1.4.2.1 Channel Assignment Scheme

Channel assignment can be seen as the process of allocating available channels to the call requests in the cells. With the prime objectives to increase the capacity and reduce the interference, a variety of channel assignment schemes have been developed which can be broadly classified into types: fixed and dynamic. One of the other channel assignment schemes is hybrid channel assignment scheme introduced by Sallberg et al. [5], which is the combination of those two schemes.

- **Fixed Channel Assignment (FCA):** In this scheme, which was studied by Mac Donald [6], each cell is allocated a predetermined set of channels. Any call request within the cell can only be served with the unused channels of that particular cell. As a result, a call is blocked in the cells with heavy traffic whereas the channels in low traffic cells remain idle. This scheme is the simplest of all the schemes but results in poor channel utilization.

- **Dynamic Channel Assignment (DCA):** While in this scheme, which was explored by Cox and Reudink [7], all channels are kept in a central pool and assigned dynamically to cells by the MSC when a call request is made. In this way, a channel can be used in any cell provided co-channel interference constraints are met. Thus, DCA reduces the likelihood of call blocking, however, it results in increased load on MSC at heavy traffic conditions.

1.4.2.2 Handoff Process

Wireless cellular networks provide mobile users freedom to move throughout the network from one cell to another. Thus, mobility is referred as the most important feature of a wireless cellular communication system. For instance, as discussed by Tarasewich [8], MTs' users may perform e-commerce transactions while moving towards another cell during the processing of their requests. It is thus very essential for network operators to provide users with continuous connections during their requests' session. Usually, such uninterrupted continuous service is achieved by supporting handoff (handover) from one cell to another. When a user moves through the coverage area of a cell while a call is in progress, the MT has to be transferred from current BS's control to another BS's control without interrupting the call. This process is termed as *handoff*. The BS transfer includes converting the call to an available channel within the new cell's allocated frequency subset. The processing of handoff is a crucial task in any cellular system. It must be fast and efficient in order to ensure the continuity of connections and the QoS perceived by the users.

Handoffs should be performed as infrequently as possible and seamlessly while being imperceptible to the users. It is often initiated when a mobile user either crosses a cell boundary or whenever the received signal quality deteriorates in the current channel. The decision-making process of handoff may be centralized or decentralized, that is, may be made at the MT or network. The selection of the new channel is made by taking into account the spectrum availability and the network load. After the successful handoff process, the user communicates through the BS in the new cell. More details on challenges and techniques associated with handoff management can be found in the survey papers by Yan et al. [9], Chandavarkar and Reddy [10], and Tayyab et al. [11].

1.4.2.3 Call Admission Control

Call Admission Control (CAC) is one of the fundamental techniques for traffic management. Whenever a call arrives (new call or handoff call) in the network, the system has to decide whether this particular call should be admitted by the network. An algorithm making these decisions is referred as a CAC algorithm. Such decisions are based not only on the available network resources but also on the QoS requirements of the call requests. For a detailed review of the work done in the domain of CAC, one may refer to the survey paper by Katzela and Naghshineh [12] and the article by Abdulova [13].

In the design of CAC schemes, blocking probability of new calls and forced termination probability of ongoing calls are the two fundamental QoS measures for traffic analysis. When a call request from a mobile user is initiated, it may either be accepted or denied due to the unavailability of channels. This denial of service request is known as call blocking, and its probability as call blocking probability. Efficient CAC schemes should achieve low blocking probability. With regard to the handoff process, if the destination cell has no available channels, an ongoing call is terminated (dropped) due to handoff failure. The probability of such an event is known as forced termination (call dropping) probability. One of the key QoS issues in cellular networks is how to control the handoff failure. From the users' perspective, terminating a call in progress is more undesirable than blocking a newly requested call. Thus, handoff calls are treated with a higher priority over new calls while allocating the channels. This can be realized by handoff prioritization CAC schemes. Various priority-based schemes have been proposed, they can be classified into two broad categories as summarized below.

- **Guard Channel Scheme:** This scheme introduced by Hong and Rappaport [14], reserves some fixed number of channels called guard channels in each cell to be used exclusively by handoff calls. The remaining channels (called ordinary or common channels) are shared both by new and handoff calls. Later there was another scheme proposed by Ramjee et al. [15], in which new calls are allowed to access the guard channels with a certain probability. This was termed as *fractional guard channel* (FGC) scheme. These channel reservation schemes can be further classified as static reservation scheme explored by Chen and Lee [16] and dynamic reservation scheme explored by Hossain et al. [17], depending on whether the set of reserved channels is fixed or varies according to the traffic conditions. As a result, such schemes help to reduce the forced termination probabilities however, they could increase the blocking for new calls. Appropriate estimation of the number of channels to be reserved is therefore considered as one of the most challenging tasks in designing an efficient CAC scheme.

