ADAPTIVE BANDWIDTH ALLOCATION IN FUTURE GENERATION WIRELESS NETWORKS FOR MULTIPLE CLASSES OF USERS

BY

HAITHAM M. ABU GHAZALEH

A Thesis

Submitted to the Faculty of Graduate Studies

In Partial Fulfillment of the Requirements for the Degree

of

MASTER OF SCIENCE

Department of Electrical and Computer Engineering

University of Manitoba

Winnipeg, Manitoba, Canada

TABLE OF CONTENTS

ABSTRACT

Future generation wireless networks are envisioned to provide ubiquitous networking to a wide number of mobile users, promising them the ability to access the various data networks anywhere and anytime. Such networks have motivated the research into efficient management and allocation of the wireless network's limited resources. Heterogeneity also exists amongst the subscribers, i.e. there are those who are willing to spend a little extra on their subscriptions in the prospect of obtaining a better level of service.

The aim of this work is to propose a framework for efficient resource management, with the aim of satisfying the heterogeneous QoS demands of the different subscribers. Part of the proposed framework was used to generate mathematical models for the purpose of analyzing the behavior of the system under two different resource management schemes.

The results obtained from the analysis have shown how the performance of one class of subscribers can influence the performance of the other classes, under a certain resource management scheme.

ACKNOWLEDGMENTS

The author would like to first thank his research adviser Dr. Attahiru Sule Alfa for all his endless support, guidance, encouragement, and efforts in helping with preparing this thesis. This project would have not been possible without him.

A great deal of thanks goes to all the academic professors whom I have been in contact with in the Electrical Engineering department at the University of Manitoba. Their advice and support has proven to be invaluable.

Many thanks goes to all my colleagues who helped me with the preparation of this thesis, and whom I have constantly annoyed!

A special thanks goes to my parents who have done the impossible to help get me to where I am today. All of my achievements have been due to their boundless love and support.

LIST OF FIGURES

1. INTRODUCTION

1.1 General Background

Wireless networks liberate their mobile users from confining their network activities with the fixed wired networks, and allows them to move almost everywhere while continuing with their tasks on the data network. Numerous market surveys have shown that in contrast with today's world business and economy, the ability to communicate as well as be mobile is becoming less of a luxury and more of a necessity (business on the move!). Mobile communication has also become very popular for leisure-use and has attracted the younger generations as a result of the reduced costs, and thus making it affordable to almost everyone nowadays. Other studies in Northern Europe and Japan [1] have found that the number of mobile phone subscribers has begun to exceed the number of fixed-line phone subscribers, indicating the public's increasing demand for mobile communication.

The internet, which is a strong driving force behind this topic, has seen a continuous exponential growth. It has become indispensable for both business and social activities, much like mobile communications. Hence, the trend is for internet access to go mobile.

Various existing wireless technologies such as GPRS, CDPD, IEEE 802.11, and more recently 3G networks, have each made wireless internet access a reality. But none of these existing networks can solely support high data-rates over a wide-coverage [2].

Communication networks have evolved from a voice-centric to a multimedia-centric world [3, 4], as a result of the continuous advances in internet access speeds and network applications. Mobile networks continue to be tailored to the current major source of wireless traffic which is voice. Future wireless networks would need to be developed to support high-speed file transfers, audio and video streaming services, at average transfer rates ranging from 200kbits/s to 2Mbits/s per user, and even higher. This evolution may even prompt the shifting of voice services to the internet through the use of VoIP applications, and users will instead have the advantage of using their IP address as their universal ID for global communication. This would be similar to our use of telephone numbers.

The evolution of mobile multimedia services has expanded the user's expectations to more sophisticated services with QoS (Quality-of-Service) comparable to wire-line access, along with demanding global mobility. This can be accomplished through the integration of several existing wireless networks into what is commonly known as "Fourth Generation" (4G) or "Beyond Third Generation" (B3G) networks [5, 6], since there is no single network technology available to cover the high user expectations. This wide-area system promises to provide users with ubiquitous data services, and aims to deliver higher data-rates as well as the ability to globally roam across the multiple heterogeneous

wireless networks. Effectively, 4G (or B3G) networks will grant their users the benefit of having access to different services, increased coverage, the convenience of a single device, and more reliable wireless access even with the failure or loss of one or more networks. Such networks can also help with satisfying the expected heterogeneous QoS demands of mobile users.

1.2 Previous Work

The topic of efficient resource management has been tackled by many researchers, all of which addressed specific issues when presenting their proposed schemes and models.

The authors in [26, 27] have developed an analytical model for bandwidth allocation to multi-class connections with different QoS constraints. They have considered multi-class connections rather than multi-class users and assumed that all users are entitled to the same fixed amount of bandwidth that is dependent on the type of application being used (e.g. voice, streaming multimedia, web-browsing). Along with assigning priorities to the different class of connections, their policies also considered giving priority to handoff connections over new connections.

In [26], the authors assumed that the total bandwidth in the network under consideration is completely shared amongst all classes of connections, and have further proposed to assign a threshold for defining a maximum occupancy to each connection class for the case of handoff connections only. In [27], the authors proposed to consider the complete partitioning of the total bandwidth for each class of connections, as well as the complete sharing of the total bandwidth amongst all the classes of connections.

The idea of multiple classes of connections with different levels of QoS was also analyzed by the authors in [31] who considered the effects of user-mobility for the different classes of connections. They also considered a fixed bandwidth allocation policy whereby each class of connections are assigned a certain amount of bandwidth. The model developed by the authors studied the case of user-mobility between two neighboring cells,

and have showed how user-mobility can have a great impact on the connection-level QoS.

Another fixed bandwidth allocation policy for multiple classes of connections was proposed in [32] and analyzed for the case of hierarchical cellular networks. The authors focused on the case of a two-layered cellular network with multiple classes of voice connections that may be transferred between the two layers as a result of mobility and overflow. A single class of data connections was also included in the model and was assumed to be admitted only into the macrocell layer. The bandwidth utilized by the existing data connections is flexibly allocated for the purpose of accepting more overflowed connections and achieving a higher system utilization, with the aim being to reduce the blocking of the existing connections.

A fixed bandwidth allocation scheme for multiple classes of subscribers was investigated by the authors in [33] with each class having distinctively different QoS requirements. A multi-cell model was developed by the authors for the purpose of investigating the effects of mobility and the bandwidth allocation scheme on the system's performance for each class of subscribers. The total bandwidth in each cell was assumed to be completely shared amongst all the classes of users. A threshold for each class of traffic is assigned such that successive requests for each class that has utilized a total amount of bandwidth greater than the set threshold is accepted with a certain probability. Such a strategy was claimed to introduce some degree of fairness for those subscribers that are requesting high bandwidths.

