NOTE TO USERS

This reproduction is the best copy available.

UMI[®]

Dynamic Fat Pipe Up-Link Resource Allocation For WCDMA Systems

Dragoslav Culum

**A Thesis Submitted to the
Faculty of Graduate Studies and Research
in Partial Fulfillment of the Requirements
For the Degree of
Master of Applied Science**

Ottawa-Carleton Institute for Electrical Engineering
Faculty of Engineering
Department of Systems and Computer Engineering
Carleton University

© Copyright

2002, Dragoslav Culum

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-79727-9

Canada

The undersigned recommend to
the Faculty of Graduate Studies and Research
acceptance of the thesis:

Dynamic Fat Pipe Up-Link Resource Allocation for WCDMA Systems

submitted by Dragoslav Culum, B. Eng. & Mgt.
in partial fulfillment of the requirements for
the degree of Master of Applied Science

Chair, Rafik Goubran, Systems and Computer Engineering

Thesis Supervisor, (R. H. M. Hafez)

Carleton University
November 2002

ABSTRACT

The third generation cellular systems' uplink resource allocation (RA) task is complicated by a number of unknowns, such as, large variance in uplink data traffic characteristics, lack of synchronization between individual mobile stations on the uplink, and rapidly changing physical channel conditions. It is believed that any long-term scheduling optimization technique would not be very efficient due to the rapidly changing traffic and channel conditions. Inefficiencies in long-term optimization would be costly in terms of throughput, delay, and the overall QoS.

In this thesis we propose to dynamically allocate, and de-allocate resources in order to adapt to the rapid changes in the channel conditions and uplink data traffic. We call the proposed scheme Dynamic Fat Pipe resource allocation scheme DFP. This scheme provides short leases on channel resources, which makes it very responsive to fast changes in the uplink traffic or channel conditions. It utilizes a fat pipe shared channel in order to provide for further optimization of system throughput [4].

The DFP RA technique ensures fairness, maximizes the system throughput, and at the same time supports QoS, as measured by minimizing the delay while supporting the required error performance. It achieves this at a significantly lower implementation complexity and signaling overhead cost than any long-term optimization scheduling technique. Furthermore, the DTDFP RA technique guarantees the stability of the system at all times, and for all possible combinations of uplink data traffic.

Acknowledgements

I would like to express my deepest gratitude to my thesis supervisor, Professor R. H. M. Hafez for his great guidance and encouragement during the course of my thesis work. He manages to strike the perfect balance between providing direction and encouraging independence.

My sister, and my parents have always encouraged, and guided me to independence, never trying to limit my aspirations. I am thankful for their support and generosity.

Table of Contents

1.	Introduction.....	1
1.1	UTRAN Architecture.....	3
1.2	WCDMA Physical Layer Background.....	5
1.3	WCDMA MAC Layer Background.....	9
1.4	MAC Layer Transmission of Data Over Air Interface.....	11
1.5	WCDMA Up-Link MAC Layer Radio Access Problem.....	13
1.6	Organization.....	14
2.	Up-Link Traffic Models and Scheduling Issues.....	16
2.1	Up-Link Data Traffic Models.....	16
2.1.1	ITU Data Traffic Models.....	16
2.1.2	Cumulative Up-Link Data Traffic Patterns.....	18
2.2	Long-Term Scheduling Resource Allocation.....	19
2.3	Short Term Scheduling.....	20
3.	Proposed Scheduling Technique.....	22
3.1	Organization.....	22
3.2	Mobile Station Scheduling Procedure.....	25
3.3	Base Station Scheduling Procedure.....	28
3.3.1	Case 1: Total Number of all Requests for Transmission in the Next Frame is Less than, or Equal to the Number of Time Slots in a Frame (i.e. 15).....	28

3.3.2	Case 2: Total Number of all Requests for Transmission in the Next Frame is Greater than the Number of Time Slots in a Frame (i.e. 15).....	29
3.4	Quality of Service Constant (QSC).....	31
3.5	Modeling of the Cell.....	34
3.5.1	Organization.....	34
3.5.2	OPNET Network Model.....	35
3.5.3	MS Physical Layout Within the Cell.....	36
3.5.4	MS Data Rates.....	37
3.5.5	Measure of Performance (Cell Throughput vs. MS Delay Relationship).....	37
4.	Simulation Results.....	39
4.1	Organization.....	39
4.2	MS Delay vs. Cumulative Offered Traffic Relationship.....	39
4.3	Definition of Instability.....	40
4.4	Causes of Instability.....	40
4.5	Cell Throughput vs. MS Delay.....	41
4.5.1	Effect of Distance Between BS and MS on the Cell Throughput vs. MS Delay.....	41
4.5.2	Effect of Increasing MS Data Rate on the Cell Throughput vs. MS Delay Curve.....	44
4.5.3	Maximum Sustainable Data Rate.....	44
5.	Simulation Validation and Capacity Calculation.....	46

5.1	ITU Source A Validation- Modeling of the System (Given that Only ITU Source A are Active in the Cell).....	46
5.1.1	Maximum Number of Unserved Requests.....	49
5.2	Modeling of the System (When a Mix of ITU Source A and B are Active in the Cell).....	50
5.3	Simulation Validation Using Confidence Intervals.....	54
5.4	Capacity Calculation.....	58
5.4.1	ITU Source B Capacity Calculation.....	60
5.4.2	ITU Source A Capacity Calculation.....	67
5.4.3	Capacity Calculation for a Combination of ITU Source A and B.....	70
5.5	Probability of Recovery.....	71
6.	Advantages of This Scheduling Technique & Conclusion.....	78
6.1	More Efficient Capacity Packing.....	78
6.2	Fairness.....	78
6.2.1	“1 flag” Request.....	78
6.2.2	“2 flag” Request.....	79
6.3	Minimum Service Rate Guarantee.....	79
6.4	Conclusion.....	80
6.5	Future Work.....	81

List of Abbreviations

BSC	Base Station Controller
CN	Core Network
DRNS	Distributed Radio Network Sybsystem
MSC	Mobile Switch Center
GGSN	Gateway GPRS Support Node
GMSC	Gateway Mobile Switch Center
PLTN	Public Land Telephone Network
RNC	Radio Network Controller
RNS	Radio Network Subsystem
SGSN	Serving GPRS Support Node
SMSC	Serving Mobile Switch Center
SRNC	Serving Radio Network Controller
SRNS	Serving Radio Network Subsystem
UE	User Equipment
UMTS	Universal Mobile Telecommunication System
UTRAN	Universal Terrestrial Radio Access Network
WWW	World Wide Web

List of Figures

Figure 1-1 UMTS Architecture.....	2
Figure 1.1-1 UTRAN Architecture.....	4
Figure 1.2-1 a.) Typical CDMA system.....	6
Figure 1.2-1 b.) Signals at each stage in a typical CDMA system (All the graphs are in frequency domain).....	6
Figure 1.2-2 Details of the Used MCS.....	7
Figure 1.2-3 Mobile Station supporting three simultaneous sessions.....	8
Figure 1.3-1 WCDMA MAC Layer Frame Structure.....	10
Figure 2.1.1-1. ITU Data traffic model.....	18
Figure 2.1.2-1. Cumulative uplink data traffic pdf [3].....	18
Figure 3.1-1. Mobile Station Behavior.....	24
Figure 3.2-1 BS Scheduling Procedure.....	27
Figure 3.3.2-1 Example of a Case 1 scheduling technique.....	30
Figure 3.3.2-2 Example of a Case 2 scheduling technique.....	30
Figure 3.5.2-1. Simulation Network Model.....	36
Figure 3.5.3-1. Cell Layout.....	37
Figure 4.5.1-1. Cell Throughput vs. MS Delay (ITU Source A, as it moves away from the BS).....	42
Figure 4.5.1-2. Cell Throughput vs. MS Delay (ITU Source C, as it moves away from the BS).....	42
Figure 4.5.1-3. Cell Throughput vs. MS Delay (ITU Source A).....	43
Figure 4.5.3-1. System Limits.....	45
Figure 5.1-1. M/D/1 Approximation of the Cell Behavior When only ITU Source A's are Active.....	47

Figure 5.1-2 M/D/1 Approximation of system behavior when only ITU Source As are active, distance between BS and MS = 10 m.....	47
Figure 5.1-3. M/D/1 Approximation of system behavior when only ITU Source As are active, distance between BS and MS = 1610 m.....	48
Figure 5.1.1-1. Number of Unserved Requests vs. Cell Loading.....	50
Figure 5.2-1. M/D/1 Approximation to the Cell Behavior When Both ITU Source's are Active.....	52
Figure 5.2-2. ITU Source A cell Loading vs. Expected Waiting Time (Each line represents different number of simultaneous ITU Source B packet calls, ranging from 0 to 15).....	53
Figure 5.2-3. ITU Source A cell Loading vs. Expected Total Normalized Waiting Time (Each line represents different number of simultaneous ITU Source B packet calls, ranging from 0 to 15).....	54
Figure 5.3-1 Confidence Intervals for the Case where the Distance Between the BS and MS is 1610m.....	55
Figure 5.3-2 90% Confidence Interval for the Case where the Distance Between the BS and MS is 1610m.....	55
Figure 5.3-3 95% Confidence Interval for the Case where the Distance Between the BS and MS is 1610m.....	56
Figure 5.3-4 99% Confidence Interval for the Case where the Distance Between the BS and MS is 1610m.....	56
Figure 5.3-5 99% Confidence Interval for the Case where the Distance Between the BS and MS is 10m.....	57
Figure 5.4.1-1. Calculation of Overlapping ITU Source B Packet calls.....	62
Figure 5.4.1-2. ITU Source B Capacity (For various Packet Call Lengths).....	67
Figure 5.4.2-2. ITU Source A Capacity	69
Figure 5.5- 1 System Finite State Machine.....	71
Figure 5.5- 2 Probability of Recovery When Only ITU Source B is Active.....	75
Figure 5.5- 3 Probability of Recovery When Only ITU Source A is Active.....	84

Chapter 1

Introduction

Today's cellular telephone systems are mainly circuit switched because they were primarily designed for voice transmission. This makes them very unsuitable for data transmission purposes. The next, third generation (3G) of cellular systems, the UMTS (Universal Mobile Telecommunications System), will have the broadband, packet based text, video, and multimedia capabilities, at data rates of up to and possibly higher than 2 Mbps. This is in addition to transmission of voice, which is already provided by the existing cellular systems. Once the UMTS is fully implemented the computer and phone users will have the capability to be connected to the internet at all times.

In order to save resources and minimize cost, the existing cellular systems (GSM/GPRS) will be incorporated into the overall UMTS structure. Within the UMTS, transmission of voice will be accomplished using circuit switched techniques, which are also used in current cellular systems. Transmission of data however, will be accomplished using separate packet switched core networks, which are already a part of the existing GPRS systems, as shown in figure 1-1. UMTS therefore uses the existing core network, which is currently part of GPRS, however it uses an entirely new radio interface. The new radio network in UMTS is called UTRAN (UMTS Terrestrial Radio Access Network), and is connected to the core network of GPRS as shown in figure 1-1.

Figure 1-1 UMTS Architecture

1.1 UTRAN Architecture

UTRAN is organized in a similar fashion as the older (GSM/GPRS) cellular systems. Within the UTRAN all of the mobile stations in one cell communicate with the base station in that cell, as is the case in current cellular systems. A number of base stations are controlled, and communicate with one Node B. Number of Node Bs are in turn controlled, and communicate with one RNC (Radio Network Controller), as shown in figure 1.1-1. A system of consisting of one RNC which controls multiple Node Bs, which in turn control multiple base stations is called the RNS (Radio Network Subsystem).

UTRAN architecture allows one mobile station to simultaneously communicate with more than one base station which may belong to more than one Node B which in turn may belong to more than one RNC, as shown in the UTRAN architecture figure. This is an important functionality, which influences the performance of any resource allocation technique, and will be discussed in greater detail later on in this thesis.

Within the UTRAN, WCDMA (Wideband Code Division Multiple Access) is the most widely adapted air interface standard. The 5 MHz bandwidth of WCDMA supports high data rates, and brings additional performance benefits such as increased multi-path diversity, which is used to minimize the effects of signal fading. In addition to other functionalities, WCDMA based systems are designed to support variable bit rates, quality of service requirements, multiplexing of services on a single connection, and location based services.

Figure 1.1-1 UTRAN Architecture

1.2 WCDMA Physical Layer Background

As in any spread spectrum system, in the WCDMA transmitter, the message signal is multiplied by one of 512 WCDMA spreading codes unique to that particular transmission. Meaning that other transmissions in the same sector use different spreading codes.

A spreading code is a sequence of chips (square pulses) valued -1 and 1 , where the number of chips in one code is the period of that code. The main characteristic of these codes is that they are all orthonormal with respect to each other. Meaning that, the integral of the product of two different spreading codes is equal to zero, and the integral of the product of two identical spreading codes is equal to one.

In the receiver, the received signal is again multiplied by the same synchronized spreading code as shown in figure 1.2-1. This multiplication removes the spreading code, and results in an original message signal. Because the spreading codes are orthonormal, and the message signal is multiplied by the same spreading code twice (once at the transmitter, and once at the receiver), the power of the transmitted signal is not altered, as shown in figure 1.2-1. Rather, only the power spectral density of the signal is changed while the signal is being transmitted. As the signal is transmitted from the mobile station to the base station, and vice versa, its power is attenuated. Strict power control algorithms are followed by the base and mobile station, so that the signal received at the base station has sufficient power required for correct reception, but does not waste unnecessary power.

Figure 1.2-1 a.) Typical CDMA system

Figure 1.2-1 b.) Signals at each stage in a typical CDMA system (All the graphs are in frequency domain)

Depending on the magnitude of the received C/I (carrier to interference ratio), as shown in figure 1.2-1, it is possible to modulate the message signal using various MCS (Modulation Coding Schemes). Different MCS offer different protection from errors caused by channel conditions, at the cost of data rate. MCS that provide higher data rate (lower frame error rate) require a larger received C/I (at the base station) than the MCS that grant lower data rates (as shown in figure 1.2-2). Therefore, it is possible to optimize the data rate by selecting a modulation and coding scheme, which results in a maximum data rate for each particular C/I magnitude level. In WCDMA standard only seven MCS are being used. They are shown in figure 1.2-2.