- **Queuing Priority Scheme:** This scheme proposed by Tekinay and Jabbari [18], gives priority to handoff requests by allowing them to queue instead of denying access if it finds no idle channel available in the target cell on arrival. The process is accomplished during the time an MT spends in the overlapping service area of cells, called the handoff area. The MT there maintains the communication via the existing channel of the source cell. If a channel in the destination cell is available before the MT crosses the handoff area, the channel is assigned to the call request. Otherwise, the call is forcibly terminated. Consequently, this scheme provides the potential to support more ongoing services, i.e., diminishes the forced termination probability however, at the expense of longer (queueing) delay.

The main objective of such CAC schemes is to achieve a good trade-off between the two QoS measures, the call blocking and forced termination probabilities so that a predefined level of QoS can be achieved. An effective CAC scheme must provide efficient spectrum utilization while satisfying users' QoS requirements.

The wireless cellular technology has come a long way since its inception. However, supporting the dramatic increase in service quality and channel capacity in cellular networks is severely limited due to the scarcity of bandwidth. Therefore, researchers are currently focusing their attention on new communications and networking paradigms that can intelligently and efficiently utilize the scarce spectrum. Cognitive radio (CR) is envisioned as a promising technology for future communications and networking that can utilize the limited network resources in a more efficient and flexible way.

1.5 Cognitive Radio and Cognitive Radio Networks

The concept of CR was coined by Joseph Mitola III [19] in 1999 in his Ph.D. thesis, which defines the CR architecture. A CR differs from conventional radio devices in that it is able to sense the best available portion of the spectrum and modify its transmission parameters such as transmission power, frequency, modulation type, and other operating parameters such as antenna beam pattern based on the interaction with the surrounding environment. Primary concern of a CR is to communicate efficiently while avoiding interference with other users. The interaction including sensing and decision making can be either established locally within the radio or can include interactions with the other

CRs in the network. A cognitive radio network (CRN) composed of multiple connected CRs equipped with cognitive features, have the capability of achieving high spectrum efficiency as described by Gavrilovska et al. [20].

The current radio devices are already capable to dynamically adjust their output power level and process incoming signals in order to overcome interference and distortion effects. However, the characteristics of a CR node in a CRN are beyond and outperform such capabilities. CRs are programmable wireless devices. They are enabled with the capability to sense, learn and acquire information from its radio frequency environment regarding transmitted waveform, type of communication network, frequency spectrum, geographical knowledge, available local resources and security policy. With this capability, CRs can identify the spectrum opportunities to communicate whenever and wherever needed. Moreover, the reconfigurable feature of CRs provides the ability to dynamically adjust their operating parameters according to the sensed information in order to attain optimal performance. Interestingly, with a learning process, CRs can adopt past experiences and radio recognition patterns, and hence the dynamic adjustment procedure can be performed without any prior plan. Accordingly, a CR can be regarded as a *smart radio* which is aware of the environment around it and has an adaptive capability different than traditional communication devices, as mentioned in the study of Biglieri et al. [21].

1.5.1 CRN Functions and Applications

In a CRN, the licensed users, also called primary users (PUs), are the users who are authorized to utilize a specific band of the spectrum. On the other hand, the unlicensed users, also called secondary users (SUs), can opportunistically exploit the spectrum holes or white spaces using dynamic spectrum access (DSA) in such a manner as not to cause harmful interference to PUs. Accordingly, the efficiency of spectrum usage gets significantly improved. Fig. 1.2 illustrates the infrastructure based CRN architecture. Resource allocation for PUs and SUs is coordinated by the BS which functions as the central controller in the network.

From mobile phones to wireless local area networks (WLANs), people want to be per-

