Cross layer design of Schedulers for Quality of Service Provisioning of Multi-Class Traffic in Wireless Networks

Abstract

In recent times, mobile broadband networks have adopted a packet switched architecture in order to accommodate the growing volume of data traffic. This requires that all traffic classes including real-time voice and video be served as packets. Traditionally, voice and video were transmitted using circuit switching. To provide circuit switched quality to these traffic classes and also to improve it, as in High Definition (HD) voice, stringent Quality of Service (QoS) metrics need to be satisfied. Some of these metrics are the probability of packet drop, average packet delay, average bit-rate, etc. Moreover, Broadband Wireless Access (BWA) networks aim to connect a large number of users and devices. Increasing capacity while guaranteeing QoS of all traffic classes is an important challenge encountered in these networks.

To address this challenge, BWA networks, such as Long Term Evolution-Advanced (LTE-A), LTE-A pro, Worldwide Interoperability for Microwave Access (WiMAX), use Orthogonal Frequency Division Multiple Access (OFDMA) as the channelization scheme and a few efficient cross-layer optimization techniques, such as Link Adaptation (LA), Packet Scheduling-Radio Resource Allocation (PS-RRA), etc. Of these, PS-RRA has emerged as a very important technique for addressing the capacity-QoS trade-off. Hence, one of the focus areas of this thesis is the design of Radio Resource Allocation (RRA) algorithms.

In the first work, we propose to use a single RRA algorithm for Voice over IP (VoIP) and video, unlike existing proposals. This is expected to simply scheduler implementation at the Base Station (BS). For this, we first extend the Dynamic FDPS of VoIP, which schedules user packets based on instantaneous channel conditions, to schedule video. Its capacity is limited by control channel. We propose to further improve its capacity by using the new Joint Time-Frequency (JTF) scheduling algorithm which looks at a user-frequency window for resource allocation. Its supported VoIP capacity is 42.5% more than the 3rd Generation Partnership Project (3GPP) requirement of 40 users per MHz in LTE systems.

The proposed generic real-time traffic scheduling algorithms, Dynamic and JTF FDPS, are then extended to schedule voice, video and best-effort traffic by combining them with a Proportional Fair (PF) best-effort scheduler. The individual traffic capacities supported by these algorithms are obtained from extensive Monte-Carlo system level simulations using a multi-cell multi-user semi-static simulator. It is calibrated with 3GPP results of signal-to-interference-plus-noise-ratio (SINR) and spectral efficiency distribution. It is found that in a mixed traffic scenario the JTF+PF algorithm, along with VoIP and video packet bundling, can support 74 VoIP users per MHz which is 58.5% more than the 3GPP specified VoIP capacity requirement of 40 users per MHz. At the same time, in the same scenario in addition to these 74 VoIP users, JTF+PF algorithm also supports 14 video users per MHz.

System capacity of the RRA algorithms mentioned before is given by the number of deployed users when at least 95% of them are satisfied. A user is satisfied when it experiences no more than 2% packet loss. This system capacity depends on network configurations, such as cell size, user distribution, area SINR distribution, interference, traffic density among different traffic/user classes, and the scheduling algorithms. These configurations cannot be known *a priori* and are different for different BSs. Being a function of these network

configurations, capacity also varies from cell to cell and, hence, cannot be predicted beforehand either. However, if users are admitted beyond the network's capacity then there is a catastrophic drop in their QoS satisfaction. So, controlling the admission of users is essential for providing QoS guarantees, which is one of the goals of this work. Since capacity depends on packet level QoS, accurate evaluation of capacity and QoS is important for efficient Call Admission Control (CAC). Capacity and QoS can be obtained through offline simulations which are exhaustive and time-consuming. On the other hand, deriving capacity and QoS from mathematical models of schedulers is faster and accurate. These models can, therefore, be used online which lead to better dynamic control of the network. So, in the next part of the work, we focus on deriving analytical frameworks of schedulers.

We first develop a queueing theoretic framework of a pre-emptive priority scheduler that serves multiple user classes in the downlink of a finite buffer multi-carrier network. User classes are differentiated by the number of frequency resources or Physical Resource Blocks (PRBs) a user needs to transmit a packet. Considering exponential arrival and service times, QoS metrics, such as probability of packet drop, probability of packet delay, average delay, throughput, and pre-emption probability, are derived from the steady state probabilities of the Continuous Time Markov Chain (CTMC) of the system. Results are validated using discrete event simulations. Packet level Key Performance Indicators (KPIs) obtained analytically from this framework can be used to analyze the system behaviour for various traffic densities of user classes at the input. The KPIs can also be used for threshold based CAC.