AMCS	Code rate	Modulation size	Data rate (kbps)
MCS 7	$\frac{3}{4}$	64 QAM 6 information bits/symbol	540
MCS 6	$\frac{3}{4}$	16 QAM 4 information bits/symbol	360
MCS 5	$\frac{3}{4}$	8 PSK 3 information bits/symbol	270
MCS 4	$\frac{1}{2}$	16 QAM 4 information bits/symbol	240
MCS 3	$\frac{3}{4}$	4 PSK 2 information bits/symbol	180
MCS 2	$\frac{1}{2}$	4 PSK 2 information bits/symbol	120
MCS 1	$\frac{1}{4}$	4 PSK 2 information bits/symbol	60

Figure 1.2-2 AMCS Details

The mobile station cannot change the attenuation its signal experiences because it depends on the channel conditions. Similarly, the mobile station cannot change the interference it experiences because the interference is caused by other sources. Therefore, the only way in which the mobile station can increase, or decrease, the C/I that is received at the base station is to increase, or decrease, its transmitted power (C).

The WCDMA standard specifies that a single mobile station has the ability to support more than one simultaneous session (channel), provided that the C/I the mobile

station experiences at the base station is large enough (the mobile station has enough transmitting power). In this thesis the maximum of 20 simultaneous sessions was selected because most of the industry made the same choice.

If the mobile station is indeed transmitting more than one channel, then each session uses a different WCDMA spreading code, and is therefore a completely separate channel from the rest of the sessions, as shown in figure 1.2-3. In this case all channels experience their own, separate C/I, and a MCS can be chosen optimally, as described above, for each channel separately.

Figure 1.2-3 Mobile Station supporting three simultaneous sessions

Furthermore it is possible to use AMCS (Adaptive Modulation and Coding Scheme) [4], which was used in this thesis, to optimize the cumulative data rate that the mobile station achieves. The AMCS selects an optimum number of simultaneous sessions the mobile station should use, and an optimum MCS for each session, for a given C/I the mobile station experiences, in such a way that the cumulative throughput of all of the channels is maximized [4]. Each MCS requires different magnitude of C/I received at the base station, which translates into different amount of transmitted power at the mobile station. Furthermore, the mobile station has a fixed maximum transmitted power of 200

mW [4]. The AMCS therefore assigns this power optimally to all of the simultaneous sessions in such a way that the maximum data rate is achieved.

1.3 WCDMA MAC Layer Background

WCDMA MAC (Medium Access Control) layer is responsible for arrangement of up-link and down-link transmissions between the mobile station and the base station. The up-link represents the transmissions originating from the mobile stations, which are destined for the base station, and the down-link represents the transmissions originating at the base station which are destined to be received by the mobile station. In this thesis only the up-link transmissions are studied.

Furthermore, the transmission of voice in WCDMA will be accomplished, separately from transmission of data. It will be done using circuit switched techniques, as it is currently the case with GSM/GPRS. For the purpose of this thesis it is assumed that voice transmission will take 20 % of the mobile station's power, and the remaining 80 % is available for the transmission of data. From this point on only the data transmissions are discussed.

WCDMA MAC layer frame is shown in figure 1.3-1. The frame is 10 ms long, and consists of 15 time slots, each of which is 0.667 ms long. Each frame is part of a super frame, which is 720 ms long, and consists of 72 time frames, as shown in figure 1.3-1.

Figure 1.3-1 WCDMA MAC Layer Frame Structure

The WCDMA MAC layer has a number of channels used to transmit information. These channels can be categorized into the ones used to transmit control information, and the ones used to transmit data. The data can be transmitted on either the DCH (dedicated channel), or the USCH (Up-link shared channel).

In the case that the dedicated channel is used to transmit data, more than one mobile stations are able to transmit at the same time, each using their own dedicated channel (their own unique WCDMA spreading code). As stated earlier, and illustrated in figure 1.2-1, each mobile station's transmission is interference to the other mobile stations. The mobile stations however, require some minimum C/I for a successful transmission. In order to provide this minimum C/I to each mobile station, each mobile station's transmitted power must be limited. Therefore, when the dedicated channel is used to transmit data, all of the mobile stations are able to transmit at the same time. However, the power of their transmissions, and therefore their data rates, are limited and shared between them.

In the case that the up-link shared channel is used to transmit data, only one mobile station is allowed to transmit at one time. However, because there is no other

mobile stations transmitting, which means that there is no interference, the mobile station in question is allowed to transmit at full power for the duration of its transmission. In this case the time duration of the transmission is shared between all of the users.

This thesis studies only the USCH method for transmitting data because the USCH can be further optimized, using AMCS, to provide higher data rates than the DCH method [2][4].

1.4 *MAC Layer Transmission of Data Over Air Interface*

Data can be transmitted over the air interface by a number of methods, all of which can generally be classified into three categories; circuit switched, random access, and scheduled.

Circuit switched technique is the one in which a virtual circuit is established between the base station and the mobile station for the duration of the data transmission. Circuit switched techniques therefore result in an initial circuit establishment delay (call setup, or system access delay), and once the circuit is established the media access delay is zero. The downside of the circuit switched techniques is that they are inefficient for the bursty data sources such as the mobile stations data sources discussed in chapter 2 of this thesis. Furthermore, because large proportion of data sources generate data for a relatively short time period (chapter 2) when compared to the time it takes for call setup, the delay the mobile stations experience (system access delay) will actually be large.

Random access techniques are opposite of circuit switched techniques. They are the techniques such as ALOHA, slotted ALOHA, CSMA/CD, and others in which the mobile station attempts to transmit a packet as soon as it is generated. Generally, random

access techniques result in minimal system access delay when offered load is relatively small when compared to maximum system throughput. However, when offered load increases to large percentage of maximum system throughput the random access techniques start exhibiting large delay and small throughput due to an increasing number of collisions and corresponding retransmission attempts.

Scheduled techniques are a compromise between circuit switched, and random access techniques. By scheduling each data transmission the scheduling techniques remove any possibility of an occurrence of a collision. This improves the maximum attainable throughput, and minimizes system access delay, when compared to the random access techniques. This is because the collisions are a main cause of increased delay and decreased throughput for random access techniques. Furthermore, scheduled techniques have substantially lower system access delay when compared to the circuit switched techniques. In addition they are more efficient (greater maximum throughput) than the circuit switched techniques, because in scheduled techniques mobile station generally transmits only when it has data to transmit, wherein circuit switched techniques it is possible for the circuit to be active while the source has no data to transmit, resulting in the waste of resources.

Generally, in up-link scheduling techniques, the base station receives mobile stations' requests for transmission for a certain time window, after which it schedules those requests. This procedure repeats its self every time window, which is in fact the scheduling period. Therefore, scheduling techniques can be further classified into long term scheduling, and short term scheduling techniques, based on the size of the scheduling period. Scheduling techniques with a "large" scheduling period are long term

scheduling techniques, while those with “small” scheduling period are short term scheduling techniques.

1.5 WCDMA Up-Link MAC Layer Radio Access Problem

UTRAN radio resources are expensive. Therefore, they should be utilized to their maximum potential. In order to achieve this, the capacity of the UTRAN must be maximized while maintaining some level of QoS (Quality of Service). Maximization of the system capacity is achieved by the maximization of the system throughput. The QoS is provided by the minimization of delay the mobile stations experience, as well as by providing the mobile stations with the ability to control their delay.

This thesis proposes a solution to the above problem in the form of DTDFP (Dynamic Time Domain Fat Pipe) resource allocation technique. The DTDFP resource allocation technique belongs to the short term scheduled class of system access techniques. It uses the USCH transmission method, which allows it to maximize the throughput of the up-link physical channel by using the AMCS [2][4].

This thesis shows that the DTDFP resource allocation technique ensures fairness, maximizes the system throughput, and at the same time provides a level of QoS, as measured by control over delay. This thesis shows that the DTDFP resource allocation scheme achieves these results, which are superior to those that can be achieved by random access and circuit switched techniques, at a significantly lower implementation complexity and signaling overhead cost than any long-term optimization resource allocation technique. Furthermore, it shows that the DTDFP resource allocation technique guarantees the stability of the system at all times, and for all possible combinations of

uplink data traffic. In addition this thesis calculates the maximum capacity of the WCDMA system, which utilizes a DTDFP resource allocation scheme, in terms of maximum number of simultaneous active sessions in one cell, and in terms of the maximum mobile station sustainable data rate at each cumulative cell offered traffic level.

The operation of the WCDMA system, which utilizes a DTDFP resource allocation scheme, was simulated using OPNET, in order to analyze its performance under various and realistic cell traffic loading, log-normal fading, and distance from base station conditions, and the results were validated using analytical techniques.

1.6 Organization

The remainder of this thesis is organized in four chapters.

- o Chapter 2 describes the individual and cumulative up-link data traffic models for the 3G cellular systems. It discusses the long term and short-term up-link scheduling techniques. Specifically it discusses the performance of those scheduling techniques when they are exposed to the expected data patterns of the 3G cellular systems.
- o Chapter 3 describes in detail the proposed scheduling technique, as well as the model which was used to simulate its operation.
- o Chapter 4 presents the results of the simulation of the DTDFP resource allocation scheme. It also discusses the meaning of those results, and their significance.
- o Chapter 5 presents the results of an analytical validation of the simulation. Specifically it models an operation of one cell which uses DTDFP resource

allocation technique with a modified M/D/1 queue system, and it explains the similarities and differences between the simulated model and an analytical model. In addition, chapter 5 calculates the capacity of one cell of a system that uses the DTDFP resource allocation scheme in terms of maximum number of simultaneous active sessions in one cell.

- o Chapter 6 concludes by reviewing the presented material, specifically the results of the simulation and analytical analysis. Finally the recommendations for further related research work is given at the end of the chapter.

Chapter 2

Up-Link Traffic Models and Scheduling Issues

The performance of any resource allocation technique very much depends on the type of traffic that the scheduling technique will be exposed to. This chapter first introduces the 3G up-link data traffic model. Then it discusses the performance of long term scheduling techniques, versus that of short term scheduling techniques, when they are exposed to the 3G up-link data traffic model.

2.1 Up-Link Data Traffic Models

Although uplink 3G WCDMA data traffic model is still unknown, for the purpose of testing the DTDFP resource allocation scheme this research paper uses the ITU 3G Data traffic predictions.

The ITU has created two possible 3G data traffic models, ITU Source A to simulate simple interactive “telnet” session, and Short Message System (SMS), and ITU Source B to simulate FTP and large file transfers.

In this thesis Source C has been added to the already existing ITU traffic model in order to test the DTDFP resource allocation scheme under extreme conditions, where one source requires significantly higher data rates than the rest of the sources.

2.1.1 ITU Data Traffic Models

ITU Class A Traffic Source

- Poisson arrivals of sessions
- Geometric number of packet calls per session (114)
- Geometric packet call inter-arrival times (1 second)
- One packet per packet call
- Geometric packet size (720 bits)

ITU Class B Traffic Source

- Poisson arrivals of sessions
- Geometric number of packet calls per session (5)
- Geometric packet call inter-arrival times (120 seconds)
- Pareto distributed number of packets per packet call
- Geometric packet inter-arrival times (0.01 second)
- Constant packet size (3840 bits)

Class C Traffic Source

- Uses same model as that of ITU Class A Traffic source
- Parameters of the model are changed so that the source transmits very large amounts of data very rarely, anywhere from 2 Mbps to 10 Mbps depending on the simulation run.

Figure 2.1.1-1. ITU Data traffic model

2.1.2 Cumulative Up-Link Data Traffic Patterns

Assuming that a cell is serving only ITU Traffic Sources (Both A and B), distributed in any proportion, the PDF for packet length request in bits is shown by figure 2.1.2-1. [3].

Figure 2.1.2-1. Cumulative uplink data traffic pdf [5]

Figure 2.1.2-1 shows that the cumulative traffic size PDF resembles exponential distribution function. It shows that each individual MS will transmit a small packet with large probability, and a large packet with a small probability.

2.2 Long-Term Scheduling Resource Allocation

In long-term scheduling techniques the entity responsible for the scheduling task completion collects all of the resource requirements of all of the mobile stations (for “longer” period of time), runs some optimization procedure and signals to the MS the amount of resources the MS has been granted. The inputs to the optimization could either be future data traffic requirements of all of the MS, or current data traffic requirements of all of the mobile stations.

In the case that the input to the optimization are the future data traffic requirements of all of the mobile stations, the individual mobile stations’ future data requirements are unknown even by the MS themselves, and are estimates at best. Therefore, errors in future data traffic requirements estimation by the MS become input errors in the optimization, which would definitely not produce an optimum result.

In the case that the input to the optimization are current data traffic requirements of all of the mobile stations, the mobile stations must collect the data and store it in its buffer for the duration of the scheduling period. As can be seen from the 3G up-link data traffic model above, most of the time the mobile stations will require to transmit a small packet. The 3G wireless systems data rates can be as high as 2 Mbps, which means that small packets can be transmitted rather quickly. If the time required for packet

transmission is significantly less than the scheduling period, which is likely to occur because of 3G cellular system's high data rate, the medium access delay is increased, even for the case where the effects of congestion are insignificant. If the scheduling period is large enough the performance of this scheduling scheme approaches that of a circuit switched data transmission method, which is definitely not optimal for the transmission of data.

Furthermore, for both of the above cases the signaling overhead is significantly larger than for the short term resource allocation method, which will be discussed later. This is because in the cases above the mobile station has to, not only signal to the base station that it has data to transmit, but it also must signal the quantity of data it needs to transmit. In any case it is necessary to minimize the amount of signaling because it is an overhead, which affects the throughput in a negative way.

2.3 *Short Term Scheduling*

Short term scheduling techniques are better suited for the 3G cellular systems' data traffic patterns, discussed above, than the long term scheduling techniques. Because of the fact, that most of the time the mobile stations will require to transmit a small packet which does not need large amounts of transmission time, due to high data rates, fast allocation and de-allocation of resources functionality of the short term scheduling techniques decreases the media access delay.