An alternative approach to employing fixed bandwidth allocation strategies would be to develop an adaptive bandwidth allocation scheme similar to the ones used by the authors in [34, 35, 36]. The models developed by these authors considered applying adaptive bandwidth allocation for multiple classes of connections in a single cell within a single network.

The adaptive bandwidth allocation scheme that was developed by the authors in [34] considered having each class of connections being allocated the minimum or maximum bandwidth, which is defined by the network for each class of connections, and subject to availability. The extra bandwidth that is utilized by those connections using the maximum bandwidth could be used to further accept more new connection requests from the same class. This is accomplished at the expense of "degrading" a connection with maximum bandwidth to using the minimum instead, and the freed-up bandwidth can be given to the new connection request. A similar approach will be considered in the work done in this thesis. The authors [34] also proposed to allow for the "borrowing" of some bandwidth that is reserved for other classes of connections whenever necessary, assuming that the total bandwidth is completely partitioned for each class.

The idea of bandwidth degradation was also considered by the authors in [35] who further derived a cost function to estimate the total revenue earned by the system employing a particular bandwidth degradation policy. They assumed that the users with a degraded service can be extremely dissatisfied and may eventually result in revenue loss for the service providers. The derived cost function could be used to compute the optimal

degradation policy with the aim of maximizing the net revenue, while attempting to minimize the reduction in user satisfaction as a result of connection degradation. Such results could be used to determine a somewhat "fair" scheme to be used since all users are assumed to be homogeneous in terms of the expected level of QoS and satisfaction. However, it might be inappropriate to apply such results for the case of multiple classes of users with different subscriptions since it is assumed that the more a user pays for his subscription, the higher the expected level of QoS.

In [36], it was proposed to have the existing connections "evenly" degraded, in order to accommodate the connection request of a new user. Such a scheme was claimed by the authors to be "fair" as opposed to randomly selecting and degrading the existing connections. However, the scheme has the disadvantage of increasing the likelihood of a user's connection being degraded. In addition, the proposed fairness assumes that all users should receive almost the same level of service. This assumption would be valid if all users had subscribed to the same level of service, and would not apply in the case for multiple classes of subscriptions. The model in [36] was developed for the purpose of attempting to trace the fluctuations of the received QoS levels of a specific user throughout their connection lifetime. However, the model assumes that the user's connection is initiated when there are no other users in the network.

The majority of the work done in the area of adaptive network resource management was focused on dealing with multiple levels of service (e.g. voice, streaming multimedia, web-browsing), with the aim being to attempt to maintain a certain level of QoS for

each class of connections. In other words, maintaining a certain level of fairness amongst all the users was one of the main criterion in those previous works. However, this is not the case when dealing with multiple classes of subscriptions with different levels of QoS expectations. Furthermore, and to the best of our knowledge, none have addressed the issue of adaptive resource management for multiple classes of subscribers in multiple networks (i.e. 4G or B3G networks).

1.3 Outline

The objective behind this research work is to develop a model for a system of integrated wireless networks which can be accessed by multiple classes of subscribers with varying QoS demands. This model will serve as a starting point for better understanding the behavior of such a system, and would help with further investigating ways of improving the resource management schemes and bandwidth allocation policies in future work.

Before presenting the work on resource management in future wireless networks, a brief description of the network architecture for future generation wireless networks will be given in *Chapter 2*, based on how the authors in [5, 6, 7, 8] envision such networks.

The proposed framework for an efficient allocation of network resources in future wireless networks will be presented in *Chapter 3*. This Section will outline how the network could behave towards the different user demands, and how the different levels of QoS required by every user can be met in accordance with their subscription profile.

A queueing model for some of the schemes discussed in *Chapter 3* will be developed and presented in *Chapter 4*. Some simple cases will be assumed for the purpose of analyzing the performance of the system and exploring its behavior under varying conditions, using the performance metrics defined in the thesis.

Finally, the work will be concluded along with a proposal of the various areas that can be further expanded in future work, and in contrast with the work that has been done so far in the area of adaptive resource management in future wireless networks.

2. FUTURE GENERATION WIRELESS NETWORKS

2.1 Network Architecture

The existing wireless data network technologies can be classified as one of two types: a network that is capable of providing high-speed connections within a limited area, or a network that is capable of servicing a wide geographic area but at low data rates. An example of the two classes of network technologies are GPRS in Cellular networks with wide coverage areas that support low data-rates, and IEEE 802.11 WLANs that support high data-rates and serve a limited number of users over a small coverage area, respectively. However, there is no single network technology that is good enough to replace all other technologies combined.

Rather than put efforts into developing new radio interfaces and technologies to expand the service coverage and data-rates of the existing wireless networks, a more feasible option would be to seamlessly integrate the existing wireless technologies onto a common platform. The result would be the unification of several heterogeneous networks of varying coverage and performance into a single logical IP-core network, with its overall coverage being the union of the networks' coverage. This approach has the advantage

of utilizing the already-established wired and wireless infrastructure without any need to replace the current generation. Such networks will also have the flexibility of further incorporating any newly developed wireless technologies into the Hybrid network.

The different wireless networks are allocated to different layers in a hierarchical manner, with respect to the network's cell size, coverage, and user-mobility, providing a globally optimized seamless service to all the users. An example of such a Hybrid network is illustrated in *Figure 2.1*, which shows how the coverage area of one network can overlap the other [7, 8]. The diagram illustrates a simple example of only two in the hierarchy, with the upper layer being the low data-rate cellular networks that provide a wide coverage area for a large number of users (e.g. GPRS), and the lower layer being the high-speed data networks that serves a limited number of users over a small area (e.g. WLANs). The base stations and wireless access points are assumed to have a link with the wired data network infrastructure (i.e. the internet). The structure of the Hybrid network can consist of more than two layers; for example, an additional layer above the GPRS network could be a satellite wide area data network, and an additional layer below the WLANs could be one that represents Ad-Hoc networks [11, 12, 13].