We can further improve upon the framework described earlier by considering instantaneous channel variations. This brings the model closer to real life. At the same time, it is also important to explore the flexibility in scheduler behaviour. So, in the next work, we relax the assumption of exponential service times and make it deterministic in keeping with the fixed length scheduling intervals of BWA networks. It is also useful for capturing the channel quality fluctuations. We design analytical frameworks for two different RRA methods. The first one is a Dynamic Priority (DP) scheduler which is characterized by a control parameter. This parameter chooses the priority level of the different user classes for resource assignment at the beginning of any scheduling interval. The second method is a bandwidth reservation algorithm for the different user classes. Users are classified based on whether their average signal to noise ratio is greater or less than a threshold value.

Instantaneous channel variations in both frameworks are modelled using Adaptive Modulation and Coding (AMC) which changes at the beginning of every scheduling interval. Each AMC mode represents the number of PRBs needed for successful transmission of a packet at any scheduling interval. AMC is modelled using a Finite State Markov Chain (FSMC). Thus, the effects of scheduler behaviour, queueing and AMC are combined into the developed framework. Since service times are deterministic, derivation and steady state solution of the Discrete Time Markov Chain (DTMC) of the DP scheduler and bandwidth reservation are used to obtain QoS metrics, such as average delay, average throughput, and the probability of packet drop. Results of both frameworks are verified using discrete event simulations. A comparison of the KPIs of the DP scheduler and bandwidth reservation reveals that the latter improves the QoS of users located closer to the transmitter and degrades that of users located farther away, in comparison to the DP scheduler. So, it may be inferred that although the framework of bandwidth reservation is more structurally simple, the DP scheduler provides a more uniform QoS across all user classes. Hence, it appears to be a more preferred choice for RRA than bandwidth reservation. It is also demonstrated how the

control parameter of the DP scheduler can be used to vary its behaviour between a strict priority and a PF scheduler.

So far we have designed the mathematical frameworks of the schedulers with the goal of accurately evaluating QoS. Since QoS guarantees are usually provided through CAC, we next envisage the use of the QoS metrics estimated by the developed frameworks in designing efficient CAC modules. In the next work, threshold based predictive CAC modules are designed. These modules use the packet level QoS metrics obtained from the pre-emptive priority scheduler as well as the DP scheduler frameworks as inputs. Time scale decomposition is used in the development of the CAC architecture. Discrete event simulation results show that the framework of pre-emptive priority scheduler can be used to predict the maximum number of ongoing calls in the network. It is also shown that the scheduler control parameter of the DP schedulercan be used to regulate the capacity of the system while satisfying QoS constraints of the different classes.

The work done in this thesis aims to design and characterize RRA methods in OFDMA based packet switched wireless access networks. The proposed JTF algorithm has been found to significantly increase the number of users supported by individual traffic classes in a mixed-traffic scenario. Hence, it can be considered as a potential algorithm in OFDMA networks such as LTE-A, LTE-A pro, etc. The mathematical frameworks of the schedulers designed can be used to provide ubiquitous service quality to heterogeneous user classes which is one of the goals of future generation wireless networks. This can be achieved by providing the QoS guarantees through CAC which may help in better dynamic network control.

Keywords: Quality of Service, Broadband Wireless Access Networks (BWA), 4G cellular systems, Packet Switcing, Orthogonal Frequency Division Multiple Access (OFDMA), Packet Scheduling and Radio Resource Allocation (PS-RRA), Adaptive Modulation and Coding (AMC), Link Adaptation (AMC), Mixed Traffic Scheduling, Packet Bundling, System Level Simulator, Queueing Theory, Continuous Time Markov Chain (CTMC), Discrete Time Markov Chain (DTMC), Finite State Markov Channel Modelling, Cross-Layer Design, Call Admission Control, Time Scale Decomposition, Exponential and Deterministic Service Times, Bandwidth Reservation, Real-Time traffic, Non Real-Time traffic