Furthermore, as the scheduling period is decreased even further, the media access delay is also decreased. However, if the scheduling period is decreased significantly, the amount of signaling overhead would increase, which would cause the system throughput

to decrease. As the scheduling period is decreased, the mobile stations' transmission time, and therefore the number of bits transmitted is also decreased to the point that the signaling overhead becomes a significant percentage of each transmission, which decreases the system throughput. Therefore, there exists some optimum range for the scheduling period, which would maximize the system throughput, and minimize the system access delay.

Furthermore, short term resource allocation schemes' dynamic fast resource allocation and de-allocation, on a frame by frame basis, functionality is necessary in order to exploit the advantages of a fat shared pipe [2].

[2] "There is a well known traffic theory result which states that it is more efficient to have one high rate server (single fat data pipe or more rigorously $M/M/1$) as opposed to multiple low rate servers (thin pipes or $M/M/n$). The RRC/DCH [RRC-Radio Resource Control, located within the RNC] (long-term scheduling) approach has NOT been designed with a fast resource allocation in mind therefore implicitly assumes a multiple thin pipe (circuit like) scenario," which by its very nature is not optimum. In this thesis the fat shared pipe was implemented using the AMCS as discussed earlier.

Chapter 3

Proposed Scheduling Technique

The DTDFP (Dynamic Time Domain Fat Pipe) resource allocation scheme is a short-term resource allocation method, in a sense that it uses short-term resource leases (few time slots on one frame) to schedule the data traffic. Furthermore it is a fat pipe resource allocation scheme since only one MS is allowed to transmit in one cell at one time, at full power that is available to it, and only for a short duration.

The DTDFP resource allocation technique uses the fact that, each individual MS will transmit a small packet with large probability and a large packet with a small probability (figure 2.1.2-1). To optimize the scheduling procedure, the DTDFP scheduling technique divides the mobile stations into two types: small user, and a large user, where it is possible for the mobile stations to change from small users to large users, and vice versa.

3.1 Organization

The following sections will describe in detail the MS and the BS scheduling procedure. These descriptions are supplemented by the figures 3.1-1 and 3.2-1, which depict the scheduling procedures using flow charts. Examples of specific scheduling cases are shown in figures 3.3.2-1, and 3.3.2-2. Section 3.4 introduces, and describes in detail the QSC (Quality of Service Constant) used by the MS to characterize its delay

tolerance threshold. Section 3.5 introduces the OPNET model that was used to simulate the operation of one WCDMA cell.

Figure 3.1-1. Mobile Station Behavior

3.2 Mobile Station Scheduling Procedure

- If the Mobile Station had been silent up to this time because it had no data to transmit.
 - Mobile Station will signal to the BS (Base Station) on the RACH that it has data to transmit, by sending either a “1 flag” (request for one time slot on one frame), or a “2 flag” (request for multiple time slots on one frame), depending on how much data it has to transmit.
- If the Mobile Station has transmitted in the current frame, and has some additional data to transmit.
 - At the end of its transmission the MS will indicate with a two bit binary flag whether it, requires another transmission for the duration of one time slot (“1 flag”), or requires another transmission for the duration of more than one time slot (“2 flag”). The absence of the flag would indicate to the BS that MS has no more data to transmit at this time.
- Mobile Station transmits the data at the time the BS signals it to start transmission.
 - If the MS requested only one time slot it transmits for the duration of only one time slot starting at the time specified by the BS.
 - If the MS requested more than one time slot.

- If the request for more than one time slot is granted by the BS the MS starts its transmission at the time specified by the BS, and transmits until the end of the current 10 ms frame.
- If the BS decided to grant a permission to the MS to transmit on only one time slot in the next frame (as opposed to multiple time slots in next frame);, then BS transmits a “flag 1” signaling to the MS that its only permitted to transmit for one time slot in current frame. The MS starts the transmission at the time specified by the BS, and only for the duration of one time slot.

Figure 3.2-1 BS Scheduling Procedure

3.3 Base Station Scheduling Procedure

In a 10 ms time window BS collects all the requests for transmissions from all of the MS, both from the RACH, as well as those at the end of current transmissions. The BS then schedules the transmissions corresponding to those requests for the next frame. Any transmission requests that have not been granted are automatically transferred as requests for transmissions in the following frame.

3.3.1 Case 1: Total number of all requests for transmission in the next frame is less than, or equal to the number of time slots in a frame (i.e. 15).

1. Allow the first “2 flag” request, if there is any, to remain a “2 flag” request.

Change all subsequent “2 flag” requests, if there is any, to “1 flag” requests, and schedule them as such. If two or more “2 flag” requests are received at exactly the same time, choose the first one randomly (i.e. if two are received at the same time flip a coin to see which one is first).

2. Schedule all of the transmissions corresponding to “1 flag” requests (including those which were changed to “1 flag” request from a “2 flag” request) on a first come first serve basis, starting at the head of the 10 ms frame. Send the corresponding scheduled transmission times to all of the MS. Along with the scheduled transmission time, send a “1 flag” to all of the MS that requested a “2 flag,” but were only granted permission to transmit for the duration of only one time slot.

3. The transmission corresponding to a “2 flag” request is scheduled for the remainder of the 10 ms frame. Send the scheduled transmission time to this MS.

3.3.2 Case 2: Total number of all requests for transmission in the next frame is greater than the number of time slots in a frame (i.e. 15)

1. Arrange all of the requests for transmission (both “1 flag” and “2 flag”) in the order in which they arrived at the BS. From the first 15 requests (15 time slots in one frame) change any “2 flag” request for transmission to a “1 flag” request for transmission.
2. Schedule all of the transmissions corresponding to first 15 requests in a first come first serve basis, each for duration of one time slot. Along with the scheduled transmission time send a “1 flag” to all of the MS that requested a “2 flag” transmission, but were scheduled for a transmission of duration of only one time slot. This is done to let the MS, which has requested permission to transmit for multiple time slots, know that it has only received permission to transmit for one time slot.
3. All of the requests for transmission that were not served in the current frame are automatically transferred to requests for transmission in the next frame (unchanged i.e. “1 flag” request remains a “1 flag” request, and a “2 flag” request remains a “2 flag” request).

Case 1

Time frame	Request for Transmission (in order of arrival)	Scheduled for Transmission in the Following Frame
Frame 0	A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P	A,B,C,D,E,F,G,H,I,J,K,L,M,N,O
Frame 1	P,A,B,C,D,E,F,G,H,I,J,K,L,M,N,O	P,A,B,C,D,E,F,G,H,I,J,K,L,M,N
Frame 2	P,O,A,B,C,D,E,F,G,H,I,J,K,L,M,N	O,P,A,B,C,D,E,F,G,H,I,J,K,L,M
Frame 3	N,O,P,A,B,C,D,E,F,G,H,I,J,K,L,M	N,O,P,A,B,C,D,E,F,G,H,I,J,K,L

frame 1						frame 2						frame 3					
A	B	C	...	N	O	P	A	B	...	C	N	O	P	A	...	B	C

Figure 3.3.2-1 Example of a Case 1 scheduling technique.

Case 2

Time Frame	Request for Transmission (In order of Arrival)	"1 flag" request	"2 flag" request	One time slot scheduled transmissions	Multiple time slot scheduled transmissions
Frame 0	A,C,D,B,E	A,B	C,D,E	A,D,B,E	C
Frame 1	A,D,B,E,C	A,B	D,E,C	A,B,E,C	D
Frame 2	A,B,E,C,D	A,B	E,C,D	A,B,C,D	E
Frame 3	A,B,C,D,E	A,B	C,D,E	A,B,D,E	C

Frame 1					frame 2					frame 3				
A	D	B	E	C	A	B	E	C	D	A	B	C	D	E

Figure 3.3.2-2 Example of a Case 2 scheduling technique.

3.4 *Quality of Service Constant (QSC)*

Whenever the MS has data to transmit it has the opportunity to request permission to transmit on one time slot in one frame (“1 flag”), or on multiple time slots on one frame (“2 flag”). It is assumed that most of the time majority of the MS will make “1 flag” request. This is because of the inherent nature of the Traffic PDF discussed earlier, and shown in figure 2.1.2-1. That is, most of the time the majority of the MS will have small amounts of data to transmit, which can be accommodated by one time slot on one frame, and very rarely will there be only a few MS that have large amounts of data to transmit. It is this kind of MS that the “2 flag” service request is meant for.

Depending on the location of the MS in the cell, the MS’s rate of transmission can vary from 0.8 Mbps to 10.3 Mbps due to the [4] C/I distribution throughout the cell. By keeping a moving average of its own transmission speed the MS can predict the amount of delay the last bit in its queue will experience (i.e. the maximum delay).

$$Delay(sec) = \frac{Queue\ Size\ (bits)}{Avg.Transmission\ Rate\ (bits/sec)}$$

Average transmission rate is needed, rather than the instantaneous transmission rate, due to the fact that the transmission rate can vary greatly because of scheduling reasons, and instantaneous congestion. For example, in the case of heavy instantaneous congestion (a lot of requests for service in a cell, only for the duration of one time frame) the MS can be denied permission for transmission in a current frame. However, it will surely be allowed to transmit in the next frame. If the instantaneous transmission rate is

used to approximate future transmission rate in this case, the approximated transmission rate would be 0 bps, which is clearly wrong. The average transmission rate is therefore a better approximation of the future transmission rates than the instantaneous transmission rate.

$$\text{Avg. Transmission Rate} = \frac{\text{Num. of Bits Transmitted in Moving Average Window}}{\text{Moving Average Window Size (num of frames)} \times 0.01 \text{ sec.}}$$

Average transmission rate can be calculated using a moving average window of the transmission rates of the last few frames.

In this simulation the ideal size of the moving average window was chosen to be 10 time frames, given that the MS has been active (made requests for transmission) in those 10 frames.

The reason for the choice of the moving size window of 10 frames is mobility. That is, it is important for the moving average to represent current transmission rate of the MS. As was already mentioned earlier the transmission rate of a MS can vary drastically depending on the physical position of that MS in the cell. Therefore if the MS moves at a rate of 100 km/h within the cell (27.8 m/s), and the moving average window of 10 time frames is used, then the transmission rate will be calculated over 0.1 ms (10 x 10 ms (one frame size is 10 ms) = 0.1 sec). In 0.1 sec, the MS will move a distance of $27.8 \times 0.1 = 2.8$ meters. From the C/I distribution discussed earlier, it can be seen that a change of 2.8 m in the distance between MS and BS does not affect the C/I significantly, and therefore the transmission rate will not be changed significantly. If the MS has not been active in last 10 consecutive frames, then the size of the moving average window should be the number of consecutive frames the MS has been active in. This means that

there is a special case in which the moving average window cannot be calculated. This happens when the MS has been silent previously (had no data to transmit), and has just received data to transmit. This is acceptable because the MS's queue has been empty, therefore MS would make a "1 flag" request either way, so that it is not impacted with a moving window size of 0.

$$\begin{aligned}
 x &= \text{Number of consecutive frames in which MS was active} \\
 W &= \text{Moving Average Window Size} \\
 \text{if } x \geq 10 &\Leftrightarrow W = 10 \\
 \text{if } x < 10 &\Leftrightarrow W = x
 \end{aligned}$$

Having established that the MS can predict its immediate future transmission rate using its recent transmissions rates, then the MS can also predict the delay that the last bit in its queue will experience.

Depending on the application the MS is running, and how sensitive that application is to the delay, the application the MS is running can determine the maximum acceptable delay. The maximum acceptable delay can be translated into MS's maximum acceptable queue size, which is in fact the QSC.

$$QSC = \text{Average Transmission rate} \times \text{Maximum Acceptable delay}$$

The decision the MS makes, as to which request for transmission to make ("1 flag," or "2 flag" request), depends on whether or not the MS's queue size is greater than the QSC.

$$\begin{aligned}
 \text{if } QSC > \text{queue size} &\Leftrightarrow \text{"1 flag"} \\
 \text{if } QSC \leq \text{queue size} &\Leftrightarrow \text{"2 flag"}
 \end{aligned}$$

Therefore, a MS will not make a request for transmission on multiple time slots in one frame unless the delay that it experiences threatens the execution of the application it is currently running. Therefore applications such as e-mail and web browsing, which are not very time sensitive, will have larger QSC than an application such as real-time video uploading, which is very time sensitive.

This is crucial for the success of this scheduling technique, because only one MS can transmit for multiple time slots on one frame during one frame. If the QSC constant was very low for all of the MS, then the MS which could tolerate greater delay would compete for the same resources (multiple time slot transmissions on one frame) with the MS that is at the edge of its delay tolerance. This would result in more delay for the MS that is delay sensitive, and less delay for the MS that is not delay sensitive. Which is clearly not an optimum outcome.

3.5 *Modeling of the Cell*

The operation of DTDFP resource allocation scheme in one sector of one cell was modeled using OPNET. For the purpose of modeling it was assumed that there was negligible inter cell interference, and that the cell, which was modeled, was not affected in any way by any of the surrounding cells.

3.5.1 *Organization*

Details of the simulation model are displayed in Appendix B. This section briefly describes the OPNET network model used to simulate the cell. It introduces the

parameters, such as, distance between BS and MS, and MS data rates, whose affect on the performance of the DTDFP resource allocation scheme was studied. Finally it introduces the MS Delay vs. Cell Throughput relationship as the sole measure of the performance of the DTDFP resource allocation technique, and the reasoning behind this choice.

3.5.2 OPNET Network Model

The OPNET network model is shown in the figure 3.5.2-1. In figure 3.5.2-1 each MS (Mobile Station) consists of a mobile source (to generate the data traffic), and mobile queue-server (to transmit the data traffic).

In the simulation Aggregate Source was used to simulate the entire traffic within the cell (except the MS which was observed). Therefore the sole purpose of the Aggregate Source was to create congestion within the cell so that the observed mobile stations performance can be monitored under specific congestion (cumulative offered traffic) conditions. This was done to simplify the model design, as well as to make the cell loading parameter easier to control. This way the cell could be loaded precisely to the exact requirement of the simulation. The aggregate cell traffic could also have been simulated with many ITU source A, and B's, but that would make the simulation more complex, harder to manage, and very little, or no extra information would be gained.

