core (EPC). As 4G networks have become limited in their capacity to sustain today’s wireless requirements, 5G was developed and is being deployed worldwide. Hence the following analysis focuses on 5G end-to-end network architecture only.

In the case of 5G networks, resource allocation mechanisms focus on allocating radio spectrum, bandwidth, and energy, and they are mostly applied between UEs and Next Generation-Radio Access Network (NG-RAN), and between NG-RAN and Virtual Evolved Packet Core (vEPC). However, since 5G networks have brought mobile edge computing, Device-to-Device communication, Software Defined Networking, Network Function Virtualization, and Heterogeneous Networks Interoperability, Limiting Resource Allocation only to those sections of the end-to-end network architecture does not cover the wide range of interacting components in 5G networks.

There are no definitions of specificities in terms of operating systems and quality of service mechanisms needed by UEs to use 5G technology and enable D2D and heterogeneous networks. Since the conventional hardware-defined RAN has been virtualized into BBUs and RRHs in the NG-RAN, the infrastructure hosting the virtualized services consumes electrical energy using only conventional mechanisms to support the BBU and RRH functions.

Concerning D2D communications specifically, resource allocation is performed on the NG-RAN so that the process of resource allocation is separated from UEs, which are the actual protagonists of D2D communications. These mechanisms add further load on NG-RAN in terms of energy expenditure, processor cycles, and memory usage considering the ad-hoc nature of D2D communications.

Different service needs are fulfilled by different service level agreements based on flexible independent network slices on the vEPC using network slicing. This flexibility aspect of network slices, provided by virtualization, emphasizes on the fact that rigid resource allocation does not necessarily apply in the 5G network context. Instead, a holistic and customized approach to resource allocation is required for the infrastructure hosting the virtualized services.

5G end-to-end architecture has been defined in terms of functionalities being provided by each part without making specifications about the hardware platform that should be used to support the Software-Defined Network itself [29]. The need to allocate resources for the hardware platform has not yet been focused on, considering the 5G architecture. Conversely, the hardware platforms' resources are assigned to virtual machines, each of which is the host of independent virtualized network services. Nonetheless, these processes are controlled by management, automation, and orchestration platforms, requiring customized policies.

NFV MANO itself requires an independent infrastructure so that its resources must be allocated in a customized manner while still being decoupled from the resource allocation process of the 5G network technology. The 5G architecture has been tailor-made to be customizable and scalable as per need, which foretells the continuous application of technologies that will require the least human intervention to fulfill those dynamic needs.

Furthermore, the previous criteria defined as delay, fairness, packet loss ratio, spectral efficiency, and throughput are not enough to cover all the resources used to deliver 5G technology. Resources such as electrical energy, processor cycles, and memory are crucial in 5G networks and, hence, criteria such as energy usage, processor speed, and memory usage need to be defined and measured.

5. Results and Discussions

Following the in-depth analysis of 4G and 5G networks architectures in the previous section, eight criteria have been identified to be taken into consideration in the resource allocation process in 5G networks. Table 1 below describes the eight criteria Delay, Fairness, Packet Loss Ratio, Spectral Efficiency, Throughput, Electrical Energy Usage, Processor Cycle, and Memory.

Table 1. Criteria for Resource Allocation

Criteria	**Description**
Delay	Delay is defined as the latency measured when packets are transmitted from a source to a destination, and it is measured in milliseconds.
Fairness	Fairness is the percentage of QoS requirements met during resource allocation.
Packet Loss Ratio	Packet loss ratio is the ratio of packets received by the receiver to the number of packets sent by a sender.
Spectral Efficiency	Spectral efficiency is the rate at which information is transmitted on a provided bandwidth and it is measured in bits/sec/Hz.
Throughput	Throughput is the number of bits in a flow processed per unit of time in a network and it is measured in Kbps.
Electrical Energy Usage	Electrical energy is the electricity used by electronic devices such as routers, switches, servers, and equipment to perform their work. It is usually measured in kWh. Different devices have different needs in terms of electricity according to their functionalities within the 5G end-to-end infrastructure, so having a differentiated approach to electrical energy resource allocation is needed.
Processor Cycle	The processor cycle is the time taken to run an elementary instruction by a processor in a machine or computer. SDN and NFV have increased the need for customized high-end processing on every device along with the 5G end-to-end infrastructure so that assigning processor cores and processor cycles no longer require a one size fits all approach.
Memory	Memory is the amount of main memory or Random-Access Memory (RAM) that is available for all devices in the 5G end-to-end architecture to store running programs. Virtual machines hosting VNFs now have main memory requirements tailor-made for their functionalities.

Based on the results obtained, the following recommendations have been made to enhance resource allocation in 5G networks.