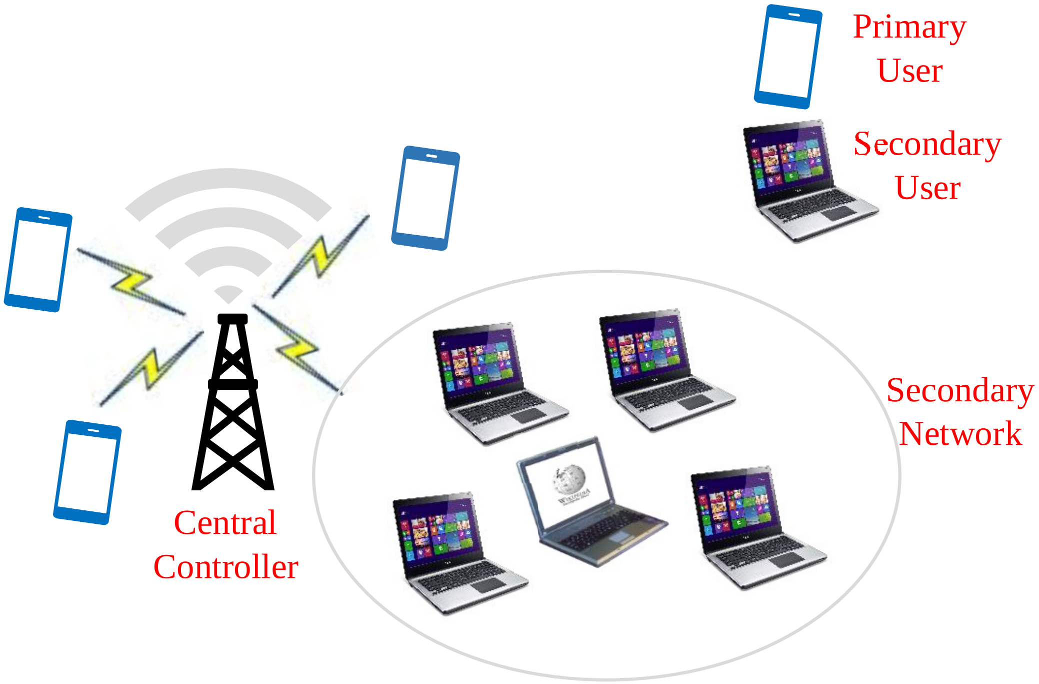

Figure 1.2: A centralized architecture for CRN.

petually networked no matter where they are. Services like mobile phones and global positioning systems (GPSs) use frequencies licensed by the Federal Communications Commission (FCC), while others like WLAN and Bluetooth use unlicensed bands. Figure 1.3 given by Nangare and Rajput [22] shows the spectrum holes and used frequencies in licensed spectrum.

A spectrum hole or white space indicates the temporarily unutilized portion of licensed frequency band, where a PU is absent and a SU is able to transmit without causing interference or degrading the performance of PUs. Hence in a CRN, an efficient spectrum management scheme is desirable which facilitates the channel allocation to both primary and secondary network users while, at the same time, avoids interfering with the primary network users. Examples of primary network include the cellular and television broadcast networks. More specifically, the important functions for spectrum management in a CRN are spectrum sensing and analysis, spectrum decision, spectrum sharing and spectrum mobility, discussed in detail by Mitola and Maguire [23]. These functions are briefly

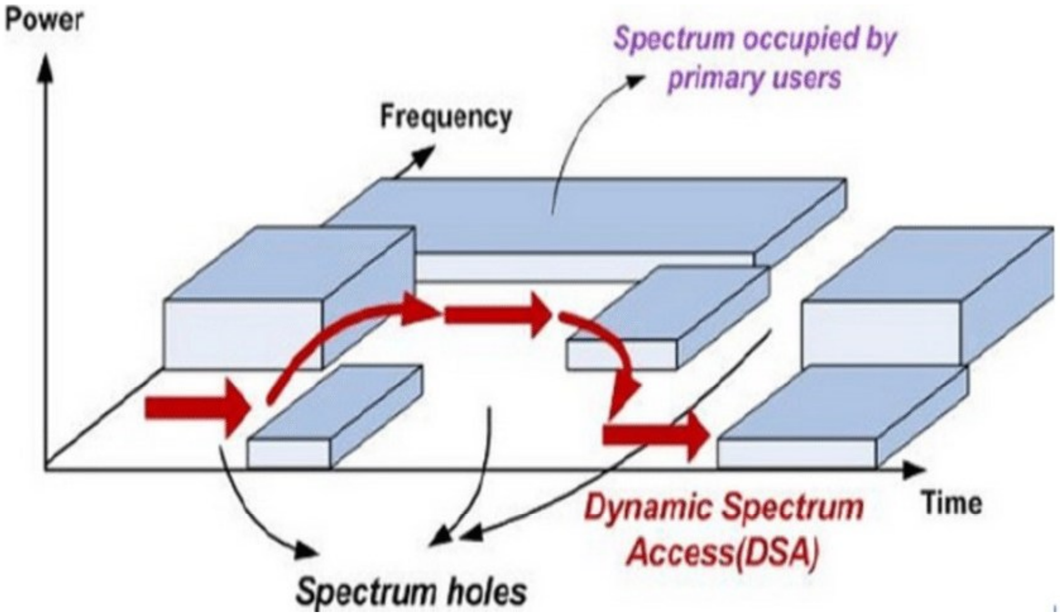

Figure 1.3: Illustration of spectrum holes in licensed band, adapted from Nangare and Rajput [22].

summarized below.

Spectrum sensing enables a CR to obtain necessary observations about its operating radio environment, such as the presence of PUs and detection of spectrum holes. *Spectrum analysis* is required for effective spectrum decision in which obtained sensing results are analyzed. In relation to spectrum sensing, the two forms of prominent errors are false alarms and miss-detections, which were described by Suliman et al. [24].

- *False alarm* occurs when while detecting the presence of PU, CR speculates that the channel is occupied by PU whereas actually no PU is present. This error restricts the SUs to access the channel, decreasing spectral efficiency.
- *Miss-detection*, on the other hand, occurs if the channel actually is occupied by PU whereas CR speculates that it is not. This error enables SUs to establish communication simultaneous to the PU and thereby results in harmful interference, increasing data collision/loss.

Imperfect sensing refers to the sensing process where the probability of false alarm and/or probability of miss-detection is (are) greater than zero, i.e., sometimes the sensing process results in a wrong decision about the state of the channel. Imperfect sensing significantly affects the operation of CRNs and therefore should be considered.