Figure 3.5.2-1. Simulation Network Model

3.5.3 MS Physical Layout Within the Cell

The mobile stations positioned in different distances from the BS experience [4] different C/I distribution values. Because the C/I value is directly proportional to the maximum data rate a MS, which utilizes AMCS (section 1.2) could achieve, the distance from the BS is an important element of a realistic simulation.

The simulation was performed with all of the mobiles, except the aggregate queue, (i.e. ITU Source A, B, and Large Source) positioned in various combinations, in a number of possible positions within the cell, depending on the requirements of each individual run, as shown in figure 3.5.3-1.

Figure 3.5.3-1. Cell Layout

3.5.4 MS's Data Rates

Every time the MS is about to begin transmission it obtains a random C/I ratio from an appropriate C/I distribution based on its distance from the BS. Based on the obtained C/I value the MS data rate is then calculated according to the table in Appendix A. The data rate for all of the MS (Source A, B, and Large, queues) was calculated that way.

3.5.5 Measure of Performance (Cell Throughput vs. MS Delay Relationship)

The purpose of the simulation was to observe the performance of the DTDFP resource allocation scheme under various conditions, which the system would encounter in reality.

The main measure of performance of the DTDFP resource allocation scheme was considered to be the individual's MS delay under various cumulative offered traffic load conditions. This relationship is affected by a number of factors depending on the physical channel conditions each MS encounters.

The reason this relationship was chosen to be the sole measure of performance of the resource allocation scheme is that all of the other performance characteristics can be related to it.

Mainly QoS (Quality of Service), which is very important to any resource allocation scheme, can directly be related to the MS delay vs. cumulative offered traffic relationship because QoS, for the most part, depends on delay the MS experiences, and the MS delay vs. cumulative offered traffic measures that delay at every system throughput point.

Specifically, the MS delay vs. cumulative offered traffic relationship depends on the location of the MS in question within the cell, and it also depends on the amount, and the type of data traffic each individual MS generates. The simulation was therefore performed in order to precisely determine the exact effect of these two parameters (MS location in the cell, and the size and type of data traffic the MS generates) on the MS delay vs. cumulative offered traffic relationship.

Chapter 4

Simulation Results

4.1 Organization

This part of the thesis briefly describes the general relationship between MS delay and cumulative offered traffic. It proceeds on to precisely define the definition of instability, which was observed, and used in this thesis. It goes on to define the exact observed causes of individual MS, and overall cell, instability. It shows the observed Cell Throughput vs. MS Delay relationship, and how its affected by changes in the distance between BS and MS, and MS data rates. It also shows that the delay that the MS encounters can be controlled by an appropriate choice of QSC. Finally it presents the observed limitation of the system by measuring the maximum sustainable data rate each MS can sustain indefinitely, as a function of the cumulative cell offered traffic, and distance between the BS and MS.

4.2 *MS Delay vs. Cumulative Offered Traffic Relationship*

In general, any throughput vs. delay relationship resembles an exponential curve, where in the early stage of the exponential curve the delay does not increase significantly by an increase in system offered load. However, at saturation the delay drastically increases for even a small increase in the offered load. This can be called region of instability, where a “small” increase in system throughput causes “large” increases in delay. This region of stability is very vague, and gray. However, it is the limit of the

system, and the throughput point at which it occurs depends on the individual MS offered traffic load among other things.

4.3 *Definition of Instability*

In this simulation there are two measures of instability, which are correlated very closely; queue size of the MS, and the variance of the delay the MS experiences. It was observed that when the variance of the delay the MS experiences increases to the order of 6 from the variances the MS experiences when the cell loading parameter is set to 0 % (i.e. no other active MS in the cell), the queue size of the MS is constantly increasing. The fact that the queue size is always increasing means that the stochastic average of the service rate is less than the stochastic average of the data arrival rate, and that the MS has reached the unstable region, in which it can't operate for a longer period of time. Therefore the MS is said to have entered the unstable region when the variance of its delay increases to the order of 6 from the variance it experiences when the cell loading is set to 0 %.

4.4 *Causes of Instability*

It was observed that a MS can enter the unstable state for one of two reasons, one is caused by the mobile station's own data rate, and the second one is caused by the congestion within the cell. However, the entire system (cell) only enters the unstable region because of congestion, which occurs when the sum of the traffic requirements of all of the users (cumulative cell offered traffic) exceeds the resources the BS possesses at its disposal.

When the mobile station enters an unstable region, because it requires more resources than the BS can afford to provide, the entire system (cell) does not necessarily become unstable. This occurs when one mobile station requires large amount of resources compared with the rest of the mobile stations in the cell. In this case the BS is unable to grant the mobile station in question, all of the resources it requires, which causes that particular MS to move into unstable region. However, the rest of the mobile stations remain in a stable state because they require “small” amount of resources, which the BS can afford to grant to them.

4.5 *Cell Throughput vs. MS Delay*

Figures 4.5.1-1, 4.5.1-2, and 4.5.1-3 show some typical cell Throughput vs. MS Delay curves for all of the sources (ITU Source A, B, and C). The stable regions (linear part of the graphs) show very good behavior in the sense that the delay remains constant, or increases slightly (not very noticeable, and can be ignored) with increasing congestion.

4.5.1 *Affect of distance between BS and MS on the Cell Throughput vs. MS Delay*

Figures 4.5.1-1 and 4.5.1-2 show the effect of the distance between BS and MS on the cell throughput vs. MS delay graphs. Namely, the shape of the graphs remains unchanged. The only effect the distance between BS and MS has on the MS delay, is that as the distance increases the average delay increases (average transmission rate decreases), so the graphs are simply shifted in the MS Delay axis (y axis), but maintain the same shape. The QSC was set to different values in these two graphs deliberately.

Figure 4.5.1-1. Cell Throughput vs. MS Delay (ITU Source A, as it moves away from the BS)

Figure 4.5.1-2. Cell Throughput vs. MS Delay (ITU Source C, as it moves away from the BS)

It was set to the equivalent of 1 second/packet of delay for ITU Source A, and an equivalent of 0.02 seconds/packet delay for Source C. This illustrates that the average delay can be controlled quite accurately by choosing an appropriate QSC.

Figure 4.5.1-3. Cell Throughput vs. MS Delay (ITU Source A)

From the figure 4.5.1-3 it can be seen that the distance between BS and the Source C MS, has very little or no affect on the cell throughput vs. MS delay curves of the other two MS (ITU Source A, and B). Similarly, the simulation shows that the distance between BS and ITU Sources A, and B has no affect on the cell throughput vs. MS delay curve of the Source C. Therefore, the distance between one MS and BS does not affect the cell throughput vs. MS delay curve of the other mobile stations.

4.5.2 Effect of Increasing MS Data Rate on the Cell Throughput vs. MS Delay Curve

Figure 4.5.2-1 shows the effect of increasing the MS data arrival rate on the cell throughput vs. MS delay curve. The main effect of increasing the MS data rate is that the MS reaches the unstable region at a smaller cell throughput value. It is important to stress however, that when one MS reaches the unstable region it has no affect on the stability of other MS, as can be seen from figure 4.5.1-3 where the same throughput vs. MS delay curve was obtained for ITU Source A for many different arrival rates of Source C active states.

In fact, the nature of this scheduling technique guarantees that even when one or more MS reach the unstable region (infinite delay), due to large data transmission rate requirements, the overall cell remains stable (it does not force other MS into unstable regions), and only becomes unstable when it is completely overloaded (i.e. more MS are present and active than the cell can sustain).

4.5.3 Maximum Sustainable Data Rate

Simulation was completed in order to find the stochastic average of the maximum sustainable data arrival rate a MS can support indefinitely without reaching the unstable state (without reaching the delay of infinity). Figure 4.5.3-1 shows the results for maximum sustainable data rate of a source when QSC was chosen to be the data equivalent of 0.02 seconds/packet (two time frames). The different lines in the graph represent the maximum sustainable data rate at specific distances between the BS and the MS ranging from 10 m (top line) to 1610 m (bottom line) with increments of 100 m. As

was expected the maximum sustainable data rate is achieved when the MS is closest to the BS, and when the cell is empty (lightly loaded).

Figure 4.5.3-1. System Limits

Chapter 5

Simulation Validation and Capacity Calculation

The details of the validation are shown in Appendix C. The results of the simulation were validated by comparison of the system (cell) with a modified M/D/1 queue system. The validation is first started only for ITU A sources, and later the validation model is altered in order to suit both ITU Source A, and ITU Source B traffic.

The confidence intervals were used in order to prove that an adequate number of simulation runs was completed, and that the observed results show convergent and meaningful results.

The capacity calculation was also calculated based on the ITU traffic model. The capacity of the system is measured by the probability of blocking, given the number of active ITU Source A and B sessions.

5.1 *ITU Source A Validation - Modeling of the System (given that only ITU Source A's are active in the cell)*

Having established that the system consisting of only ITU Source A's behaves as a Poisson random process (when observed on frame to frame basis), as shown in Appendix C.2 in greater detail, then the cell can be modeled by an M/D/1 queue, where arrivals arrive according to a Poisson process, and packets are served according to some constant rate.

Figure 5.1-1. M/D/1 Approximation of the Cell Behavior When only ITU Source A's are Active

Figures 5.1-2 and 5.1-3 show the M/D/1 queue (Pollaczek-Khinchin) approximation of the Cell throughput vs. MS Delay curve. The distance between the BS and MS (ITU Source A) was set to 10 meters in figure 5.1-2, and 1610 m in figure 5.1-3. This results in mobile station's average transmission rate of 10.3 Mbps in figure 5.1-2 and 1.5 Mbps in figure 5.1-3.

Figure 5.1-2 System behavior when only ITU Source As are active, distance between BS and MS = 10 m

Figures 5.1-2 and 5.1-3 show that an M/D/1 queue provides a close and fair approximation of the real system behavior. The difference between the two curves (observed delay, and M/D/1 approximation) is greater in figure 5.1-2 than in figure 5.1-3. This stems from the fact that in an M/D/1 approximation it is assumed that there is precisely one packet per transmission time slot, as well as that the entire time slot is occupied by one packet. In reality this assumption is not correct due to the packet length distribution of ITU Source A. The ITU Source A packet length is a random process which has geometric distribution with a mean of 720 bits.

Figure 5.1-3. System behavior when only ITU Source As are active, distance between BS and MS = 1610 m

Due to the fact that in simulation the distance between the MS and BS was set to 10m in figure 5.1-2, and to 1610m in figure 5.1-3, the average transmission rate is 10.3Mbps for figure 5.1-2, and 1.5 Mbps for figure 5.1-3. The length of one time slot is 0.000667 sec., which results in an average 6867 transmitted bits per time slot in figure

5.1-2, and 1000 in figure 5.1-3. However, the mean packet size of an ITU Source A is $E[\text{Packet Size}] = 720$ bits. Then, at saturation (which is where the M/D/1 curve diverges from the simulated result) when the MS has a lot of data to transmit, on the average it will transmit approximately 9.537 packets/time slot in figure 5.1-2, and 1.39 packets/time slot in figure 5.1-3. Because of this the M/D/1 approximation of the real system behavior is much better in figure 5.1-3, than in figure 5.1-2.

However, from figures 5.1-2 and 5.1-3 it can be said that the simulation behaves like an M/D/1 queue system. The difference arises because the MS in the simulation is capable of transmitting multiple packets on one time slot as the cell approaches saturation. This allows the MS in the simulation to maintain almost constant delay before the cell reaches saturation, and therefore diverge from an M/D/1 queue approximation. However, once the cell reaches saturation, the simulated result almost instantaneously increases in the delay (moves into the unstable region), and therefore converges to the M/D/1 approximation.

5.1.1 Maximum Number of Unserved Requests

From the $E[Q]$ (Expected number of packets in queue) it is possible to calculate the expected (mean) number of unserved requests for transmission in one frame at each point of the cell throughput. Figure 5.1.1-1 shows this for the case where only ITU Source A MS are active.

Figure 5.1.1-1. Number of Unserved Requests vs. Cell Loading

From figure 5.1.1-1 it can be observed that in steady state conditions, once the number of observed packet requests reaches value of seven the system becomes unstable and the delay reaches infinity. This is however an extreme, and even at this cell loading the delay is large and probably unacceptable. The comparison between figure 5.1.1-1 and figures 5.1-2 and 5.1-3 shows that the system is still stable (constant delay) when the number of unserved packet request at the end of the frame reaches the maximum value of 2.

5.2 Modeling of the System (when a mix of ITU Source A's and B's are active in the cell)

Within one ITU Source B packet call the packet inter-arrival times are of geometric distribution with a mean of 0.01 second, which is equivalent to one transmission request per frame.

Furthermore, ITU Source B packet size is constant 3840 bits, which requires a transmission rate of 5.76 Mbps, so that the entire packet can be transmitted in one time slot (0.667 ms).

Assuming that the entire packet can indeed be transmitted in one time slot, the average ITU Source B station will make a request for transmission, on one time slot, every frame for the duration of a packet call. Therefore, in steady state conditions one cell can handle only 15 ITU Source B simultaneous packet calls.

By fixing the number of overlapping (in time) ITU Source B packet calls, and assuming steady state conditions with respect to ITU Source As (some constant Poisson distributed packet transmission request arrival rate), it is possible to model the cell as an M/D/1 queuing system.

When these assumptions are used the Pollaczek-Khinchin equations remain the same with the exception that the $E[S]$ (Expected packet service time) becomes dependent on the number of simultaneous ITU Source B packet calls that are in progress.

Figure 5.2-1. M/D/1 Approximation to the Cell Behavior When Both ITU Source's are Active

Then the expected cell throughput vs. MS delay curves are shown in the following figures.

Figure 5.2-2 shows the expected cell throughput vs. MS delay curves for various combination of ITU Source As and Bs. Each line in figure 5.2-2 represents a different number of simultaneous ITU Source B packet calls (in steady state ranging from 0 to 15) that are active at the same time, with the ITU Source A load shown on the x-axis.

Figure 5.2-2. ITU Source A cell Loading vs. Expected Waiting Time (Each line represents different number of simultaneous ITU Source B packet calls, ranging from 0 to 15)

Figure 5.2-3 shows the cell throughput vs. expected total waiting time for all the different load combinations of ITU Sources A and B.