1. Starting with user equipment, to use 5G technology, UEs need to have a specific 5G enabled processor and antenna. Such UE processors should be designed to provide efficient energy resource allocation, processor cycle allocation, and memory space allocation to support full-fledged 5G offerings.
2. The operating system running on the 5G UEs should also include energy, processor cycle, memory space, and spectrum resource allocation mechanisms, and offer users and sub-processes the option to use those mechanisms so that quality of service requirements by users can be defined by users themselves on the go, and via subscription as control-plane and user-plane have been split in 5G networks.
3. On the next generation-radio access network, since base stations are no longer responsible for decision making, instead of being controlled by baseband units, time, spectrum, and energy allocation for base stations can be dynamically defined at Baseband Units (BBU) so that they become energy efficient. This allows remote radio heads to avoid resource wastage.
4. To maximize the usage of Device-to-Device communications (D2D) among UEs, NG-RAN could offload the resource allocation process to UEs so that base stations do not consume resources. UEs could use their processor, memory, and battery to process resource allocation only towards D2D communication requirements.
5. Virtualization platforms required to virtualize vEPCs functionalities usually run on converged or hyper-converged infrastructure, which consists of computing modules, storage modules, and network modules, which also need holistic energy, processor cycle, and memory space resource allocation process described above.
6. Resource allocation in terms of electrical power and network bandwidth for the hardware platform supporting the virtualization technology itself, must be considered by optimizing the supply of electrical power and network bandwidth allocation of the hardware platforms according to the service they provide along with the end-to-end 5G architecture.
7. Virtual machines, hosting the Virtual Network Functions (VNF), should have a dynamic customized resource allocation process through automation and orchestration platforms to suit their functionalities. Customized policies can allocate resources such as processors and memory to virtual machines based on specific scenarios.
8. Including Network Function Virtualization Management and Orchestration (NFV MANO) in 5G network architectures further accentuates the burden on the underlying infrastructure, so that a holistic resource allocation policy must be inbuilt in the MANO platforms, or in the infrastructure management systems.

9. Virtual infrastructure managers, which are part of MANO should not only manage NFV infrastructures, but also police over the virtual resource allocation process among VNFs so that this process is decoupled from the other components of the 5G architecture.

10. Since the core of the vEPC is software-defined, and runs on a virtualized platform, data analytics services and deep neural networks hosted on virtual machines can be integrated into MANO platforms for intelligent resource allocation analysis and optimization.

6. Conclusions

In this paper, the history of wireless telecommunication networks from 1G to 3G has been described to understand the rise of 4G and 5G networks. The key issues that gave rise to the need for 4G and 5G networks have been discussed together with their requirements. The end-to-end architectures of 4G and 5G networks have been explained, highlighting the differences between components from each generation, contrasting with the needs they respectively need to fulfill. Also, resource allocation in computer networks has been defined, and resource allocation mechanisms in 4G and 5G networks have been described alongside the goals they achieve. Research has shown that, alongside bandwidth, memory space for the buffer, processor cycles, and electrical energy, form part of the resources allocated along with end-to-end network architectures. Various criteria, such as delay, fairness, packet loss ratio, spectral efficiency, and throughput, have been observed to measure resource allocation in 4G and 5G networks. However, it was discovered that they primarily focus on Radio Resource Allocation besides the end-to-end networks, while neglecting other types of resources. Since the application of Software Defined Networking, Network Function Virtualization and cloud-defined infrastructure have defined new bottlenecks for resource allocation in 5G networks. Hence, it is proposed to expand the evaluation criteria to include electrical energy usage, processor cycle, and memory. Finally, since 5G technology is replacing 4G networks, ten recommendations have been proposed to holistically enhance resource allocation coupled with the multiple components of the 5G architecture. Hence improving resource allocation can deliver the promised 5G services.

References

[1] D. I. Kaklamani, Athanasios D. Panagopoulos, Panagiotis K. Gkonis, “Antennas and Propagation Aspects for Emerging Wireless Communication Technologies,” Electronics, Vol. 10, No. 8, 2021.

[2] U. Varshney, R. Jain, “Issues in emerging 4G wireless networks,” Computer, Vol. 34, No. 6, pp. 94-96, 2001.

[3] N. Akkari, N. Dimitriou, “Mobility Management Solutions for 5G Networks: Architecture and Services,” Computer Networks, Vol. 169, 2020.

[4] A Javed, J. Harkin, L. McDaid, J. Liu, “Predicting Networks-on-Chip traffic congestion with Spiking Neural Networks,” Journal of Parallel and Distributed Computing, Vol. 154, pp. 82-93, 2021.

[5] J. Feng, Q. Pei, F. R. Yu, X. Chu, J. Du, L. Zhu, “Dynamic Network Slicing and Resource Allocation in Mobile Edge Computing Systems,” IEEE Transactions on Vehicular Technology, Vol. 69, No. 7, pp. 7863-7878, 2020.

[6] Y. Zhou, “Resource Allocation in Computer Networks: Fundamental Principles and Practical Strategies,” Doctoral Dissertation, Drexel University, 2003.

[7] S. Patel, V. Shah, M. Kansara, "Comparative Study of 2G, 3G and 4G," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Vol. 3, No. 3, pp. 1962-1964, 2018.