Further, based on spectrum analysis, a decision is made to capture the best available spectrum opportunity considering the user and QoS requirements. *Spectrum sharing* then provides a fair scheduling method to efficiently utilize and allocate the available licensed spectrum. *Spectrum mobility* is another important function of CR as it provides seamless communication requirements during the transition from one frequency to other. This transition refers to spectrum handoff and is possible due to PU's detection at the already occupied channel by a CR user or detection of some other better spectrum opportunity. Following this notion, cognitive communications can improve spectrum usage efficiency and support higher bandwidth services. As a result, CR technology can be employed in many applications as it provides seamless, adaptive and secure communications. Few examples of prospective areas are identified as, e-health services, military networks, intelligent transportation systems, emergency and public security networks.

Another important application is in the next generation wireless networks, which we explore in our work. The CR technology provides intelligence to the user devices that enable them to select the most appropriate network for the service requests. When making a Wi-Fi call, user device utilizes the connected Wi-Fi network when available, or else uses the carrier's cellular network. In addition, depending on which signal has higher strength, it can switch automatically between these two networks throughout the call. This is one example of how CRs can improve communication quality as discussed by Masonta et al. [25]. Beyond this, CR technology provides a broad range of services to meet the challenges of 5GB in the future. Thus, the 5G and beyond cellular networks employing CRs are viewed as a solution that will revolutionize the telecommunication industry and will overcome the limitations imposed by current wireless services design techniques.

1.5.2 Dynamic Spectrum Access

Dynamic spectrum access has been identified as a promising technique to allocate radio spectrum in an efficient and effective manner. The technique of dynamically accessing the unused spectrum band is referred as DSA. CRs plays a vital role to realize DSA in wireless communications. With DSA, SUs are allowed to dynamically search for idle spectrum bands and utilize them to communicate. The licensed spectrum bands are opportunistically accessed by SUs such that the interference caused to PUs is negligible. Through DSA, CRNs can improve the spectrum efficiency as well as the capacity of wireless networks.

According to FCC's Spectrum Policy Task Force (SPTF) [26, 27], at any given time and location, a significant amount of spectrum remains idle without being utilized. A motivating example is the current broadcast television frequency bands where out of a large number of channels, very few channels are utilized in any given TV market. Therefore, the general belief about running out of available radio spectrum does not hold. This aforementioned issue of spectrum under-utilization occurs because the current spectrum allocation policy is static which is unable to allocate the licensed spectrum in a flexible way. The static allocation policy is effective to avoid interference among spectrum bands but also comes at the cost of inefficiency in radio spectrum utilization. Thus, in order to improve spectrum utilization, there is more need of advanced approaches such as dynamic or flexible spectrum allocations so as to accommodate the increasing bandwidth demands in the next generation wireless networks. With the recent advances in CR technologies, DSA is moving towards reality for this purpose.

1.5.3 Design of Dynamic Channel Access Strategies

The effectiveness of a DSA scheme depends on the availability of channels as well as the features of the channel allocation scheme taken into consideration. The channel allocation process is responsible for finding the most suitable spectrum band while satisfying the interference constraints. Recent studies by Do et al. [28] and Azarfar et al. [29] reveal that in order to achieve effective dynamic channel access (DCA), there is a need to facilitate the spectrum allocation strategy with the following requirements. While making use of the unused spectrum, the available channels should be assigned to SUs with the aim of maximizing the overall channel occupancy in the CRN. Also, the DCA scheme is expected to ensure not to compromise the PUs' privilege for accessing the licensed channels when needed and as long as needed. Moreover, an effective channel distribution procedure is required to allocate the available channels to a group of SUs in an environment of fairness.

Several resource management techniques that can be employed with the DCA mechanism have been proposed in order to further enhance the overall performance of CRNs. Among these techniques, the following are utilized in this thesis.

- *Spectrum handoff (handover)* that occurs when a higher priority PU appears at the licensed channel and finds it occupied by SUs. In this case, the ongoing transmission of an SU must be switched seamlessly and immediately to other vacant channel, if one is available, without degrading the QoS.
- *Call buffering* strategy which can be used to reduce the blocking probability and the forced termination probability of SU services. This strategy enhances traffic flows such that the preempted/interrupted SUs can possibly be queued in a buffer to wait until releasing of an occupied channel. In this case, when a queued SU finds a new available channel, it is is allowed to continue its transmission. Consequently, the forced terminating probability of the preempted/interrupted user will get reduced but at the cost of longer queueing delay. Also, a buffer may be used to queue new service arrivals to improve blocking probability.
- *Channel reservation* which is a prioritization mechanism developed for QoS provisioning in CRNs. Specifically, it aims to reduce the forced termination probability of handoff calls by reserving a certain number of channels for spectrum handoff. As a result, blocking probability of new arrivals will increase in many cases however, the percentage of increment depends on the adopted channel access scheme. Thus, channel reservation allows the trade-off between forced termination probability and blocking probability based on the QoS requirements of the users.
- *Selective interruption* which means if an ongoing SU is interrupted by a PU ar-

rival, the flexible resource allocation (FRA) strategy prioritizes the real-time SUs by selectively interrupting the elastic/non-real time SUs prior to real-time SUs.