Figure 5.2-3. ITU Source A cell Loading vs. Expected Total Normalized Waiting Time (Each line represents different number of simultaneous ITU Source B packet calls, ranging from 0 to 15)

ITU Source A load is displayed on the x-axis, while each line represents different ITU Source B load (ranging from 0 to 15 simultaneous packet calls).

5.3 Simulation Validation Using Confidence Intervals

As mentioned earlier, the confidence intervals were computed for all of the simulation results. The graphs in figures 5.3-1 to 5.3-5 show the confidence intervals for the simulation result displayed in figures 5.1-2, and 5.1-3, which was compared to the M/D/1 queue system performance.

Figure 5.3-1 Confidence Intervals for the Case where the Distance Between the BS and MS is 1610m

Figure 5.3-2 90% Confidence Intervals for the Case where the Distance Between the BS and MS is 1610m

Figure 5.3-3 95% Confidence Intervals for the Case where the Distance Between the BS and MS is 1610m

Figure 5.3-4 99% Confidence Intervals for the Case where the Distance Between the BS and MS is 1610m

Figure 5.3-5 99% Confidence Interval for the Case where the Distance Between the BS and MS is 10m

Namely the graph in figure 5.3-1 shows the simulated result as compared to the 90%, 95%, 99% confidence interval for the case, which is shown in figure 5.1-3, where the distance between the mobile station and the base station is 1610 meters. This same result is also presented separately for the sake of clarity in figures 5.3-2, 5.3-3, and 5.3-4 for the 90%, 95%, and 99% confidence interval case respectively. Graph in figure 5.3-5 shows the simulated result as compared to the 99% confidence interval for the case, which is shown in figure 5.1-2, where the distance between the base station and the mobile station is 10m.

Both sets of simulation results, the cases where distance between the base station and the mobile station is 10m and 1610m were obtained by calculating an average of 28

identical simulation runs (only the random number generator seed was changed for each run) each of which recorded 10000 individual statistics.

From the results displayed in figures 5.3-1 to 5.3-5 it can be concluded that the simulation was run “long enough,” as well as that the obtained observations show meaningful results, because the simulated result is always located between the confidence intervals, as well because the stringent confidence intervals (99% confidence intervals) are “close” to the observed outcome.

5.4 *Capacity Calculation*

In DTDFP resource allocation scheme there are 15 time slots in one frame. The blocking probability can be characterized as the probability that more than 15 requests for packet transmission from 15 different MS occur within one time frame. The probability of blocking is defined as the probability that a MS will not be allowed to transmit in the current frame, given that the MS has made a request for transmission on that frame.

Furthermore, the maximum capacity of one cell can be characterized as the maximum number of active MS that the system can support indefinitely, either without the system entering the unstable state, or the capacity in which the system is able to recover (return to the stable state, given that it was in the unstable state previously).

The probability of blocking will be calculated on a frame by frame basis, and without memory. That is, the calculation for the probability of blocking in the current time frame does not take into consideration any unserved requests for transmission (blocked packets) from the previous frame. Therefore, the probability of blocking is the same for each frame, provided that the offered load has not changed (the system is in

steady state). In that case the probability of blocking is equal to the probability that the system enters the unstable state, and therefore represents the maximum capacity that the system can support with a certain probability of blocking (probability of entering the unstable state).

Although the probability of blocking is an accurate calculation for the probability that the system enters the unstable state (in steady state conditions), it does not say anything about the ability of the system to recover from the unstable state in steady state conditions. Probability of recovery calculates the probability that the system will have less unserved requests for transmission in the next, than in the current frame, given that the system has entered the unstable state.

Therefore, these two calculations, probability of blocking, and recovery probability, depict an accurate picture about the maximum capacity of the system (in steady state), which is using DTDFP resource allocation scheme.

The following sections calculate the probability of blocking for all combinations of active MS (ITU Source A, and ITU Source B). First the capacity of the system, in terms of the probability that the system enters the unstable state, is calculated for the case where only ITU Source B mobile stations are active. Then the same capacity of the system is calculated for the case where only ITU Source A mobile stations are active, and for the case in which any combination of ITU Source A and B mobile stations is active.

Finally, the capacity of the system, in terms of the ability of the system to recover from an unstable state given that it has entered the unstable state, is calculated for any combination of active ITU Source B and ITU Source A mobile stations separately.

5.4.1 ITU Source B Capacity Calculation

ITU Source B packets arrive according to geometric packet inter-arrival times within one packet call. Geometric packet inter-arrival times are discrete, and have duration times T . If the discrete packet inter-arrival times are decreased so that they approach 0 (i.e. inter-arrival times = T/n , as n approaches infinity), then the geometric distribution approaches exponential distribution. Thus if the $p[X>t]$ is a geometric r.v. then:

$$\begin{aligned} p[X > t] &= (1 - p)^{\frac{nt}{T}} \\ &= \left(\left(1 - \frac{\alpha}{n} \right)^n \right)^{\frac{t}{T}} \Rightarrow e^{\frac{-\alpha t}{T}} \quad \text{as } n \Rightarrow \infty \end{aligned}$$

Where:

$$\frac{\alpha}{n} \Rightarrow \text{probability of a packet arrival event occurring in one time } T$$

Therefore, geometric inter-arrival distribution in discrete time intervals is equivalent to exponential inter-arrival distribution in continuous time intervals. For the purpose of capacity calculations, continuous time will be used which allows the use of exponential inter-arrival times.

Furthermore, geometric inter-arrival distribution can be defined as the number of failures until the occurrence of an event, where the occurrence of an event is defined as an arrival of a packet. The packet arrival event can be characterized as a Bernoulli random variable where a definition of success is an arrival of a packet, and a failure is the absence of a packet arrival. Furthermore, by the definition of a packet call, it is

impossible for two or more packet arrival events (successes) to occur at the same time within one packet call, which is also a requirement of a Bernoulli random variable.

Having established that a packet arrival event within one packet call is a Bernoulli random process, then the sum of the number of successful Bernoulli trials (number of packet arrival events) in a given time period is a Binomial random process.

By once again dividing the time into intervals T , and allowing the size of the intervals to approach zero (T/n , as n approaches infinity), the Binomial random process approaches Poisson random process. Therefore, if the time is viewed continuously, the packet arrivals within one packet call follow Poisson distribution.

Given that N ITU Source B packet calls are active at the same time. The probability that an ITU Source B, which requires a transmission in a current time frame, is rejected due to congestion (is forced to wait for at least one time frame) is equal to the probability of more than 15 (15 time slots in one time frame) packet arrival events within one time frame. Having established that packets within one packet call arrive according to Poisson distribution, then the $p(\text{blocking})$ can be calculated as follows.

$$p(\text{blocking}) = 1 - \sum_{k=0}^{15} \frac{(\alpha_{\text{pkts}} T_F N_B)^k}{k!} e^{-(\alpha_{\text{pkts}} T_F N_B)}$$

Where:

- T_F = One frame time 0.01 sec.
- α_{pkts} = Poisson Rate of arrival of packets within one packet call
- N_B = Number of Packets Calls in one ITU Source B Session
- and $T_F \alpha_{\text{pkts}} N_B$ is the cumulative packet arrival rate

ITU Source B packets calls arrive according to a geometric inter-arrival process within one ITU Source B session. Same approximations and assumptions can be used as for packet arrivals within packet calls to prove that ITU Source B packet calls arrive according to the Poisson distribution within one ITU Source B session when time is viewed as continuous. Then, from the Poisson distribution, the probability of k packet calls occurring at the same time at time t (at least partially overlapping, figure 5.3.3-1) can be calculated as follows:

Figure 5.4.1-1. Calculation of Overlapping ITU Source B Packet calls

Where packet call arrivals follow a Poisson distribution, and packet call length is a mixture of distributions of packet arrivals within packet call (Poisson), packet inter-arrival rate within packet call (exponential), and number of packets in a packet call (Pareto).

$$\text{Packet call length} = (\text{\# of packets in a packet call})(\text{packet length}) + \sum_{0}^{\text{\# of packets in a packet call} - 1} (\text{Packet Interarrival time})$$

Where number of packets in a packet call is a Pareto distribution, and packet inter-arrival time is an exponential distribution.

The characteristic function of an exponential random variable is calculated to be:

$$\begin{aligned}\Phi_X(\omega) &= \int_0^{\infty} \alpha e^{-\alpha x} e^{j\omega x} dx \\ &= \int_0^{\infty} \alpha e^{-(\alpha - j\omega)x} dx = \frac{\alpha}{\alpha - j\omega}\end{aligned}$$

Then the sum of n independent identically distributed exponential random variables is:

$$S_L = X_1 + X_2 + \dots + X_L$$

And in terms of the characteristic function:

$$\begin{aligned}\Phi_{S_L}(\omega) &= E[e^{j\omega S_L}] = E[e^{j\omega(X_1 + X_2 + \dots + X_L)}] \\ &= E[e^{j\omega X_1}] \dots E[e^{j\omega X_L}] \\ &= \Phi_{X_1}(\omega) \dots \Phi_{X_L}(\omega) \\ &= \left(\frac{\alpha}{\alpha - j\omega} \right)^L\end{aligned}$$

This is exactly an m-Erlang random variable. Then the pdf is obtained by calculating the inverse Fourier transform.

$$\begin{aligned}
f_{S_N}(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \Phi_{S_N}(\omega) e^{-j\omega x} d\omega \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(\frac{\alpha}{\alpha - j\omega} \right)^L e^{-j\omega x} d\omega \\
&= \frac{\alpha e^{-\alpha x} (\alpha x)^{L-1}}{(L-1)!} \quad x > 0,
\end{aligned}$$

Where :

L = Number of packets in one packet calls

α = Poisson Arrival Rate

Then the packet call length distribution can be calculated as follows:

$$\begin{aligned}
& p(\text{Packet call length} < T) \\
&= \sum_{n=2}^{\infty} p(n \text{ packets in a packet call}) p(\text{Sum of interarrival times} < (T - n(6.67 \times 10^{-4}))) \\
&\quad + p(1 \text{ packet in a packet call}) \times \begin{pmatrix} 1 & \text{if } T \geq 6.667 \times 10^{-4} \\ 0 & \text{otherwise} \end{pmatrix} \\
&= \sum_{n=2}^{\infty} \frac{ab^a}{n^{a+1}} \left(\int_0^{T-n(6.67 \times 10^{-4})} \left(\frac{\alpha_{pkts} e^{-\alpha x} (\alpha_{pkts} x)^{n-1}}{(n-1)!} \right) dx \right) + ab^a \times \begin{pmatrix} 1 & \text{if } T \geq 6.667 \times 10^{-4} \\ 0 & \text{otherwise} \end{pmatrix}
\end{aligned}$$

Where :

a, b = Pareto Distribution Variables

T = shown in graph 5.3.1-1

Then the probability of k overlapping sessions can be calculated as follows (figure 5.3.3-1):

$$\begin{aligned}
& p(k \text{ overlapping packet calls} / M_B) \\
& = p(\text{session arrivals at time } \tau) p(\text{session length } t - \tau) \\
& = \int_0^t \frac{(\alpha_B (t - \tau) M_B)^k}{k!} e^{-(\alpha_B (t - \tau) M_B)} \left(1 - \sum_{n=2}^{\infty} \frac{ab^a}{n^{a+1}} \left(\int_0^{t-\tau-n(6.67 \times 10^{-4})} \left(\frac{\alpha_{pkts} e^{-\alpha x} (\alpha_{pkts} x)^{n-1}}{(n-1)!} \right) dx \right) \right. \\
& \quad \left. + ab^a \begin{pmatrix} 1 & \text{if } (t - \tau) \geq 6.667 \times 10^{-4} \\ 0 & \text{otherwise} \end{pmatrix} \right) d\tau
\end{aligned}$$

Where :

M_B = Numboer of Active ITU Source B Sessions

α_B = Poisson Rate of Arrival of ITU Source B Packet Calls

Then the probability of blocking can be calculated as

$$p(\text{Blocking} / M_B) = \sum_k p(k \text{ overlapping packet calls}) p(\text{Blocking} / k \text{ overlapping packet calls})$$

$$\begin{aligned}
& p(\text{blocking} / M_B) \\
& = \sum_{k=0}^{\infty} \left(\int_0^t \frac{(\alpha_B (t - \tau) M_B)^k}{k!} e^{-(\alpha_B (t - \tau) M_B)} \left(1 - \sum_{n=2}^{\infty} \frac{ab^a}{n^{a+1}} \left(\int_0^{t-\tau-n(6.67 \times 10^{-4})} \left(\frac{\alpha_{pkts} e^{-\alpha x} (\alpha_{pkts} x)^{n-1}}{(n-1)!} \right) dx \right) \right. \right. \\
& \quad \left. \left. \times \begin{pmatrix} 1 & \text{if } \tau \geq 6.667 \times 10^{-4} \\ 0 & \text{otherwise} \end{pmatrix} + ab^a \right) d\tau \right) \\
& \quad \times \left(1 - \sum_{n=0}^{15} \frac{(\alpha_{pkts} T_F k)^n}{n!} e^{-(\alpha_{pkts} T_F k)} \right)
\end{aligned}$$

α_B = Poisson arate of arrival of packet calls within one session

T_p = Time duration of one packet call

M_B = Number of Active ITU Source B sessions

However, it is possible to simplify this calculation by providing an approximation of the probability of blocking. This can be achieved by using the expected packet call length rather than the exact packet call length. The expected packet call length can be calculated as follows.

$$\begin{aligned}
T_2 &= E[\text{Packet Call Length}] \\
&= E\left[(\# \text{ of packets in a packet call})(\text{packet length}) \right. \\
&\quad \left. + (\# \text{ of packets in a packet call} - 1)(\text{Packet Interarrival time}) \right] \\
&= \sum_{n=b+1}^{\infty} n \frac{ab^a}{n^{a+1}} (6.67 \times 10^{-4}) + (n-1) \frac{ab^a}{n+1} \int_0^{\infty} \frac{x}{\alpha_{pkts}} e^{-\frac{x}{\alpha_{pkts}}} dx \\
&= \sum_{n=b+1}^{\infty} n \frac{ab^a}{n^{a+1}} (6.67 \times 10^{-4}) + (n-1) \frac{ab^a}{n+1} \left(\frac{1}{\alpha_{pkts}} \right)
\end{aligned}$$

Where:

a, b Are Pareto Distribution variables

Then the probability of blocking can be calculated as follows:

$$p(k \text{ overlapping sessions} / M_B) = \frac{(\alpha_B T_2 M_B)^k}{k!} e^{-(\alpha_B T_2 M_B)}$$

Then, given the Number of Active ITU Source B sessions:

$$p(\text{Blocking} / M_B) = \sum_{k=0}^{\infty} \left(\frac{(\alpha_B T_2 M_B)^k}{k!} e^{-(\alpha_B T_2 M_B)} \right) \left(1 - \sum_{n=0}^{15} \frac{(\alpha_{pkts} T_F k)^n}{n!} e^{-(\alpha_{pkts} T_F k)} \right)$$

The results of the above calculations are shown in figure 5.3-1. Figure 5.3-1 shows the various probability of blocking curves given the expected number of ITU Source B packets in one ITU Source B packet call.