In such a configuration, the cells in a mobile cellular network will be used to provide "blanket coverage" for the wireless LAN pico-cells. Wireless LAN technologies [9] are quite attractive for dense urban environments, i.e. places where there is a high demand for multimedia applications, such as hotels, airports and shopping malls. They serve as a cost-effective complement to the Cellular networks in terms of increased QoS (data-

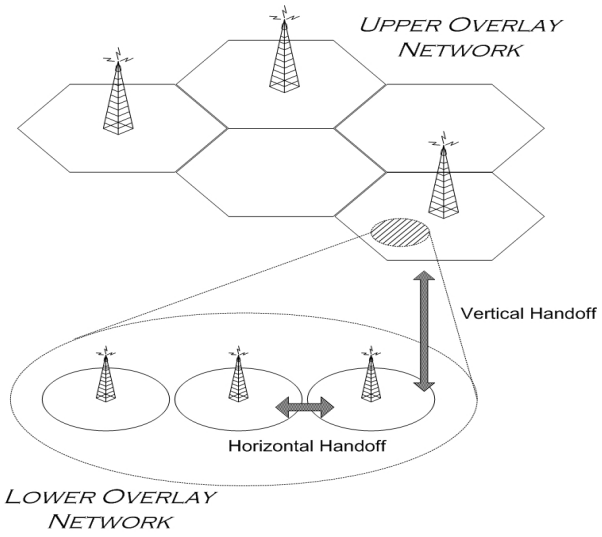

Fig. 2.1: The Hybrid Wireless Network

rate, performance, etc.) for multimedia applications accessed in dense urban areas where mobility is relatively low. The Hybrid network architecture will also allow the mobile users to freely move between the wireless networks, while maintaining their running applications in a manner that may satisfy their QoS requirements [10].

The deployment of 4G systems will therefore allow the Cellular networks and wireless LANs to coexist together, with each playing a complementary role. The result is a wide-area coverage network that is able to provide users with ubiquitous data services ranging from low to high data-rates in strategic locations.

2.2 Coupling of the Existing Wireless Networks

The authors in [14, 15] have looked at two different ways of integrating the different types of wireless network technologies into a single network. In a *Tightly-Coupled Interworking*, the authors propose to have the WLANs integrated into the Cellular (or 3G) networks, such that these WLANs will behave as other access points on the Cellular network. This would require placing a WLAN gateway on the access point between the two networks, which has the function of hiding the WLAN protocols from the Cellular network. The users under the coverage and service of the WLAN can use their own protocols for communication, but this would only be valid for the data traffic within the network. Any traffic that is intended for nodes outside the WLAN would require the implementation of the Cellular protocols.

A significant problem encountered with such an integration is the increased traffic load that the Cellular network must sustain, since the majority of the traffic from the WLAN will need to be injected into the Cellular network (WLAN nodes are likely to communicate with other nodes outside their network). This makes it necessary for Cellular network operators to modify the entire network to withstand such high loads of traffic.

A more favorable approach would be to implement a *Loosely-coupled Interworking* architecture, with the WLAN gateway being directly connected to the core network (internet), with no direct link to the Cellular network.

2.3 Mobility & Vertical Handoff

4G networks will grant their mobile users the chance to seamlessly transfer their ongoing sessions from one network to another, once a "better" network becomes available, or the connection with the current network is about to be lost. Traditionally, Cellular networks allow mobile users that are roaming between the cells to transfer their connections from one cell to another through a process known as *handoff*. With the emerging of 4G systems, this will now be termed as *Horizontal handoff*, which is defined as the process of transferring a mobile user's connection between base stations (BS) in the same network.

For users that are transferring their connections between different networks as a result of a "better" network becoming available, they will have to undergo what is known as *Vertical Handoff* [7, 18]. This type of handoff could also avoid having an ongoing connection being lost as a result of the user having exited the service coverage of its current network.

Vertical handoff is executed between BS that are using different wireless network technologies, and occurs when a different network becomes *consistently* better than the current network in terms of the offered bandwidth. A significant change in QoS will likely be experienced by the users and will affect the performance of both upper-layer protocols and applications. The level of impact would depend on the type of networks that are involved in the connection-transfer process, along with the user's running application. Synchronization [16] of the sessions during the network transfer process should also

be considered, in order to eliminate the possibility of packet-losses [17] due to vertical handoff. Multi-mode terminals [19, 20, 21] are being equipped with multiple network interfaces and have the responsibility of seamlessly initiating the appropriate type of handoff whenever necessary.

Vertical handoff can occur in two different ways; an *Upward* and *Downward* vertical handoff. A Downward vertical handoff is initiated as a result of a Mobile Host (MH) moving into the coverage of a lower overlay network with a higher date-rate. The MH would proceed to switch its connection from the higher (Cellular) to the lower (WLAN) overlay network. There is no risk of performance loss with this type of handoff since the connection with the upper overlay network is never lost during the process.

An Upward vertical handoff is when the MH switches its connection from a lower (WLAN) to a higher (Cellular) overlay network. This occurs when the MH moves out of the coverage of the lower overlay network, and attempts to maintain its connection with the data network by switching to an upper overlay network. This type of handoff can be very critical since the MH may lose its connection with the data network for some period of time while switching between the overlays, which can severely impact the MH's performance, especially if the process is done while the MH is undergoing data transfer. A late decision in initiating upward vertical handoff can degrade the performance of an on-going session. Therefore, this type of handoff is of crucial importance for reliable operation, and any proposed scheme should not be limited to reactions towards link-disconnections.

The general policy of having the MH's connection attempting to always associate with the lowest reachable overlay network (due to the high performance offered) tends to ignore some system dynamics, such as the current traffic load of the network [23, 24, 25], cost of using the network, and the power consumption of network interface . Such metrics should be exploited in a suitable manner for determining the "best" network within the MH's reach [22], along with deciding on the type of handoff to initiate.

3. THE PROPOSED FRAMEWORK FOR RESOURCE ALLOCATION

Each type of wireless network is tailored for a specific group of applications and promises to provide its users with a particular level of QoS. Due to the diversity of network applications becoming available, each with their own QoS measures, future generation networks propose to introduce a variety of QoS levels to support the heterogeneous requirements of their mobile users. This Chapter will focus on developing the framework for an efficient resource allocation scheme, along with describing how such networks should adapt their available resources to satisfy the various QoS demands for each user among the networks. The framework in this Chapter will be presented for the case of a single network. It is assumed that the same resource management scheme will be applied in all of the networks of the 4G (or B3G) systems.

The main resource to be considered here is the amount of bandwidth that should be allocated to each user. The aim is to maximize the usage of these resources, while satisfying the heterogeneous QoS demands of different users. The proposed scheme attempts to also maximize the number of on-going connections in the network without interfering with the minimum QoS demands of each user.