The existing literature on queueing modeling and analysis for cellular networks is very vast. The next section is restricted to the investigation of the studies relevant to the objectives of this thesis work. Apart from the below literature survey, each chapter contains a detail discussion on the previous studies.

1.6 Literature Review

1.6.1 Historical Development of Queueing Theory

Queueing theory had its origins in the seminal work of Danish Mathematician and engineer, A.K. Erlang [30] in 1909 in the context of telephone traffic congestion. Erlang was responsible for mathematically representing the notion of stationary equilibrium, for the earliest attempt at optimization of a queueing system and for introduction of the Poisson distribution to congestion theory. He established many principal results which are still in use today. Work on the application of the queueing theory to telephony continued after Erlang. Molina [31], in 1927, published “Applications of the Theory of Probability to Telephone Trunking Problems”. Fry [32] extended much of Erlang’s work in the telephone-oriented book entitled “Probability and Its Engineering Uses”, which provided an excellent survey of congestion problems and dominated the literature until the 1950s. In early 1930s, F. Pollaczek [33] did some pioneering work on Poisson input, arbitrary holding time, and single/multiple channel problems, followed by Kolmogorov [34] in 1931, Khintchine [35] in 1932 and Palm [36] in 1938. A survey of early queueing models for communication networks can be found in the work of Kleinrock [37].

A major breakthrough by Kendall [38] occurred in 1951, who gave a general review of congestion theory and examined the $M/G/1$ queueing system, where G stands for general service time distributions. He [2] further extended his work to analyze the properties of multi-server $GI/M/s$ queueing system in order to handle non-Markovian arrival processes, where GI stands for general independent arrivals. According to Baskett et al. [39], the 1970's revolution in the world of communication gave rise to diverse problems and congestion is one of them, which motivated the study of queueing theory as a powerful analysis and design tool. Since then, interest in queueing theory has been fueled by its wide application in telecommunication and data networks (see e.g., Bertsekas et al. [40]; Giambene [41]; Alfa et al. [42]). Therefore, it is no surprise that it has also found equal applicability in cellular networks including CRNs.

1.6.2 5G and Beyond Research Directions

The mobile cellular era started in the early 1980s. Since then, four cellular generations were implemented offering several services. The first generation used analog transmission for speech services, which was studied by Agrawal and Zeng [43]. Then, second-generation (2G) mobile systems were introduced in the late 1980s offering digital speech service. Subsequent generation, 3G supported the proliferation of smartphones by introducing data connections and allowing access to the Internet, as explored by Mishra [44]. The fourth generation (4G) service networks improved data connections, making them faster and better by providing greater bandwidth for uses such as streaming. Mobile Internet connectivity gained widespread popularity with the fourth generation Long Term Evolution (LTE) communication systems introduced in 2009 as discussed by Capozzi et al. [45]. The last decade has seen the dramatic increase in the demand of mobile data due to an increase in the number of new services and applications. Examples of such services/application include Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I), and other industrial applications including sensor networks. This explosive demand of mobile data results in several challenges which shifted the research directions to 5G networks (see e.g., Prasad [46]; Soldani and Manzalini [47]). A White Paper on "Enabling 5G in India", released in February 2019 by the Telecom Regulatory Authority of India (TRAI) [48] includes detailed insights into enabling 5G deployment in India. Recently, Dang et. al [49] presented a vision for 5GB such as 6G communications from a human-centric perspective that could serve as a research guide in the post 5G era.

Each successive generation of cellular networks aims to provide higher capacity and improved QoS. To achieve this, one of the most challenging problems is how to provide efficient spectrum management in a mobile environment as discussed by Zhao et al. [50]. The large amount of radio resources used for handoffs and the effect of handoff techniques on system interference, user satisfaction level and capacity create a demand for efficient handoff techniques. To aid with spectrum management, performance analysis of cellular systems using handoff techniques is carried out by a number of authors.

1.6.2.1 Handoff Techniques

The simplest way of assigning priority to handoff calls is to reserve some channels (guard channels) for them. The guard channel concept was introduced in the mid-'80s for mobile systems however, policies based on guard channels have long been used in telecommunication systems. The guard channel scheme, with and without queueing of handoff requests was investigated by Hong and Rappaport [14] under the assumption that handoff call arrivals follow a Poisson process. Further, Guerin [51] studied a CAC scheme in which the new calls are queued if the number of vacant channels are less than a specified number of guard channels, while the handoff calls are served without restriction, but are blocked if all channels are occupied. In order to provide better QoS to handoff calls while maintaining high throughput for new calls, Choi and Sohraby [52] introduced and analyzed a cut-off priority scheme with guard channels. Therein, under such scheme, when the number of guard channels exceeds a given threshold, the new calls are queued in a buffer. Moreover, based on the buffer occupancy, hysteresis control analysis is carried out then to control the number of guard channels, assuming the handoff traffic following Poisson process. Later, an iterative algorithm to determine the optimal number of channels to be reserved for handoff calls was developed by Jain et al. [53].