Figure 5.4.1-2. ITU Source B Capacity (For various Packet Call Lengths)

Therefore the capacity of a cell, in terms the number of simultaneous active sessions the cell can support depends on the length of an ITU Source B packet call.

5.4.2 ITU Source A Capacity

By using the same traffic approximation assumptions for ITU Source A as for ITU Source B in moving from slotted time to continuous time it can be said that ITU Source A packet calls arrive according to Poisson distribution within one ITU Source A

session. In addition, ITU Source A expected packet size is 720 bits, which is very small and can easily be transmitted in one time slot by even the smallest data transmission rates. Moreover, ITU Source A packet call consists of only one packet, which simplifies the calculation even more because the sum of the exponential inter-arrival times need not be calculated.

Then the maximum capacity (blocking) occurs when more than 15 MS request permission for transmission in one time frame.

$$p(\text{blocking} / M_A) = 1 - \sum_{k=0}^{15} \frac{(\alpha_A T_F M_A)^k}{k!} e^{-(\alpha_A T_F M_A)}$$

Where:

T_F = *One Frame Time 0.01sec.*

α_A = *Poisson rate of arrival of ITU Source A packet calls within one session*

M_A = *Number of active ITU Source A Sessions*

The result of the capacity calculation when only ITU Source A's are active in the cell is shown in figure 5.3-2. From figure 5.3-2, it can be seen that the cell capacity when only ITU Source A's are active does not depend on the length of the ITU Source A packet call. This of coarse is because ITU Source A packet call is always made up of only one packet. The peculiar result however, is that for most part the cell can sustain more ITU Source A simultaneous sessions that ITU Source B simultaneous sessions, given that ITU Source B packet calls are made up of less than 100 packets on the average. This occurs because ITU Source B packet call is a much more rare event than an ITU Source

A packet call. Therefore, even though ITU Source B packet calls are longer, (given that they are less than 100 packets long,) they occur less frequently than ITU Source A packet calls, and for that reason present a lower burden on the cell.

Figure 5.4.2-3. ITU Source A Capacity

By comparing the ITU Source A Capacity graph (figure 5.3-2) from this section, with the Expected Waiting Time Graph (figure 5.2-2) from the Validation section, it can be observed that the two results agree. Namely, in the Validation section the Expected Waiting Time graph (figure 5.2-2) when zero ITU Source B packet calls are active, the cell reaches instability (rapidly increasing delay) when approximately 1000 ITU Source A sessions are active. From the ITU Source A Capacity graph (figure 5.3-2) in this section it can be seen that the blocking probability starts to rapidly increase when around 1000 ITU Source A sessions are active, and 0 ITU Source B Sections are active. Therefore, even though the two results above were obtained by using completely

different methods and assumptions, they are correlated, which further substantiates the correctness of both of the sets of results.

5.4.3 Capacity Calculation for a Combination of ITU Source A and B

Keeping in accordance with the traffic assumptions and approximations used in individual capacity calculations for both ITU Source A, and B the cell capacity for different combinations of ITU Sources A, and B can be calculated in the following way.

$$p(\text{Blocking} / M_B, M_A) = p(k \text{ overlapping src B packets} / M_B) \times p(((k \text{ overlapping src. A packets}) + (k \text{ overlapping src. B packets})) > 15 / M_A)$$

Where :

M_B = Number of active Source B sessions

M_A = Number of active Souce A sessions

Then :

$$p(\text{Blocking} / M_A, M_B) = \sum_{k=0}^{\infty} \left(\int_0^{\tau} \frac{(\alpha_B \tau M_B)^k}{k!} e^{-(\alpha_B \tau M_B)} \left(1 - \sum_{n=2}^{\infty} \frac{ab^a}{n^{a+1}} \left(\int_0^{\tau - \tau - n(6.67 \times 10^{-4})} \left(\frac{\alpha_{pkts} e^{-\alpha_{pkts} x} (\alpha_{pkts} x)^{n-1}}{(n-1)!} \right) dx \right) \right) + ab^a \begin{pmatrix} 1 & \text{if } \tau \geq 6.667 \times 10^{-4} \\ 0 & \text{otherwise} \end{pmatrix} \right) d\tau \right) \times \left(1 - \sum_{n=0}^{15} \frac{((\alpha_{pkts} k + M_A \alpha_A) T_F)^n}{n!} e^{-((\alpha_{pkts} k + M_A \alpha_A) T_F)} \right)$$

Then the approximation of probability of blocking becomes:

$$p(\text{Blocking} / M_A, M_B) = \sum_{k=0}^{\infty} \left(\frac{(\alpha_B T_p M_B)^k}{k!} e^{-(\alpha_B T_p M_B)} \right) \left(1 - \sum_{n=0}^{15} \frac{((\alpha_{p_{kts}} k + M_A \alpha_A) T_F)^n}{n!} e^{-((\alpha_{p_{kts}} k + M_A \alpha_A) T_F)} \right)$$

5.5 Probability of Recovery

As mentioned earlier the probability of recovery is the probability that the system will have less unserved requests for transmission in the next frame, as opposed to the current frame, given that the system has unserved requests for transmission in the current frame. All of the calculations assume steady state offered load conditions. Therefore, the probability of recovery is equivalent that the number of unserved requests for transmission will not increase indefinitely under steady state offered load.

Figure 5.4-1 shows the finite state machine of the system, which has entered the

Figure 5.5- 1 System Finite State Machine

unstable state. In figure 5.4-1 m represents the number of unserved requests for transmission at the end of each frame. The finite state machine, in figure 5.4-1, changes states at the end of each WCDMA time frame.

Therefore the finite state machine (figure 5.4-1) has the ability to enter any state N , which is greater than m . This occurs when the number of request for transmission arrivals in the next frame is equal to $N + 15$ (i.e. at the end of the current frame 15 requests for transmissions are served, which leaves $m-15$ unserved requests for transmission, and when the $N+15$ new requests for transmissions arrive the final state is $m-15+N+15 = m + N$).

However, the finite state machine (in figure 5.4-1) can only enter 15 states, which are smaller than m . This occurs when the number of request for transmission arrivals in the next frame is equal to L , where L is greater than or equal to zero, and less than or equal to 14 (i.e. at the end of the current frame 15 requests for transmissions are served, which leaves $m-15$ unserved requests for transmission, and when the L new requests for transmissions arrive the final state is $m-15+L$).

The probability of recovery in the following calculations is defined as the probability that the next state the system enters is smaller than, or equal to m , given that the system is currently in state m . Then the probability of recovery ($p(R/M_B)$) is calculated as follows:

$$\begin{aligned}
p(R/M_B) &= \sum_{l=0}^{15} p(l \leq 15/M_B) \\
&= \sum_{l=0}^{15} \sum_{m=0}^{\infty} p(m/M_B) p(l \leq 15/m, M_B) \\
&= \sum_{l=0}^{15} \sum_{m=0}^{\infty} p(m/M_B) \frac{(\alpha_{pkts} T_F N_B)^l}{l!} e^{-(\alpha_{pkts} T_F N_B)}
\end{aligned}$$

Where :

l = Number of new arrivals of requests for transmission in current frame

m = Current state of the system (Figure 5.5 – 4)

Where the probability that the system enters the state m ($p(m/M_B)$), under steady state conditions, can be calculated by using $p(k \text{ overlapping packet calls } / M_B)$ from equation (18). This calculation assumes that the system offered load is instantly increased from zero to the required steady state load. In that case the probability that the system is in state m from figure 5.4-1 is calculated as follows.

$$\begin{aligned}
p(m/M_B) &= \sum_k p(k \text{ overlapping packet calls } / M_B) \times p(m \text{ new packet arrivals } / k \text{ overlapping packet calls}) \\
&= \sum_{k=0}^{\infty} \left[\int_0^t \frac{(\alpha_B (t-\tau) M_B)^k}{k!} e^{-(\alpha_B \tau M_B)} \left(1 - \sum_{n=2}^{\infty} \frac{ab^a}{n^{a+1}} \left(\int_0^{t-\tau-n(6.67 \times 10^{-4})} \left(\frac{\alpha_{pkts} e^{-\alpha x} (\alpha_{pkts} x)^{n-1}}{(n-1)!} \right) dx \right) + ab^a \right) d\tau \right. \\
&\quad \times \left(\begin{matrix} 1 & \text{if } (t-\tau) \geq 6.667 \times 10^{-4} \\ 0 & \text{otherwise} \end{matrix} \right) \left. \right] \\
&\quad \times \left(\frac{(\alpha_{pkts} T_F k)^{m+15}}{(m+15)!} e^{-(\alpha_{pkts} T_F k)} \right)
\end{aligned}$$

Then the

$$p(R/M_B) = \sum_{l=0}^{15} p(m/M_B) \left(\frac{(\alpha_{pkts} T_F k)^l}{l!} e^{-(\alpha_{pkts} T_F k)} \right) \begin{matrix} \text{probability of} \\ \text{recovery } p(R / \end{matrix}$$

M_B) is calculated as follows:

By applying the same procedure for the case when only ITU Source A sessions are active it is possible to calculate the recovery probability ($p(R/M_A)$) when only ITU Source A sessions are active.

Then $p(m/M_A)$ is:

$$p(m/M_A) = \frac{(\alpha_A T_F M_A)^m}{m!} e^{-(\alpha_A T_F M_A)}$$

And $p(R/M_A)$ is:

$$p(R/M_A) = \sum_{l=0}^{15} \sum_{k=0}^{\infty} \left(\frac{(\alpha_A T_F M_A)^k}{k!} e^{-(\alpha_A T_F M_A)} \right) \left(\frac{(\alpha_A T_F M_A)^l}{l!} e^{-(\alpha_A T_F M_A)} \right)$$

Figure 5.4-2 shows the recovery probability ($p(R/M_B)$) when only ITU Source B mobile stations are active (Expected number of ITU Source B packet calls in one ITU Source B session was chosen to be 93), and figure 5.4-3 shows the same information for the case in which only ITU Source A mobile stations are active.

Figure 5.5- 2 Probability of Recovery When Only ITU Source B is Active

Figure 5.5- 3 Probability of Recovery When Only ITU Source A is Active

By comparing figure 5.4-2 from this section with figure 5.3-1 it can be observed that the probability of blocking starts increasing rapidly at around 500 active ITU Source B sessions, where the recovery probability starts decreasing rapidly at around 700 active ITU Source B sessions. From this it can be concluded that the capacity of the system, in which only ITU Source B sessions are active, and in which the expected number of packet calls in one ITU Source B session is 93, is 700 simultaneously active ITU Source B sessions. Clearly, in the region between 500 and 700 active ITU Source B sessions blocking does occur, but the system is able to recover from it with probability almost equal to 1. As the number of active ITU Source B sessions is further increased beyond 700, the probability of recovery decreases rapidly, and it can be said that the system is likely to enter (if it has not already entered) the unstable state (always increasing delay).

By comparing figure 5.4-3 from this section with figure 5.3-2 it can be observed that the probability of blocking starts increasing rapidly at around 800 active ITU Source A sessions, where the recovery probability starts decreasing rapidly at around 1000 active ITU Source A sessions. From this it can be concluded that the capacity of the system, in which only ITU Source A sessions are active, is 1000 simultaneously active ITU Source A sessions. Clearly, in the region between 800 and 1000 active ITU Source A sessions blocking does occur, but the system is able to recover from it with probability almost equal to 1. As the number of active ITU Source A sessions is further increased beyond 1000, the probability of recovery decreases rapidly, and it can be said that the system is likely to enter (if it has not already entered) the unstable state (always increasing delay).

Furthermore, figure 5.4-3 is consistent with figures 5.2-2, and 5.2-3 from validation section (section 5.2). Namely, figure 5.4-3 shows that the system enters the unstable state (always increasing delay), under steady state offered load conditions, at around 1000 active ITU Source A sessions, and 0 active ITU Source B sessions, as do the figures 5.2-2, and 5.2-3, which exhibit a large increase in delay starting at around 1000 active ITU Source A, and 0 active ITU Source B sessions. Therefore, these results further validate the accuracy of the validation section, and the assumptions used in that section.

Chapter 6

Advantages of This Scheduling Technique & Conclusion

6.1 *More Efficient Capacity Packing*

[2] The characteristics of the shared channel can be adjusted dynamically in order to counteract rapid changes in channel conditions. In this thesis the maximum throughput of the shared channel is achieved by the [4] optimum allocation of C/I. This dynamic adaptation of the shared channel to ever changing channel conditions is very difficult, if not impossible to achieve with any other resource allocation scheme that is not dynamic, and that does not have the fat pipe characteristics. [2] By allowing only short leases on channel resources the system remains very responsive to changes in the channel conditions.

6.2 *Fairness*

6.2.1 “1 flag” Request

The nature of the first come first serve scheduler is inherently fair, in a sense that the MS that asks for permission to transmit first will be given the permission to transmit

first. That way, when a particular frame has more requests for transmission, then available time slots, it will assign time slots for transmission in the order in which the requests were received. This means that the MS, which issued the last request, will not be allowed to transmit in the next frame. However, that request will be the first in line for the following frame, and surely will be granted a first time slot in that frame, as shown in figure 3.3.2-1 (this specific scheduling example).