The authors in [26, 27, 28] have proposed assigning a *Service-Class* to each user as a starting point for satisfying the heterogeneous QoS demands. Each class defines the level of the user's QoS demands based on the user's running application. In other words, the more bandwidth-sensitive a user's running application is, the more the bandwidth that should be guaranteed to that particular user by the network. However, in this work, the focus is instead on multiple subscriptions rather than multiple services, whereby a user subscribes to a particular connection class which defines its level of QoS demands, based on QoS contracts between the user and the network. Therefore, the more the user pays for a higher class subscription, the more the bandwidth that the network should guarantee that particular user, on average. The motivation behind employing multiple class subscriptions is that differential subscriptions (or pricing) could be a key tool through which the network can induce efficient use of network resources [30], while keeping high performance services available for those who require it.

3.1 Classification of Users

We begin by assuming that there are N different classes of users, such that $1 \leq N < \infty$. Each user is assigned a particular class C_i , where $1 \leq i \leq N$, which is based on the user's type of subscription. Therefore, the set of classes that are available for subscriptions are $\mathbf{C} = \{C_1, C_2, C_3, \ldots, C_N\}$. Moreover, it will be assumed that a class C_i user has a higher class than a class C_{i-1} user, with C_N being the highest class and C_1 the lowest. This would imply that the network should attempt to guarantee a C_i user with a QoS that is better than or at least similar to that of a C_{i-1} user. In other words, C_i users are guaranteed to be served with a QoS that is at a minimum equivalent to the QoS for C_{i-1} users.

In general, it will be assumed that $\underline{\mathbf{u}}\,(C_i) \;\geq\; \underline{\mathbf{u}}\,(C_{i-1})$, where $\underline{\mathbf{u}}\,(x)$ in utility theory is a measure of the satisfaction gained from a good or service x. This assumption implies that users with a higher subscription class are expecting an overall better performance than those with lower class subscriptions, and at no time will they experience a performance lower than what is experience by the users with a lower class subscription. Alternatively, the network operators may choose a subscription class policy with $\underline{\mathbf{u}}\,(C_i) \;>\; \underline{\mathbf{u}}\,(C_{i-1})$.

Consider the example where $N = 4$. In this case, the set of available classes are given by $\mathbf{C} = \{C_1, C_2, C_3, C_4\}$. A suitable label can be associated with each of the different classes, namely { *Bronze, Silver, Gold, Platinum* } , with the *Platinum* class of users having a higher subscription class than the others.

3.2 Type of Mobile Terminal

The type of terminal used to access the network will also play a vital role in determining the level of QoS that is required by each user. Some terminals are limited by the types of services they can offer and excessive bandwidth would be wasted, e.g. limitations of terminal performance may be due to memory capacity, screen resolution, etc.

It is assumed that there are M different types of terminals that can be used to access the network, with $1 \leq M < \infty$. Any user in class C_i can access the network through any of these terminals. Therefore, we have that each user in class C_i is utilizing a terminal D_j, where $1 \leq j \leq M$, with D_j terminals having at the least a similar performance to the D_{j-1} terminals. To illustrate, consider the example where $M = 2$. The two different types of terminals could be described as follows,

D_1 = a low performance terminal, such as a PDA or cellular device.

D_2 = a high performance terminal, such as a laptop.

In general, it will be assumed that $\underline{\mathbf{u}}(D_i) \geq \underline{\mathbf{u}}(D_{i-1})$, with the choice of having the equality being left up to the service providers.

A user with subscription class C_i and utilizing a D_j-type terminal is said to have the profile $\{C_i, D_j\}$. Note that the user-profile $\{C_i, D_j\}$ can also be thought of as another set of classes C_z, where $1 \leq z \leq (M \times N)$.

3.3 Allocation of Bandwidth Units

As mentioned earlier, the main network resource that affects the QoS perceived by the users is the amount of bandwidth that is made available to each user. The total bandwidth in each cell of the network can be divided into a discrete number of equal units U, which is similar to the approach found in [26, 27, 28, 31, 32, 33, 34, 35, 36]. Each U represents a particular amount of bandwidth which is analogous to a wireless channel, e.g. 1 $U = K$ MHz, where K is some constant defined by the network operators. A certain number of units can be assigned by the network to a subscriber in accordance with its utilized terminal and class.

A maximum amount of bandwidth units U^{max} can be allocated by the network to each subscriber with a class C_i subscription and terminal D_j. For a subscriber with the profile $\{C_i, D_j\}$, let U^{max} be $U^{max}_{i,j}$. A minimum U^{min} will also be set and guaranteed by the network for each subscriber, with U^{min} being $U^{min}_{i,j}$ for a subscriber with the profile $\{C_i, D_j\}$.

Hence, all users with subscription class C_i utilizing a terminal type D_j is expecting to be allocated $U_{i,j}$ units, subject to availability, where,

$$U^{min}_{i,j} \leq U_{i,j} \leq U^{max}_{i,j} \tag{3.1}$$

The following properties apply to the definitions given:

- $U_{i,j}^{max} \geq U_{i,j}^{min}$

  (where $U_{i,j}^{max} = U_{i,j}^{min}$ implies that a fixed bandwidth allocation policy is employed).

- $U_{i,j} \leq U_{i+1,j}$ assuming that $\underline{\mathbf{u}}(C_i) \leq \underline{\mathbf{u}}(C_{i+1})$

  implying that higher class of users are assured at the least the same amount of units as those for the lower class.

- $U_{i,j} \leq U_{i,j+1}$ assuming that $\underline{\mathbf{u}}(D_i) \leq \underline{\mathbf{u}}(D_{i+1})$

  implying that higher performance terminals are assigned at the least the same amount of units as those for the lower performance terminals.

An example of how the network can allocate a number of these units U to the subscribers with different classes and terminals (with N=4 and M=2) is given in *Table 3.1.* A further assumption is made in the example given in *Table 1* which is that $U_{i,j}^{min} = U_i^{min}$, for all j.

Class (C_i)	**Bronze**(C_1)	**Silver**(C_2)	**Gold**(C_3)	**Platinum**(C_4)
$U_{i,2}^{max}$	3	4	5	6
$U_{i,1}^{max}$	2	3	4	5
U_i^{min}	1	2	3	4

Tab. 3.1: Resource (Unit) Allocation for Different Classes of Users

What *Table 3.1* shows is the maximum and minimum U a particular user is guaranteed to have by the network based on the user's QoS contract. For example, C_1 users with terminal D_2 have their $U^{max} = 3$, i.e. network guarantees that class of user a maximum of 3 units, whenever possible. A user should not concern itself with the network traffic conditions. It should simply expect to connect with the network at a rate

between the maximum and minimum allowed for his class of service, and relative to the type of terminal being used. The reason for having a lower U^{max} for each class of users that are using terminal D_1 is due to the limited resources of that terminal compared with that of terminal D_2. Therefore, the excessive U that would have been given to D_1 (if it was made equal to the U^{max} allowed for D_2) would be wasted by the network and of little benefit to the user.