Performance analysis of wireless cellular systems adopting the guard channel scheme has been carried out by a number of authors under the Poisson assumption for handoff traffic (e.g., see Ma et al. [54–56]; Jain and Rakhee [57]). Chlebus and Ludwin [58] were the first ones to question the assumption of handoff arrivals being Poissonian. Rajaratnam and Takawira [59] demonstrated that the Poisson assumption may not be appropriate for calls that handoff a great many times. In their later work [60], they presented two-moment analysis of cellular networks with channel reservation and general arrival process of handoff. Such analysis necessitates determining the variance in addition to the mean of the offered traffic processes. Dharmaraja et al. [61] evaluated the performance of cellular networks employing the guard channel scheme and assuming handoff calls' arrival process as a renewal process. More specifically, they obtained the results for exponential, Erlang and hyperexponential interarrival time distributions. Their results validated that the variance (or equivalently coefficient of variation) along with the mean is important to determine the handoff call blocking measures.

In the conventional guard channel scheme, the blocking probabilities vary largely with the change in the number of reserved channels. To overcome this problem, the FGC scheme was proposed and employed by Ramjee et al. [15] for optimal admission control policies in cellular networks. Recursive formulas of new call blocking and handoff dropping probabilities for FGC scheme were derived by Vazquez-Avila et al. [62]. Such formulas allow simple and stable computing of the aforementioned probabilities, especially when the number of channels is large. Later, the particular examples of FGC scheme were analyzed by Goswami and Swain [63], and Safwat et al. [64] to diminish the deteriorating effects of guard channel scheme on new call blocking probability. More specifically, a finite population model employing Limited FGC (LFGC) scheme was developed and analyzed by Goswami and Swain [63], wherein the number of guard channels can vary by a fraction of one. Performance analysis of LTE-Advanced networks was carried out by Safwat et al. [64] in the presence of Uniform FGC (UFGC) scheme, which accepts new calls with an admission probability independent of channel occupancy.

Several variations of the guard channel scheme, with queueing of handoff requests and/or of new call requests are another important schemes that are frequently used in the literature (e.g., see Kim et al. [65]; Halabian et al. [66]; Jain and Mittal [67] for later important work and Ahmed [68]; Lawal et at. [69]; Palle and Takawira [70] for surveys). Dudin et al. [71] demonstrated a highly positive effect of buffering the handoff calls and the channel's reservation using guard channel policy, when analyzing the performance of multi-channel retrial queue for 5G networks. Garah and Oudira [72] recently proposed a queueing handoff scheme involving fixed channel allocation strategy. Based on the measurement of the user's residual time within the handoff region, dynamic queuing of handoff requests is considered in the proposed scheme.

Moreover, an overview about the issues related to handoff initiation and decision to build a handoff management solution in the literature was given by Kassar et. al [73]. Sgora and Vergados [74] provided an inclusive survey of the basic elements, the different types and phases of the handoff procedure. Later, Kumar and Purohit [75] gave a comparison of all the handoff strategies based on execution time, Signal-to-Interference (S/I) ratio, Received Signal Strength (RSS), call handling difficulty, handoff made and generation methods. Additionally, for the spectrum handoff decision mechanism, few authors have studied handoff schemes taking the quality of signal into account, including Koushik et al. [76] and Sibomana et al. [77]. Based on the link quality and user priority, the multi-class spectrum handoff issue was addressed in CRNs by Koushik et al. [76], using the mixed preemptive/non-preemptive $M/G/1$ queueing model. Such model assumes that if the remaining service time is below a predefined threshold, the SUs cannot be interrupted otherwise, it will be moved to the preemptive mode. Sibomana et al. [77] considered the fact that service time varies according to the quality of the wireless channel and hence, modeled the traffic at the base station as a $GI/G/1$ queueing system to evaluate the performance of SU transmission in CRNs.

1.6.2.2 Cognitive Radio Research Directions

As part of the 5G and 5GB paradigm, inspired by the fact that a significant amount of the wireless spectrum remains underutilized, the CR technique has been proposed by Mitola [78]. Due to its great promise to improve the efficiency of spectrum utilization in future wireless networks, the CRs which are conducted in basis of DSA have been consistently attracting extensive research efforts. There are also quite a few recent surveys on CRNs, including by Kumar et al. [79] and Amjad et al. [80]. With the rapid proliferation of CR research, a wide range of innovative research themes and standards have been pursued. Among them DSA explored by Liang [81] and advanced spectrum sensing techniques investigated by Aggarwal et al. [82] are two popular directions. A comprehensive review on the critical ongoing efforts and various challenges associated with CR towards the use of DSA is presented by Kumar et al. [83].