6.2.2 “2 flag” Request

The DTDFP scheduling technique exhibits fairness even when multiple MS issue “2 flag” requests. Figure 3.3.2-2 shows an example of scheduling technique when multiple “2 flag” requests are received. From figure 3.3.2-2 it can be seen that the “2 flag” request that was received first will be scheduled last, and the remaining “2 flag” requests will be scheduled as “1 flag” requests in the order in which they were issued. This means that in the frame 2 the “2 flag” request will be granted to the MS that issued the second “2 flag” request in frame 1, and that in frame 3 the “2 flag” request will be granted to the MS that issued the third “2 flag” request in frame 1, and so on. This is obviously fair because all of the MS that issued “2 flag” requests will be granted the requested resources in a first come first serve basis, and will rotate (share) those resources.

6.3 Minimum Service Rate Guarantee

Given that the cell is not overloaded, all of the MS will be guaranteed on the average the minimum of one time slot per frame, as can be seen from figure 3.3.2-1 and figure 3.3.2-2.

6.4 Conclusion

The DTDFP resource allocation scheme was designed, simulated, validated, and analyzed in terms of the number of various types of sessions it can support for an indefinite amount of time. Furthermore, its performance was tested, and analyzed under numerous realistic conditions, including log-normal fading, the distance between MS and the BS, various cell offered traffic conditions, and various MS individual offered traffic conditions. Throughout these tests the DTDFP resource allocation scheme showed positive results.

The main reason the DTDFP resource allocation scheme was tested, was to compare its performance with an optimum result, which can “ideally” be approached by a long-term optimization based method. The performance of the DTDFP resource allocation scheme showed very good results in the sense that the MS delay was maintained constant for various cumulative cell offered traffic (cell throughput). Only at around 90% to 95% of throughput did the MS start to experience any delay increases.

The DTDFP resource allocation scheme was observed to be fair under all cumulative, and individual MS offered traffic levels. Furthermore, the DTDFP resource allocation scheme exhibited fairness even when one or more MS required significantly

higher data traffic rates than the rest of the MS within the cell. In this case the DTDFP resource allocation method provided the maximum data rates it can afford to the MS requiring large traffic rates, without a significant increase in the delay the rest of the mobile stations experience. In the case where the BS was unable to grant the required data rate to one of the MS, which requested significantly higher data rates, only that mobile station would experience a significant increase in the delay (reached instability). The other mobile stations were unaffected, and most importantly the overall cell remained stable.

It was shown that the DTDFP resource allocation scheme provides a mechanism by which the MS can control the maximum amount of delay it experiences (QSC), and therefore can guarantee some minimum QoS, given that the cell is not overloaded. Furthermore it was shown that the DTDFP resource allocation scheme allows for greater shared channel throughput optimization through [4] Optimum Allocation of C/I.

In addition the DTDFP resource allocation provides for insignificant waste of resources, and thereby provides large throughput value, very close to 100 percent, given that there is enough offered traffic. This is possible because the only waste of bandwidth occurs in signaling overhead which consists of three bits every time slot, which for all intensive purposes can be considered insignificant. Therefore, the DTDFP resource allocation scheme provides an optimum result, if the optimum result is to maximize throughput and QoS, while at the same time minimizing delay. This result however, is achieved at a very low signaling overhead, and implementation complexity, when compared to any long-optimization based resource allocation methods.

6.5 Future Work

This thesis analyzed the performance of the DTDFP resource allocation scheme in terms of the delay the MS experiences under various offered traffic conditions. The concept of the QoS was explored very briefly by noting that delay can indeed be controlled to a certain extent by the mobile station (using the QSC). The accuracy with which the delay can be controlled was never explored. This analysis would be very interesting, especially the ability of the QSC to control the delay under various MS input data traffic, and cumulative cell offered traffic conditions. It would also be very interesting to test the impact of errors in the MS calculation of QSC on the other mobile stations' delay and QoS.

Finally, the QSC in this thesis was designed to be controlled by the maximum acceptable delay, which the application the MS is running can sustain. The maximum acceptable delay was chosen to be the controlling parameter due to the fact that some mobile stations needed greater data rates more than others, and the only way to ensure that, only the MS which needs larger data rates most urgently gets them, was to select the maximum acceptable delay as the controlling parameter. Although there is nothing wrong with this method of controlling the QSC it is not optimum. It may be possible to control the QSC in an adaptive way, which depends on the cell loading (cumulative cell offered traffic), such that the delay the MS experiences is minimized even further. Although this kind of control of the QSC would not gain much in terms of minimum QoS (minimum level of QoS is already provided by the QSC) it would absolutely minimize the delay the MS experiences at any time, and at any cumulative cell offered traffic level, and therefore maximize the QoS.

Appendix A

Mapping of C/I to the Data Rates

Depending on the C/I ratio the MS encounters, the utilization of the AMCS (Adaptive Modulation and Coding Schemes) from section 1.2 of this thesis, provides the MS with the ability to achieve the data rates shown in figure A-1. These data rates were used in the simulation discussed in section 3.3 of this thesis. The data was obtained from [4].

Carrier to Interference Ratio	MS Data Rate (bps)
$C/I < -16$	0
$-16 < C/I < -12$	59940
$-12 < C/I < -11$	119880
$-11 < C/I < -9$	179820
$-9 < C/I < -8$	299700
$-8 < C/I < -7$	359640
$-7 < C/I < -6$	419580
$-6 < C/I < -5$	599400
$-5 < C/I < -4$	719280
$-4 < C/I < -3$	899100
$-3 < C/I < -2$	1138860
$-2 < C/I < -1$	1493100
$-1 < C/I < 0$	1846260
$0 < C/I < 1$	2317140
$1 < C/I < 2$	2620080
$2 < C/I < 3$	2925180
$3 < C/I < 4$	3293460
$4 < C/I < 5$	3836160
$5 < C/I < 6$	4375620
$6 < C/I < 7$	4975020
$7 < C/I < 8$	5542020
$8 < C/I < 9$	6261300
$9 < C/I < 10$	7100460
$10 < C/I < 11$	7432560
$11 < C/I < 12$	7912080
$12 < C/I < 13$	8451540
$13 < C/I < 14$	9170820
$14 < C/I < 15$	9950040

$15 < C/I$	10789200
------------	----------

Figure A-1 MS Data Rates vs. Carrier to Interference Ratio

Appendix B

Details of the Simulation of the DFP Resource Allocation Scheme

MS's Data Rates

Every time the MS is about to begin transmission it obtains a random C/I ratio from an appropriate C/I distribution based on its distance from the BS. Based on the obtained C/I value the MS data rate is then calculated according to Table A-1 provided in Appendix A. The data rate for all of the MS (Source A, B, and Large, queues) was calculated in this way.

Load_C_I

The load_C_I process model imports C/I distributions, which were calculated externally, as functions of distances.

Process Models (Finite State Machines)

In OPNET a final state machine can have one of two possible states, forced or unforced.

When the process enters a forced state it must leave it immediately after the completion of all of the tasks in that particular state. Because this is an event driven simulation it means that no time is wasted in a forced state. In other words the simulation time when the process enters the forced state is identical to the simulation time when the process exits the forced state.

On the other hand, the simulation time when the process enters the unforced state is not necessarily the same as the simulation time when the process exits the unforced state. The unforced state is only exited when an interrupt, which was scheduled in local or some other (foreign) process, occurs which causes the process to exit the unforced state.

Therefore the simulation time only changes when a process enters the unforced state. This means that all of the tasks, which are completed in the forced state, occur immediately with zero delay.

cell_data_src

The cell_data_src is the aggregate cell data traffic source. The purpose of the aggregate cell data traffic source is to simulate the “rest” of the MS effect on the MS in question. Simulating the aggregate data traffic by one source allows for the exact control over the cell loading (aggregate offered traffic), which is more precise, then simulating the aggregate traffic by direct summation of a number of different MS.

Aggregate source was simulated in OPNET as OPNET's own simple source with an addition of a cell-loading parameter. Cell-loading parameter is defined to be the average number of time slot requests in one frame issued by the "rest of the users within the cell" (i.e. the aggregate source). Cell-loading parameter was designed to be a random process following Poisson distribution with a mean ranging from 0 time slot requests in one frame (i.e. cell loading 0%, empty cell) to 15 time slot requests in one frame (i.e. cell loading 100%, full cell), which was changed depending on the requirements of each individual simulation run.

Furthermore the distribution in time of those requests was designed to be uniform in time between 0 sec. (beginning of a time frame), to 0.01 sec. (end of a time frame).

Figure B-1. Finite State Machine of the Aggregate Source

cell_queue

The **cell_queue** behaves as an aggregate queue. It accepts the data generated by the aggregate source (**cell_data_src**), and queues it accordingly. In that sense aggregate queue functions exactly as the Source A, B, and Large queues, with one exception. The

aggregate queue is allowed to make as many requests for transmission (requests for time slots) in one time frame as the data equivalent of the number of time slots that it receives from the aggregate source. The times those requests are made correspond to the times data was generated by the aggregate source (i.e. there is zero transmission delay between aggregate source and aggregate queue).

In order to minimize computing complexity, and therefore the time duration of the simulation, the cell_queue (aggregate queue) does not really transmit any data to the MS. The aggregate queue receives the traffic generated by the aggregate source, queues it accordingly, and then makes requests for transmission based on the amount of traffic in its queue. After receiving the permission to transmit from the BS the aggregate queue destroys the data in question, and updates its queue size.

For the purpose of the simulation it is irrelevant that the aggregate queue doesn't transmit any data to the BS. Rather what is important, is that the aggregate queue makes requests for transmission based on the amount of traffic in its queue, reserves those time slots for it self (causes congestion) so that no other MS can transmit during those time slots.

MS A

MS A (Mobile Station A) simulates the behavior of the MS whose uplink data traffic follows the parameters specified by the ITU Source A traffic model. MS A consists of two units, tr_src_a to generate the data traffic specified by the ITU Source A, and mob_q_a to transmit that data traffic according to the ATD Resource Allocation scheme.

tr_src_a (Traffic Source A)

Traffic source A generates the data traffic according to the ITU Source A traffic model. After generating the data traffic the traffic source A transmits that data to the mob_q_a (mobile queue A) with no time delay.

Figure B2. Finite State Machine of Traffic Source A

Operation of the tr_src_a (Traffic Source A)

The traffic source A is a modified fifo queue, in which all of the states, except *pk_generate*, are unforced.

Init

In the “init” state the traffic source A initializes all the variables which are needed for its operation. *Init* is an unforced state in which the mobile completes its initialization tasks and waits for the simulation start interrupt.

Session

The *session* state is an unforced state in which the process schedules an arrival of a new ITU Source A session, as well as computes the number of packet calls in that session according to the parameters specified in the ITU Source A traffic model.

Generate

The *generate* state is an unforced state in which the traffic source A generates an interrupt for the next packet call arrival according to the statistics outlined in the ITU Source A traffic model. In the *generate* state the process also checks whether or not the session has ended, and generates a session end interrupt if the session has indeed ended.

pk_generate

The *pk_generate* state is a forced state in which the traffic source A generates a packet according to the statistics outlined in the ITU Source A traffic model, and generates an interrupt to deliver the packet to the mobile queue A without any time delay.

Stop

The *stop* state is an unforced state, which the process enters only when the simulation has reached the end. The *stop* state cancels any event that was scheduled previously, but has not been completed, and stops the scheduling of any new future interrupts.

mob_q_a (Mobile Queue A)

Mobile queue A receives the traffic generated by the traffic source A, segments the received packets into bits which can be transmitted on one time slot, makes requests for transmission to the BS, and transmits on time slots which were granted to it by the BS.

Mobile queue B (mob_q_b) and large queue (lrg_user) behave in the same way as the mobile queue A.

Packet Segmentation

As soon as the packets are received by the mobile queues from the traffic sources they are segmented into bits, and stored in a SAR (segmentation and reassembly) buffer.

Before every transmission the mobile queues obtains a random C/I ratio value from the C/I distribution corresponding to its location within the cell. From the resulting C/I value the mobile queues determine the maximum transmission rate according to the optimized C/I allocation procedure, and based on that transmission rate it calculates the number of bits to be transmitted in the next transmission.

$$N = R \times \begin{pmatrix} \text{one time slot (0.000667 sec.) for flag 1 transmission} \\ k \text{ time slots (k(0.000667 sec.)) for flag 2 transmission} \end{pmatrix}$$

Where

$N \Rightarrow$ Number of Bits to be transmitted

$R \Rightarrow$ Calculated transmission rate

When the mobile queues receive the permission for transmission they transmit the calculated number of bits.

In addition, before issuing a request for transmission the mobile queues compare their queue size with the QSC corresponding to the application each MS is running. Then, if the QSC is greater than, the queue size of the mobile queue the mobile queue issues a “1 flag” request, otherwise it issues a “2 flag” request.

Figure B-3. Mobile Queue A Finite State Machine

Operation of the Mobile Queue

The mobile queue is also a modified fifo queue, in which only the “init,” and “idle” states are unforced. All the other states are forced.

Init

In the *init* state the mobile queue initializes all the variables which are needed for its operation. *Init* is an unforced state in which the mobile completes its initialization tasks and waits for the first packet arrival.

Load_C_I

Load_C_I is a forced state. The mobile queue enters the *Load_C_I* state only once at the beginning of the simulation (first packet arrival), and never enters it again for the rest of the simulation.

In the *Load_C_I* state the mobile queue loads the particular C/I distribution depending on the setting of the “Distance from BS” variable, which is adjusted for the requirements of each individual simulation run.

Segment

The mobile queue enters the “segment” state after every arrival (after “arrival” state), except for the first packet arrival event in which the segment state is entered from the “*Load_C_I*” state.

The segment state is a forced state, meaning that all of its tasks are completed instantaneously without delay. In the “segment” state the mobile queue segments the newly arrived packet into bits, and stores them in the SAR buffer.

Arrival

The *arrival* state is a forced state in which the mobile queue accepts the newly arrived packet, and queues it with no priority into the queue.