If we consider the 4G (or B3G) network architecture, the value of each unit U may be different for every network in the hierarchy of 4G, but the number of units that have been contracted can remain the same for each network. For example, the value of the bandwidth unit in a cellular network can be $U = K_1$ MHz, whereas $U = K_2$ MHz in a WLAN, such that $K_2 \geq K_1$. Also note that *Table 3.1* shows a linear increase in the number of units between the different classes of users. But the argument can be further extended to display any other non-linear increase in the units between the classes, which is a decision that remains under the control of the network operators.

3.4 Bandwidth Allocation Policy

Upon initial connection, the network is expected to grant a class C_i user that is utilizing a D_j terminal with $U_{i,j}$ units. The amount $U_{i,j}$ would depend on the number of units that are free and available in the network. This amount may change throughout the entire duration of the user's ongoing session, when compared to what was initially allocated by the network. For example, a user may initially obtain 3 units, which may go down to 1 unit at a later time. Hence, the users should expect their allocated units to be increased (upgraded) or decreased (degraded) throughout the session, subject to availability and policy. This would imply that the network is allowed to control the number of units that are allocated to each user subject to the specific allocation policy that is employed, and provided that the changes are within the limits given by (3.1).

The initial allocation of units for a new connection-request can be handled by the network in mainly two different ways. One approach would be to initially allocate the user with $U_{i,j}^{min}$ units, and then proceed to increase the number of units whenever possible. This depends on the amount of units that are freely available to the users in the network. The other approach would be to have the network initially grant the new user with $U_{i,j}^{max}$ units, or as close as possible to $U_{i,j}^{max}$ if the maximum cannot be allocated. The approach of initially attempting to obtain $U_{i,j}^{max}$ (or as close to it as possible) upon initial connection , has the advantage of maximizing the usage of the available resources, along with user satisfaction. The alternative approach of trying to first obtain $U_{i,j}^{min}$ upon initial connection, and then proceed to have the connection

upgraded to a number of units not greater than $U_{i,j}^{max}$, may prevent the user from the benefits gained by using a higher number of units. This could also be true even if the network increases the units to the maximum for that particular user after the initial connection is made. The reason being that if we assume a TCP connection, then TCP may not rapidly converge towards the new optimum after increasing the allocated bandwidth. Therefore, this type of approach is *not* maximizing the available resources and may yield to wastage of resources allocated.

3.4.1 Connection Degradation Process for Multiple Classes

If the user could not even obtain $U_{i,j}^{min}$ upon initial connection due to the high traffic and unavailability of units in the network, then the network will undergo the process of *"degrading"* those connections that are using more units than their minimum, until enough units are available to be given to the new user. The network should only look at freeing-up enough units to grant the new user with $U_{i,j}^{min}$ units. This has the benefit of reducing the number of existing connections to be degraded. The policy may also be modified to free-up enough units to grant the new user with higher number of units, but such a policy suffers from the disadvantage of increasing the likelihood of having the existing connections being degraded. If the network is unable to free any/enough units to give to the new user, then the new user's connection-request will be blocked.

The network resource (or total bandwidth units) could be partitioned in such a way that each class of subscribers have access to a separate pool of bandwidth units. The

partitioning of the resources allows the network to treat each class of subscribers independently. Hence, only those users from the same class as the new user are selected in the connection degradation process. The network resource could instead be completely shared by all the classes of subscribers. In this case, the connection degradation process should also consider selecting users from classes other than the class of the new request. One approach would be to have the network begin with degrading the lowest class of users first and then work their way up to the highest class, in the attempt of freeing enough units for a new user. Therefore, the network will start degrading C_1 users whenever possible, and then proceed to degrade the class C_2 users once all the users of class C_1 are only given $U_{1,j}^{min}$, and so on.

While the term *connection degrading* was used to indicate a reduction in the utilized bandwidth units (as was done by many authors such as [34, 35, 36]), it does not necessarily imply that user-satisfaction is at a risk. The reason being that unlike the assumptions given by the previous authors, the proposed policy assumes that user classification is subscription-based, and a lower level of service could be considered acceptable by the user, as long as it does not violate the QoS contract between the user and the network. A higher number of units could be considered as a *bonus*, and is made available whenever possible and in accordance with the allocation policy.

3.4.2 Connection Degradation Process for Multiple Terminal-Types

Within each class of subscribers, there are also those that are using different types of terminals. For M different types of terminals, there are M different groups of users with the same subscription class. This indicates that the connection degradation policy should take into consideration the fact that the type of terminal could influence the selection process. The network operators may choose to assign a selection process that does not distinguish between the different groups of users with the same subscription class. In other words, all users with the same subscription class are equally likely to be selected to have their connections degraded. Alternatively, users with lower-type terminals could be selected to have their connections degraded with a higher priority over those users that are using "better" terminals.

To constantly select one group in preference to the other would mean that the other group would eventually dominate, i.e. its best to always use a particular type of terminal to increase the chances of keeping U^{max} throughout the connection. It may seem plausible to apply such a scheme since users with high-performance terminals are more likely to be running applications that are bandwidth-sensitive, when compared to those users that are utilizing low-performance terminals. Nevertheless, such a scheme will introduce some unfairness into the selection process.

3.4.3 The Degradation Process for Multiple Classes and Terminal-Types

This sub-Section proposes a degradation policy that combines the policies described in the previous sub-Sections, along with some further developments. To allow for some degree of fairness amongst all of the users with the same subscription class, it is proposed to allow the network to select a user at random from the class of C_i users that are utilizing D_j terminals with a probability $p_{i,j}$, such that $\sum_{j=1}^{M} p_{i,j} = 1$, $\forall i$.

Consider the example with $N = 4$ and $M = 2$. *Table 3.2* shows the probabilities and order of selection for connection degradation, assuming that the network's resource is completely shared amongst all the classes of users, and given that the priority in selection is given to lower class subscribers. Hence, the network will begin to degrade the C_{i+1} users after all the C_i users are using $U_{i,j}^{min}$.