[8] S. Frattasi, H. Fathi, F. H. P. Fitzek, R. Prasad, “Defining 4G Technology from the User’s Perspective,” IEEE Network, Vol. 20, No. 1, pp. 35-41, 2006.

[9] S. H. Ahmadpanah, A. J. Chashmi, M. Yadollahi, “4G Mobile Communication Systems: Key Technology and Evolution,” 3rd National Conference on Computer Engineering and IT Management, Teheran, Iran, 2016.

[10] A. H. Khan, M. A. Qadeer, J. A. Ansari, S. Waheed, “4G as a Next Generation Wireless Network,” International Conference on Future Computer and Communication, Kuala Lumpur, Malaysia, 2009.

[11] K. Gomez, L. Goratti, T. Rasheed, L. Reynaud, "Enabling disaster-resilient 4G mobile communication networks," IEEE Communications Magazine, Vol. 52, No. 12, pp. 66-73, 2014.

[12] P. K. Agyapong, M. Iwamura, D. Staehle, W. Kiess, A. Benjebbour, "Design considerations for a 5G network architecture," IEEE Communications Magazine, Vol. 52, No. 11, pp. 65-75, 2014.

[13] Huawei, “White Paper: 5G Network Architecture - A High-Level Perspective,” Huawei Technologies Co. Ltd, Shenzhen, China, 2016.

[14] A. Gohil, H. Modi, S. K. Patel, "5G Technology of Mobile Communication: A Survey," International Conference on Intelligent Systems and Signal Processing (ISSP), Vallabh Vidyanagar, India, 2013.

[15] P. S. Khodashenas, J. Aznar, A. Legarrea, C Ruiz, M. S. Siddiqui, E. Escalona, S. Figuerola, "5G Network Challenges and Realization Insights," 18th International Conference on Transparent Optical Networks (ICTON), Trento, Italy, 2016.

[16] R. N. Mitra, P. Agrawal, “5G Mobile Technology: A Survey”, ICT Express, Vol. 1, No. 3, pp. 132-137, 2015.

[17] L. B. Le, V. Lau, E. Jorswieck, N-D. Dao, A. Haghighat, D. I. Kim, T. Le-Ngoc, “Enabling 5G mobile wireless technologies,” EURASIP Journal on Wireless Communications and Networking, Article 218, 2015.

[18] B. Bangerter, S. Talwar, R. Arefi, K. Stewart, “Network and Devices for the 5G Era,” IEEE Communications Magazine, Vol. 52, No. 2, pp. 90-96, 2014.

[19] T-H. Ting, T-N. Lin, S-H. Shen, Y-W. Chang, “Guidelines for 5G End to End Architecture and Security Issues,” arXiv, 2019. https://arxiv.org/abs/1912.10318.

[20] B. Sokappadu, A. Hardin, A. Mungur, S. Armoogum, "Software Defined Networks: Issues and Challenges," Conference on Next Generation Computing Applications (NextComp), Mauritius, 2019.

[21] A. Ahmad, M. T. Beg, S. N. Ahmad, “Resource Allocation Algorithms in LTE: A Comparative Analysis,” International Conference of IEEE India Council (INDICON), New Delhi, India, 2015.

[22] S. M. Alamouti, A. R. Sharafat, “Resource Allocation for Energy-Efficient Device-to-Device Communication in 4G Networks,” 7th International Symposium on Telecommunications (IST 2014), Tehran, Iran, pp.1058-1063, 2014.

[23] Y. Cheng, H. Jiang, W. Zhuang, Z. Niu, C. Lin, "Efficient resource allocation for China's 3G/4G wireless networks," IEEE Communications Magazine, Vol. 43, No. 1, pp. 76-83, 2005.

[24] A. Gupta, R. K. Jha, “A survey of 5G Network: Architecture and Emerging Technologies,” IEEE Access, Vol. 3, pp. 1206-1232, 2015.

[25] T. Dlamini, “Core Network Management Procedures for Self-Organized and Sustainable 5G Cellular Networks,” arXiv, 2019. https://arxiv.org/abs/1909.09097

[26] S. K. S. Raja, A. B. V. Louis, G. A. Dalton, “Optimal Resource Allocation Scheme in Wireless 5G Networks,” TEST Engineering & Management, Vol. 83, pp. 18529-18535, 2020.

[27] S. Haryadi, D. R. Aryanti, "The fairness of resource allocation and its impact on the 5G ultra-dense cellular network performance," 11th International Conference on Telecommunication Systems Services and Applications (TSSA), Lombok, Indonesia, 2017.

[28] R. Vannithamby, S. Talwar, "Distributed Resource Allocation in 5G Cellular Networks," in Towards 5G: Applications, Requirements and Candidate Technologies, Wiley Telecom, pp. 129-161, 2017.

[29] J. Al-Azzeh, A. Mesleh, Z. Hu, R. Odarchenko, S. Gnatyuk, A. Abakumova, "Evaluation Method for SDN Network Effectiveness in Next Generation Cellular Networks," International Journal of Communication Networks and Information Security, Vol. 10, No. 3, pp. 472-479, 2018.