In particular, there are works that treat CRN while taking sensing errors i.e. imperfect sensing into account. For instance, Tang and Mark [84] analyzed the effects of imperfect sensing using a two-dimensional continuous-time Markov chain. Therein, a multichannel system is assumed and the steady-state probabilities are determined by applying the matrix-analytic method. While a discrete-time Markov chain is used to model a CR system with imperfect sensing by Gelabert et al. [85], where sensing periodicity and sensing accuracy are taken to be key parameters. Later on, Ngatched et al. studied [86] the performance of a multichannel CRN with imperfect spectrum sensing, wherein spectrum handover and queueing mechanism are enabled to save blocked and interrupted SUs. Homayouni and Ghorashi [87] considered the spectrum sensing errors in CRN under the sub-banding strategy, such that if all the channels are being used by the PUs and SUs, while a new call request of SU arrives, then the SUs' channels can be divided into two sub-bands and both SUs can coexist in the same channel with half rate, simultaneously. Recently, Wang et al. [88] examined the impacts of sensing failures in CR system where a retrial phenomenon with balking is used to characterize the SUs' repeated transmissions due to the sensing failures and the interruption by PUs.

The growing importance of cellular networks has also stimulated active research into how data can reliably be transmitted over the mobile communication network as discussed by Popovski et al. [89]. As an integral part of 5G and beyond paradigm, ultra reliable communication (URC) has recently been identified as an important emerging area and according to Popovski et al. [90], the research work in this field is still in its initial phase. The approach by Siddiqi et al. [91] to achieve reliable communication suggested allocating channels to the end users in the presence of various failures/breakdowns in the form of uncertainties. Few recent studies which explore the significance of reliability while evaluating the performance exists. For instance, a systematic view on URC in 5G wireless systems was presented by Popovski [92]. Sattiraju et al. [93] introduced a reliability analysis by considering the wireless transmission as a renewal process with variable failure and repair rates. A comprehensive survey that characterizes the reliability analysis methods was given by Ahmad et al. [94]. Therein, the pros and cons of reliability analysis and modeling techniques are also discussed.

Achieving reliable communication in today's cellular network is particularly challenging due to their intrinsic nature. Avizienis et al. [95] mentioned that these networks are interconnected and dependent, one network component failure may impact other network components, or even the entire network. With its salient advantages for DSA, CR is again regarded as a promising technology which can support and maintain highly dependable future communication systems as discussed by Azarfar et al. [96]. In the context of a communication system, according to Ni et al. [97], and Rubino and Sericola [98], dependability is governed by its capacity to provide services that can justifiably be trusted while its components are prone to failures. The theory to analyze such a phenomenon is called as dependability theory in the research community.

In the literature, with respect to metrics, there are two approaches for providing reliable communication i.e. from the traditional QoS viewpoint (e.g., Urgaonkar and Neely [99]; Beyranvand et al. [100]) and from the dependability theory viewpoint (e.g., Schotten et al. [101]; Balapuwaduge et al. [102]). An opportunistic scheduling and allocation scheme for CRNs was proposed by Urgaonkar and Neely [99] to maximize SUs' throughput while providing reliability guarantees for PU transmission. Focusing on the 5G key attributes of very low latency and ultra-high reliability, an analytical framework to investigate how these attributes can be achieved was developed and analyzed by Beyranvand et al. [100]. To estimate the presence or absence of link reliability, availability as a novel metric for URC was introduced by Schotten et al. [101]. Several reliability and availability metrics for channel access based on the dependability theory in multi-channel CRNs were described by Balapuwaduge et al. [102]. Al-Kuwaiti et al. [103] suggested that the common measurable attributes include availability and reliability which needs to be supplemented together with some fault-tolerance feature while evaluating the performance of a dependable system. Recently, Mendis and Li [104] gave the definition of availability of user equipments in the space domain devoted to URC, from the dependability theory's perspective.

The optimization problems of joint access of PUs and SUs can be effectively solved by means of queueing theory. So, although the term CR was introduced by Mitola and Maguire [23] in 1999, only recent literature in CR devoted to application of queueing theory was extensive. The comprehensive survey by Akyildiz et al. [105] devoted to CR was published in 2006. Now, the literature in CRNs with applications of queueing theory is huge.

1.6.3 Traffic Modeling

As mentioned earlier, queueing theory and models assume fundamental role in the performance evaluation of wireless networks. Notably, Markov chain based modeling has been widely used in the literature to evaluate the performance of cellular mobile networks. Modeling can be based on either discrete-time Markov chain (DTMC) or continuous-time Markov chain (CTMC). CTMC based modeling has proved effective due to its accurate modeling, as well as easy computation characteristics as discussed by Dudin et al. [106]. The formulation of Markov traffic models and the relevant performance analysis is discussed in more detail in the next chapter. The selection of the queueing model needs to be aligned with the network scenario under consideration.

1.6.3.1 Queues with Priority

In CRNs, as a rule, it is suggested in the considered models that PUs are always assigned the highest priority over all types of SUs for channel access (e.g., see Wang et al. [107]). For this reason, priority based queueing models are the most frequently adopted models for CRNs as they can capture the different priority requirements of the PUs versus the SUs. For instance, Heo et al. [108] analyzed the performance of SU traffic in a priority preemption network. Therein, to mathematically model the waiting time of a SU in the network, the SU arrival time, the PU activity as well as the preemption probabilities are taken into account. Bassoo and Khedun [109] evaluated the waiting times for PUs and SUs in both preemptive and non-preemptive priority queue. In non-preemptive case, a PU which finds the channel occupied by a SU cannot preempt it instead it waits until channel becomes free, whereas in the preemptive case, it can. The SUs' spectrum handoff behavior based on $M/G/1$ queueing model was analyzed by Yang et al. [110], under preemptive repeat priority discipline. Based on the analysis, they derived closed form expressions for sojourn time and extended data delivery time.