	D_1	D_2
Bronze(C_1)	$p_{1,1}$	$p_{1,2}$
Silver(C_2)	$p_{2,1}$	$p_{2,2}$
Gold(C_3)	$p_{3,1}$	$p_{3,2}$
Platinum(C_4)	$p_{4,1}$	$p_{4,2}$

Tab. 3.2: Probabilities for Selection of Performance Degradation

Therefore, the network should first look at C_1 users and choose to degrade those that are using terminal D_2 at a probability of $p_{1,2}$ and those that are using terminal D_1 at $p_{1,1} = 1 - p_{1,2}$, until all the users are allocated $U_{1,j}^{min}$. The network would then proceed in the same manner with C_2 users, then C_3, and finally C_4 users, until all of them are allocated U^{min}. Any further connection requests would be blocked or queued since no more units are available. This selection process is assumed to be independent of the

class and terminal-type of the new request that is causing the network to execute the connection degradation process.

The assignment of the probabilities $p_{i,j}$ could be further assumed to be independent of the class of subscriptions, i.e. $p_{i,j} = p_j \ , \forall \ i$. Therefore, using the previous example, the probability of selecting a user with a D_1 and D_2 terminal to have its connection degraded is given by p_1 and p_2, respectively, and regardless of the subscription class.

The probabilities $p_{i,j}$ are assumed to be statically assigned by the network operators. However, these probabilities need not necessarily be fixed and could instead be adaptive with respect to certain conditions. An example would be to define $p_{i,j}$ as a function of the user's elapsed connection-time t in the system, i.e. how long the user has been in the system. The function $p_{i,j}(t)$ could also be developed in such a way that the probability $p_{i,j}$ increases with t, implying that the network is more likely to select the user with a high t to degrade his performance. $p_{i,j}(t)$ can assume any functional form, as long as it satisfies the following conditions,

$$p_{i,j}(t_2) \ \geq \ p_{i,j}(t_1) \quad for \quad t_2 \ > \ t_1 \tag{3.2}$$

$$p_{i,j}(t) \mid_{t \to \infty} \ \to 1 \qquad and \qquad 0 \leq p_{i,j}(t) \leq 1 \tag{3.3}$$

In this case the network would have to select at random a user with D_j , for all j, and determine each of their elapsed connection times t. The relevant probability functions

for each of those selected users would then have to be evaluated by the network, and the one with the highest probability would be selected to have its connection degraded.

In addition to computing the probability functions $p_{i,j}(t)$, a weight factor w_j could also be incorporated into those functions. These weights could be assigned by the network operators in such a way that favors the selection of a certain group of users (using a particular type of terminal) over others. For example, the network may be partial to selecting users with D_2 terminals to have their connections degraded over those that are using D_1 terminals. The reason being that degrading the connections of those subscribers using D_2 may free-up more units than those obtained from choosing to degrade the connections of subscribers using D_1. This could allow for the degrading of fewer users. Such an approach may even prove to be socially appealing since it attempts to degrade the minimum number of connections.

3.4.4 Connection Upgrade

The upgrading of the existing connections in the network would only be allowed if one or more bandwidth units become available as a result of a service completion, and assuming that there are no new requests that require those units. The network would also have to decide on which group of users are to be selected for upgrading their existing connections. Such a decision could be based on the number of units that become available, and it can also take into consideration the different levels of priority amongst the different subscription classes and terminal-types. Once the network has

decided on which group of users with the subscription profile $\{C_i, D_j\}$ are to be selected for the connection upgrade, the choice of a particular existing connection is assumed to be made at random.

To allow for the existing connections to be upgraded may aid in promoting maximum utilization of the network resource. The allowance of connection degrading and upgrading will likely introduce fluctuations in the QoS levels experienced by the existing users [36]. These fluctuations could also be a burden on the network's processors. Hence, the connection degrading/upgrading policy might have to be tuned to certain requirements that are relevant to a particular network. The issue of "fairness" in selecting which connections to have upgraded need also to be considered.

3.5 System Description

In general, the system we are looking at considers the number of users in a single network for each class, with respect to the type of terminal used and units allocated. The state Q of the system is given as,

$$Q = \{X_{i,j,k} \ ; \quad 1 \leq i \leq N, \quad 1 \leq j \leq M \ , \quad U_{i,j}^{min} \leq k \leq U_{i,j}^{max}\} \tag{3.4}$$

where $X_{i,j,k}$ is the number of users in the network with class C_i, using terminal D_j, and currently assigned k units of bandwidth. Let T_Q be the total amount of resources (i.e. bandwidth units) that is utilized by the users in state Q. Moreover, due to the limited amount of bandwidth available, and in conjunction with the proposed resource allocation strategy, the set of all possible states Q in the state space $\mathbf{S}$ must satisfy the following condition,

$$\mathbf{S} \ = \ \{ \ Q \ : \ T_Q \ \} \qquad with \qquad T_Q \ = \ \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=U_{i,j}^{min}}^{U_{i,j}^{max}} k.(X_{i,j,k}) \leq B_{TOT} \ , \tag{3.5}$$

where B_{TOT} is the total number of units available in the network for all connections.

3.5.1 Examples of System Transitions

For illustrative purposes, a simple case will be examined with $N = 2$ and $M = 2$. In addition, it will be assumed in the example that the network has assigned the same

unit allocations for C_1 and C_2 subscriptions, given in *Table 3.1*, i.e. $U_{1,1}^{min} = U_{1,2}^{min} = 1$, $U_{1,1}^{max} = 2$, $U_{1,2}^{max} = 3$, $U_{2,1}^{min} = U_{2,2}^{min} = 2$, $U_{2,1}^{max} = 3$, and $U_{2,2}^{max} = 4$. The total number of units, B_{TOT} , is also assumed to be completely shared by both classes of users. Therefore, we have the following general system state

$$Q = \Big(X_{1,1,1} \ , \ X_{1,1,2} \ , \ X_{1,2,1} \ , \ X_{1,2,2} \ , \ X_{1,2,3} \ , \ X_{2,1,2} \ , \ X_{2,1,3} \ , \ X_{2,2,2} \ , \ X_{2,2,3} \ , \ X_{2,2,4} \Big) \tag{3.6}$$

If the system is in a state where the network has enough units available to grant each user with $U_{i,j}^{max}$, then we have a system with a general state Q_n given as:

$$Q_n = \Big\{ 0 \ , \ X_{1,1,2} \ , \ 0 \ , \ 0 \ , \ X_{1,2,3} \ , \ 0 \ , \ X_{2,1,3} \ , \ 0 \ , \ 0 \ , \ X_{2,2,4} \Big\} \tag{3.7}$$

with $T_{Q_n} \leq B_{TOT}, \quad X_{1,1,2} \geq 0, \quad X_{1,2,3} \geq 0, \quad X_{2,1,3} \geq 0, \quad X_{2,2,4} \geq 0.$

The state Q_n implies that every user in the system is getting $U_{i,j}^{max}$ units for their connection. The other variables given by 0 are necessarily zero, until the network has no more enough units for allocating U^{max} to any new user.