Sense_chan

The *sense_chan* is a forced state in which the mobile queue chooses a random C/I value based on the C/I distribution that was loaded in the *Load_C_I* state, and computes the transmission rate, and the number of bits to be transmitted in the next transmission. Furthermore, in the *sense_chan* state the mobile queue compares its queue size with the QSC, and decides which request for transmission to make (“1 flag,” or “2 flag” request), and makes that request. The *sense_chan* state also puts a time stamp, and mobile queue ID on that request, so that the MS knows the time corresponding to that request.

Svc_start

The *svc_start* state is a forced state in which the mobile queue gets a handle on the packet which is being transmitted, computes the amount of time the packet will require to be transmitted (depending on the “1 flag,” or “2 flag” request), and schedules an interrupt (service completion interrupt) at that time.

Svc_compl

The *svc_compl* is a forced state signifying the end of packet transmission. In the *svc_compl* state the mobile queue generates an interrupt in order to force the BS process model to accept a packet arrival interrupt.

Reassembly

The *reassembly* state is a forced state in which the mobile queue removes the transmitted packet from the SAR buffer and the queue, and updates all of the variables.

Idle

The *idle* state is an unforced state in which the mobile queue awaits for any number of possible interrupts. It allows the simulation to exit this process model, and enter some other process model, which has a scheduled interrupt. Because the *idle* state is the only unforced state which the process model enters more than once, it is the only state in which the simulation time is allowed to change.

tr_src_b (Traffic Source B)

Traffic source B generates the data traffic according to the ITU Source B traffic model. After generating the data traffic the traffic source B transmits that data to the mob_q_b (mobile queue B) with no time delay.

Figure B-4. Traffic Source B Finite State Machine

Operation of the tr_src_b (Traffic Source B)

The traffic source B is also a modified fifo queue, in which all of the states, except *pk_generate*, are unforced.

init

In the “init” state the traffic source B initializes all the variables which are needed for its operation. *Init* is an unforced state in which the mobile completes its initialization tasks and waits for the simulation start interrupt.

Session

The *session* state is an unforced state in which the process schedules an arrival of a new ITU Source B session, as well as computes the number of packet calls in that session according to the parameters specified in the ITU Source B traffic model.

Packet_call

The *packet_call* state is a forced state in which the process checks weather or not the session has come to an end, in which case it generates a *session_terminate* interrupt. If this is not the case the *packet_call* state generates the packet call distribution, packet call interrupt to begin the first packet call, and the number of packets in a packet call according to the statistics outlined in the ITU Source B traffic model.

Generate

The *generate* state is an unforced state in which the traffic source B generates an interrupt for the next packet arrival according to the statistics outlined in the ITU Source B traffic model. In the *generate* state the process also checks weather or not the current packet call has ended, and generates a packet call end interrupt if the packet call has indeed ended.

pk_generate

The *pk_generate* state is a forced state in which the traffic source B generates a packet according to the statistics outlined in the ITU Source B traffic model, and generates an interrupt to deliver the packet to the mobile queue B without any time delay.

Stop

The *stop* state is an unforced state, which the process enters only when the simulation has reached the end. The *stop* state cancels any event that was scheduled

previously, but has not been completed, and stops the scheduling of any new future interrupts.

Lrg_src_old (Large Traffic Source)

Large traffic source generates the data traffic according to the Large Traffic Source model. After generating the data traffic the large traffic source transmits that data to the mob_q_lrg (large mobile queue) with no time delay.

Large traffic source was designed in order to observe the performance of the DTDFP resource allocation scheme when one source requires significantly larger amount of resources when compared to the rest of the mobile stations. Large source traffic process model was designed to be identical to the tr_src_a (ITU Source A) process model, only the magnitude of the traffic pattern parameters was changed.

Namely, the large source process model enters the active *session* state very “rarely” (once or twice in the entire simulation run). However, the large traffic source’s sessions were designed to have large amount of packets, with very small packet inter-arrival times (sometimes zero inter-arrival times).

These traffic patterns make the large traffic source very bursty, requiring anywhere from 2 Mbps to 10 Mbps transmission rates depending on the requirements of each individual simulation run.

BS_queue (Base Station)

Process model for the BS_queue is shown in figure 5. In the simulation the BS was designed to operate according to the DTDFP Resource Allocation scheme outlined in the proposed scheduling techniques part of the report.

Namely, the BS is implemented as a FIFO queue with an addition of a scheduling state.

Figure B-5. BS Finite State Machine

Therefore the BS acts as a FIFO queue. However, the BS schedules an interrupt at the end of every time frame (**FRAME_END**), which causes the BS to go into the *schedule* state. The flow chart for the BS scheduling procedure is shown in the figure B-5.

Appendix C

Simulation Validation Details and Assumptions

C.1 Validation

The validation is first started only for ITU A sources, and later the validation model is changed, and modified so that it describes a combination of a various mix of both sources (ITU source A and B).

C.2 ITU Source A Validation

As stated in section 2.1.1 the mean packet call inter-arrival rate for ITU Source A is 1 second, and there is precisely one packet per packet call. This packet call has a mean length of 720 bits, which can easily be transmitted within one time slot even by the MS with smaller transmission rates. This means that, given that one ITU Source A is in an active session it will require one slot on one frame, every 100 frames.

$$E[\text{Interarrival Time}] = 1 \text{ sec.} = 100 \text{ frames}$$

$$\text{For Geometric Dist.} \rightarrow E[x] = \frac{p}{1-p} = 100 \text{ frames} \quad \therefore p = \frac{1}{101}$$

$$p[x \leq X] = \sum_{k=0}^X p(1-p)^k \quad \text{also} \rightarrow p[x=0] = p(1-p)^0 = p = 0.0099 \approx 0$$

Therefore, the probability that one ITU Source A will require a transmission on two adjacent frames (inter-arrival time of zero) will happen with a probability of 0.0099 (1/100, once every one hundred transmissions). Given that transmissions occur on the average every 100 frames, then this is equivalent to one MS requiring transmissions on adjacent frames every 10000 transmissions (almost never). Figure C-1 shows the ITU Source A packet call inter-arrival time pmf.

Figure C-1. ITU Source A Packet Call Inter-arrival pmf

From figure C-1 it can be seen that ITU Source A will require transmissions on adjacent time slots extremely rarely, as well as that times between transmissions will be a "large" (as compared to one frame time) number of frames.

Because of this large packet call inter-arrival time, as well as the fact that sessions arrive according to the Poisson distribution, it is clear that when the BS is observing requests for transmissions on one frame, that those requests will arrive according to the Poisson distribution. The mean Poisson rate of arrival solely depends on the loading of the cell (i.e. number of active MS within the cell). Further validation of this assumption is shown in part 5.3 of the thesis.

C.3 Modeling of the System (given that only ITU Source A's are active in the cell)

Having established that the system consisting of only ITU Source A's behaves as a Poisson random process (when observed on frame to frame basis), then the cell can be modeled by an M/D/1 queue, where arrivals arrive according to a Poisson process, and packets are served according to some constant rate. Then Pollaczek-Khinchin formula can be used to analyze the system in the following way:

Figure C-2. Modeling of the Cell (Only ITU Source A are Active)

$$\begin{aligned}
E[W] &= \frac{\lambda E[S^2]}{2(1-\rho)} \\
E[T] &= \begin{cases} E[S] + E[W] + 0.01/2 & \Leftrightarrow \text{if } E[W] < 0.01 \\ E[S] + E[W] & \Leftrightarrow \text{if } E[W] \geq 0.01 \end{cases} \\
E[Q] &= \lambda E[T]
\end{aligned}$$

Where:

$$\begin{aligned}
E[W] &\Leftrightarrow \text{Expected Waiting Time In Queue} \\
E[S] &\Leftrightarrow \text{Expected Service Time for One Packet} \\
E[S^2] &\Leftrightarrow \text{Second Moment of } E[S] \\
E[T] &\Leftrightarrow \text{Expected Total Waiting Time In Queue} \\
E[Q] &\Leftrightarrow \text{Expected Number of Unserved Packets} \\
\lambda &\Leftrightarrow \text{Poisson Packet Request Arrival Rate} \\
\mu &\Leftrightarrow \text{Packet Service Rate} \\
\rho &\Leftrightarrow \frac{\lambda}{\mu}
\end{aligned}$$

The Pollaczek-Khinchin formula states that $E[T] = E[S] + E[W]$. This is certainly true for an M/D/1 queuing system. The difference in $E[T]$ that arises between M/D/1 queuing system and this scheduling procedure occurs if $E[W] < 0.01$ sec. (one frame time). If $E[W] < 0.01$ sec. and a MS makes a request for transmission on the current

frame, then that MS has to wait until the next frame to begin transmission. Given that arrivals of requests within one frame time are equiprobable and have uniform

$$f(x) = p(x = X) = \frac{1}{0.01}$$

$$E[x] = \int_{-\infty}^{\infty} xf(x) dx$$

$$E[x] = \frac{1}{0.01} \int_0^{0.01} x dx = \frac{0.01}{2}$$

distribution, then the expected additional delay is:

Hence $E[T] = E[W] + E[S] + 0.01/2$ when $E[W] < 0.01$.

Figures C-3 and C-4 show the M/D/1 queue (Pollaczek-Khinchin) approximation of the Cell throughput vs. Delay curve. In figure C-3 the distance between the BS and MS (ITU Source A) was set to 10 meters, and in Figure C-4 it was set to 1610 m. This results in mobile station's average transmission rate of 10.3 Mbps in figure C-3 and 1.5 Mbps in figure C-4.

Figures C-3 and C-4 show that an M/D/1 queue provides close and fair approximation of the real system behavior. The difference between the two curves is greater in figure C-3 than in figure C-4. This stems from the fact that in an M/D/1 approximation it is assumed that there is precisely one packet per transmission time slot, as well as that the entire time slot is occupied by one packet. In reality this assumption is not correct due to the packet length distribution of ITU Source A.

Figure C-3 M/D/1 Approximation of system behavior when only ITU Source As are active

Figure C-4 M/D/1 Approximation of system behavior when only ITU Source As are active (Distance between BS and MS = 1610 m)

References

- [1] Japan's Proposal for Candidate Radio Transmission Technology on IMT – 2000: WCDMA, Japan (MPT), June 1998
- [2] Comparison of RRC and MAC Based Methods of Dynamic Scheduling Between Users on the Uplink, ETSI SMG2 UMTS L2/3, Stockholm, 8-11 March 1999, Agenda item: 6.2, Source: Motorola
- [3] R. H. M. Hafez, T. Kuntz, "IP-Based Radio Services, Evaluation of IP-based Wireless Access," Project Report Submitted to Industry Canada, July 2000.
- [4] H. Lu, "Optimal Allocation of Power to AMCS for Maximum Throughput in WCDMA," Thesis for Master of Applied Science, Ottawa-Carleton Institute for Electrical Engineering, Faculty of Engineering, Department of Systems and Computer Engineering, September 2002.
- [5] A. Leon-Garcia, "Probability and Random Processes for Electrical Engineering, Second Edition," Addison-Wesley Publishing Company, Inc., 1994
- [6] 3G TR25.848 v0.6.0 (2000-05) TSG-RAN Working Group 1 meeting #19, 3rd Generation Partnership Project, Technical Specification Group Radio Access Network, Physical Layer Aspects of UTRA High Speed Downlink Packet Access (Release 2000)
- [7] TSG-RANWG1#17, TSGR1#17(00)1395, "Adaptive Modulation and Coding (AMC)," Motorola, Oct 2000
- [8] 3GPP TSG RAN WG1, TSGR1#18(01)0146, "Complexity of Node B for MIMO Architectures", Lucent Technologies

- [9] R. De Bernardi, D. Imbeni, L. Vignali and M. Karlsson, "Load Control Strategies for Mixed Services in WCDMA," Vehicular Technology Conference Proceedings, 2000. BTC 2000-Spring Tokyo. 2000 IEEE 51st, volume: 2, 2000.
- [10] R. Parsad, W. Mohr, W. Knoch, "third Generation Mobile Communication Systems", Boston: Artech House, 2000.
- [11] T. S. Rappaport "wireless Communications: Principles and Practice," Upper Saddle River, N.J.: Prentice Hall PTR, C 1996.
- [12] 3GPP TSG RANWG1 TSGR1#14(00)0881, "High Speed Downlink Packet Access Simulation Assumptions," Nokia, July, 2000
- [13] M. Haardt, A. Klein, R. Koehn, S. Oestreich, M. Purat, V. Sommer, T. Ulrich, "The TD-CDMA Based UTRA TDD Mode", IEEE Journal on Selected Areas in Communications, Volume 18, No. 8, August 2000
- [14] TSG-RAN Working Group 1 MEETING #16, TSGR1#16(00)1182, "AH24: Link Level Simulation Results of HSDPA," Panasonic, September 2000.
- [15] TSG-RAN Meeting #17, RP-000126, "Detail of High Speed Downlink Packet Access," Motorola, March 2000.
- [16] T.J. Mouldsley, "Throughput of High Speed Downlink Packet Access for UMTS," 3G Mobile Communication Technologies, 26-28 March 2001, Conference Publication.
- [17] S. C. Yang, "CDMA RF System Engineering," 1998 ARTECH HOUSE, INC
- [18] TSG-RAN Working Group 1 Meeting #17, TSGR1#17(00)1377, "HS-DSCH Simulation Results," SONY Corporation, November 2000.

- [19] 3GPP TSG GERAN Adhoc#2, Tdoc GAHW-000041, Oct 9-13, 2000, Agenda item 6.1.3, Munich, Germany, Source: Lucent Technologies, Title: On GERAN Data Link Control Signaling for a Dedicated Channel.

- [20] Antti Toskala, 3GPP TSG RAN WG1 Chairman, Nokia Networks, AMR Mode Command Transmission Options, R1/S4 Meeting, Paris France, Source: Nokia.

- [21] 3GPP TSG GERAN Adhoc, Tdoc GAHW-000021, Munich Germany, Agenda Item 6.1.3, 9th to 13th October 2000, Source Ericsson

- [22] 3GPP TSG RAN WG2 #1, Tdoc RAN WG1 020/99, 20-22 January 1999 Helsinki, Finland, Source: TSG RAN WG2, Date: 21 January 1999.