Furthermore, such priority models have also been applied to scenarios where the SUs are differentiated into separate classes, such as SUs which have stringent delay constraints i.e. real-time traffic e.g., audio and video, versus SUs with higher delay constraints i.e. non-real-time or elastic traffic e.g., email and file transfer (see e.g., Alqahtani [111]). In general, real-time services have higher priority over non-real time services whereas best effort services e.g., ethernet, under which no QoS guarantees are supported, are given the lowest priority (see e.g., Lee and Akyldiz [112]). Canberk et al. [113] designed an admission control algorithm to improve the overall throughput and fairness among four different types of CR users. The four different CR user types in the order of their priorities include E1/T1 applications based on constant bit rate traffic having highest priority, video conference users, VoIP users and best effort users having lowest priority. Chen et al. [114] modeled the CRN as an $M/M/1+GI$ queueing system with heterogeneous SUs. SUs with higher priority have preemptive resume priority over lower priority ones, and the former ones have generally, independently distributed impatient waiting time. The analysis therein is further extended to multi-channel queueing model and hence the closed form of expected system delay for lower priority SUs is derived. Recently, the CRN was modeled using an $M/M/1$ multi-class queueing system under a hybrid preemptive/non-preemptive priority discipline by Fahim et al. [115]. The discipline therein assumes that lower priority SUs can no longer be preempted by higher priority SUs when their number of interruptions reaches a certain threshold value. The closed form expressions for the overall system time, the extended data delivery time and the dropping probability for each class of SUs are also derived.

1.6.3.2 Queues with Retrials

Research on cellular networks with retrial queues has captured much attention. In practice, the users finding the channels occupied or failed may temporarily leave the service area and make repeated attempts (retrials) at random intervals to obtain service. Meanwhile, before retrying to occupy a channel again, the blocked or interrupted calls are assumed to be staying at some virtual waiting space called orbit, as depicted in Fig. 1.4. A complete description of retrial queues can be found in the earlier monographs of Falin & Templeton [116] and Artalejo & Gómez-Corral [117]. To estimate the blocking probabilities in cellular networks with retrials, an approximate technique was proposed by Marsan et al. [118]. The technique is based on CTMCs with state spaces whose cardinalities are proportional to the maximum number of ongoing calls within a cell. Artalejo and Corral [119] investigated the effect of channel idle period after each service completion, on the QoS of multichannel retrial cellular system. Later, Tien Van Do [120] analyzed a retrial queueing model for cellular networks, involving the FGC policy.

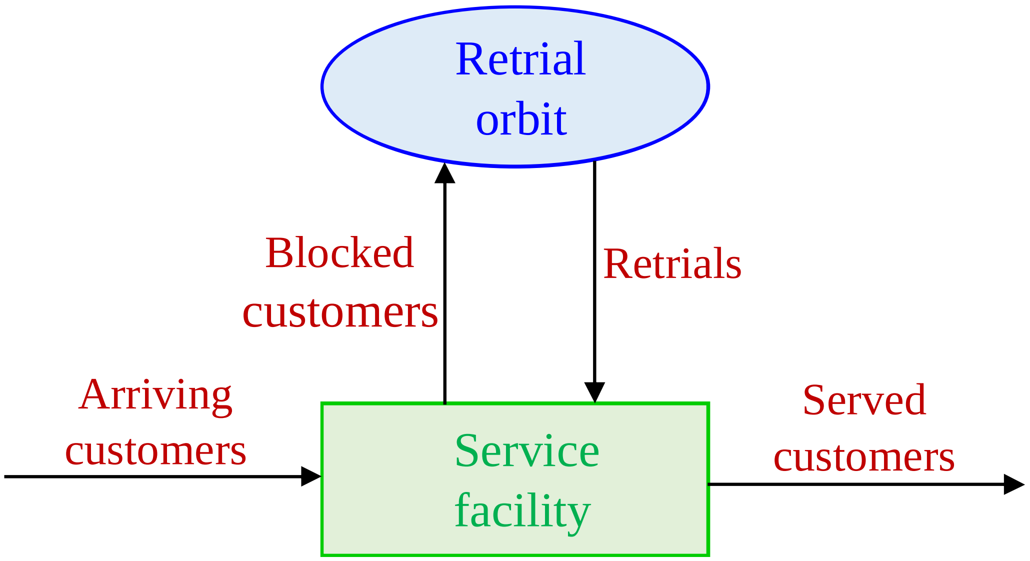

Figure 1.4: General structure of a retrial queue.

With queueing analysis of CRNs, Chang and Jang [121] quantified the spectrum occupancy, delay and throughput under preemptive priority for PUs. They seem to be the first