If the system with state Q_n has enough units to accommodate the needs of a few more new connection-requests with $U_{i,j}^{max}$, then the further arrival of another user of class C_1 using terminal D_1 causes the state transition of $Q_n \rightarrow Q_{n+1}$, where

$$Q_{n+1} = \Big\{ 0 \ , \ (1 + X_{1,1,2}) \ , \ 0 \ , \ 0 \ , \ X_{1,2,3} \ , \ 0 \ , \ X_{2,1,3} \ , \ 0 \ , \ 0 \ , \ X_{2,2,4} \Big\} \tag{3.8}$$

with $T_{Q_{n+1}} \leq B_{TOT}$.

This shows that the new connection request will also be allocated $U_{1,1}^{max}$ for its class, due to the sufficient availability of units.

Alternatively, the system with state Q_n could be in such a state that there is not enough units available to grant a certain new connection request with $U_{i,j}^{max}$. However, it still could be in a state of granting U^{max} units for further requests with lower classes of subscriptions or terminal-types. If we assume that for a certain connection-request with class C_i and terminal D_j , the network with state Q_n has insufficient units to grant that new request with $U_{i,j}^{max}$, then one of the two possible outcomes can occur. The network could attempt to allocate the new connection-request with $U_{i,j}$ units, within the limits given by *(3.1)*, such that

$$T_{Q_n} + U_{i,j} \leq B_{TOT} \tag{3.9}$$

with $U_{i,j}$ being assigned as close to $U_{i,j}^{max}$ as possible. This outcome can only occur if $T_{Q_n} < B_{TOT}$. The other possible outcome occurs when the system with state Q_n is in such a state where even $U_{i,j}^{min}$ cannot be allocated to the new request, such that

$$T_{Q_n} + U_{i,j}^{min} > B_{TOT} \tag{3.10}$$

In this case, the system would have to execute the connection degradation process. Consider the example where the condition given by *(3.10)* is true for the system with state Q_n, and a new user with class C_1 using terminal D_1 arrives into the system. The

following system state transition $Q_n \rightarrow Q_{n+1}$ could occur, where

$$Q_{n+1} = \Big\{ (X_{1,1,1} \ + \ 2) \ , \ (X_{1,1,2} - 1) \ , \ 0 \ , \ 0 \ , \ X_{1,2,3} \ , \ 0 \ , \ X_{2,1,3} \ , \ 0 \ , \ 0 \ , \ X_{2,2,4} \Big\} \quad (3.11)$$

with $T_{Q_{n+1}} \ \leq \ B_{TOT}$.

Initially $X_{1,1,1} = 0$, and the arrival of the new user has caused the network to select one user from the group $X_{1,1,2}$ to have its connection degraded (assuming that $X_{1,1,2} > 0$), and thereby reducing the number of users in that group by one. The user with the degraded connection is added to the group $X_{1,1,1}$ along with the new user.

The further arrival of a new class C_2 connection-request with terminal D_1 into the system with state Q_{n+1} could trigger the network to select a few existing class C_1 connections to be degraded, assuming that the network has insufficient units to grant the new request with $U_{2,1}^{min}$ units. If the connection degradation policy requires that only the minimum amount of units be given to the new connection-request, i.e. $U_{2,1}^{min}$, then the network needs only to free-up two units for the new request. These units can be obtained by either degrading the existing connections of two users from the group $X_{1,1,2}$ to $X_{1,1,1}$ (assuming that $X_{1,1,2} \geq 2$), or one user from the group $X_{1,2,3}$ to $X_{1,2,1}$ (assuming that $X_{1,2,3} > 0$), or one user from the group $X_{1,2,3}$ to $X_{1,2,2}$ (assuming that $X_{1,2,3} > 0$) and one user from the group $X_{1,1,2}$ to $X_{1,1,1}$ (assuming that $X_{1,1,2} > 0$). The choice from either options will depend on the degradation scheme adopted by the

network. If the second degradation option was selected, then the new system state transition $Q_{n+1} \rightarrow Q_{n+2}$ would be,

$$Q_{n+2} = \Big\{ \quad X_{1,1,1} \ , \ X_{1,1,2} \ , \ (X_{1,2,1} \ + \ 1) \ , \ 0 \ , \ (X_{1,2,3} \ - \ 1) \ , \ (X_{2,1,2} \ + \ 1) \tag{3.12}$$

$$X_{2,1,3} \ , \ 0 \ , \ 0 \ , \ X_{2,2,4} \ \Big\} \tag{3.13}$$

with the other variables remaining unchanged from state Q_{n+1}, and $T_{Q_{n+2}} \leq B_{TOT}$.

3.6 Distribution of The Total Network Resources

The complete sharing of the total bandwidth units amongst all the classes of subscribers in the network may introduce some degree of unfairness to the users in the system. For example, networks with many C_N users (highest class of subscription) will likely occupy the majority of units available, and leave very little for the rest. One way of overcoming this problem of unfairness is to reserve a certain number of units for each class of users, and thereby guaranteeing that the network will be able to accommodate at least a particular number of users for each class. This can be accomplished through the complete partitioning of the total resources for each class of users. The partitioning of units could also be assigned by the network in a such way that is relative to certain conditions, e.g. the assignment may be based on the total traffic of users of all classes in the nearby cells.

3.6.1 The Partitioning and Borrowing of the Network Resources

The higher class users have opted to pay more for their subscriptions, and in return are relying on the network to provide them with a performance that should be better than those in the lower classes. Hence, the network should focus more of its attention to those users, while simultaneously trying to satisfy the needs of the lower class users. If the higher class users find that they are getting less attention than what they expect, or even getting similar performances to those in the lower classes, then they are led to believe that the extra amount that they have paid in their subscriptions are worthless,

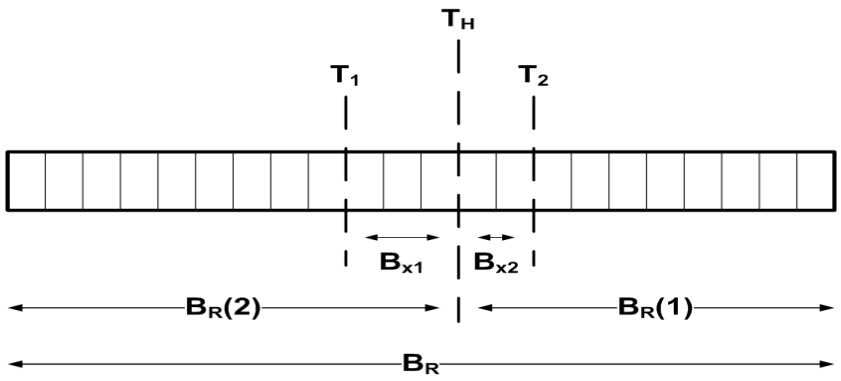

Fig. 3.1: Resource Partitioning and Borrowing for Two Classes of Users

and the network may risk losing a "valuable" client. To avoid such consequences, it is necessary for the network to distribute a good fraction of its resources to the higher class users, with the possibility of having some of these unused resources "borrowed" [34] by the lower class users, rather than being wasted.

To help explain the proposed method for resource partitioning and borrowing, consider the simple case of having only two different classes of users in the network, namely C_1 and C_2. Assume that out of the total number of units B_{TOT} that are available to all the users, a fixed number of B_R units in total have been reserved, such that $B_R \leq B_{TOT}$. These B_R units are partitioned in such a way that $B_R(2)$ is reserved for C_2 users and $B_R(1)$ for C_1 users, such that $B_R = B_R(2) + B_R(1)$, as illustrated in *Figure 3.1*. The unreserved units of B_{TOT} are assumed to be completely shared by all classes of users, with the reserved units behaving as guard channels. If $B_R = B_{TOT}$, then the total network resource is said to be *completely* partitioned.

Let T_H in *Figure 3.1* be the dividing partition that splits the total reserved units.

T_H may remain fixed or could be periodically adjusted in manner that depends on the traffic of each class of users. It is not necessary for each cell in the network to have an equal T_H. For example, the assignment of T_H could be in favor of allowing more C_2 users in a location where they are most likely to be heavily populated.

T_H could also be temporarily extended up to T_2 for C_2 users, or up to T_1 for C_1 users, as seen necessary by the network. This could occur at a time when: (i) the network requires more than $B_R(2)$ units to serve the available C_2 users, while not all of the $B_R(1)$ units are used up, or, (ii) the network requires more than $B_R(1)$ units to serve the available C_1 users, while not all of the $B_R(2)$ units are used up, respectively. This type of behavior is described as "borrowing" units from the pool of units reserved for other classes of users.

For the case of C_1 users, even though there is a certain number of units $B_R(1)$ that are reserved for them (given by T_H), a further number of extra units from the set of units $B_R(2)$ can be "borrowed" by the C_1 users whenever needed. A maximum of B_{x1} units can be borrowed provided that no C_2 users are using up those extra units. This approach may offer an efficient way of utilizing any of the unused units in B_{x1} by the C_2 users, in the effort of serving as many of the C_1 users as possible, and without jeopardizing the performance of the C_2 users. However, once the traffic of C_2 users reaches a level where they require those B_{x1} units that are temporarily utilized by the C_1 users, the network should then have to force some of the connections of the C_1 users to be immediately dropped. The network would have to select a number of C_1 users at random to have their

connections dropped, with the number of users corresponding to the number of units that need to be returned to B_2. The selection of C_1 connections to be dropped could also be based on certain parameters such as *"residual connection duration"*, whereby the one with the most (or least) would be chosen to have his connection to be instantaneously dropped.

A similar approach applies to the case of having excessive C_2 users exploiting the extra B_{x2} units whenever possible. However, in this case, when the C_1 users require the use of the borrowed B_{x2} units, it might be preferable to have the network wait until the C_2 user's connection is completed (provided that it does not exceed a certain length of time) before it can return the "borrowed" units back to $B_R(1)$. One can argue that such a method may be deemed unfair, when compared with the previous method of prematurely terminating the connections of the C_1 users whenever necessary. But one must bear in mind that these higher class users are higher paying customers, and are expecting a better treatment by the network.

Generally, the borrowing of units from those reserved for other classes of users should only be allowed by the network once all the users are utilizing their minimum allowable number of units, assuming that adaptive bandwidth allocation is employed by the network.

The previous example can be further extended for the case of three classes of users, as shown in *Figure 3.2*. The diagram shows how the network may assign the partitioning and borrowing thresholds, and behaves in a similar manner to the previous case of two

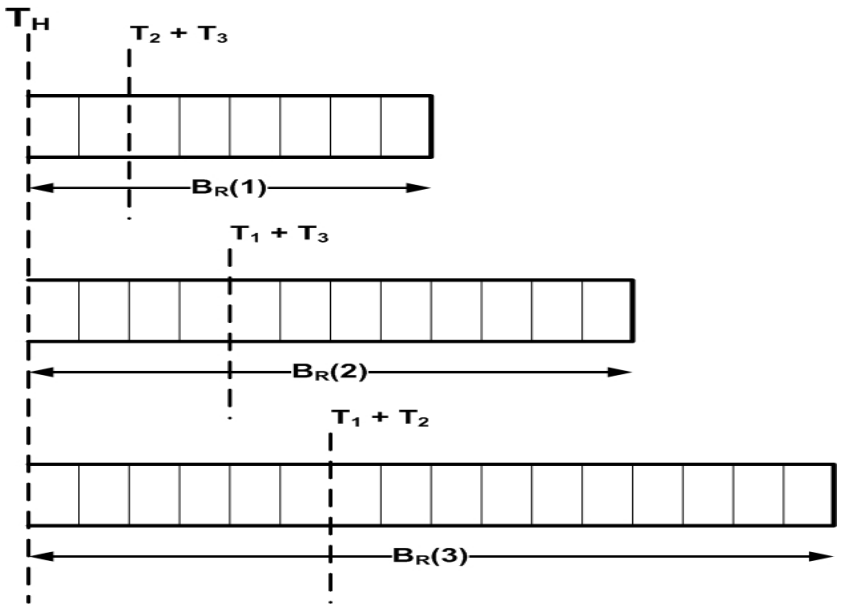

Fig. 3.2: Resource Partitioning and Borrowing for Three Classes of Users

classes of users. Note that for this example, $B_R = B_R(1) + B_R(2) + B_R(3)$.

3.6.2 Static vs. Adaptive Partitioning of the Total Network Resources

The static assignment of the total network resources would imply having a fixed number of units reserved for all users in each class. However, it may be difficult to choose how to fairly allocate a fixed number of units for each class. No matter how one chooses to assign these units (e.g. more units reserved for C_2 users than C_1, or each with an equal number of units), and no matter what the distribution of the different user populations are, some degree of unfairness will likely be introduced to some or all classes of users. Such a scheme may also make the network unappealing to some users. Hence, not only can static partitioning be inefficient, but it can also be an unattractive and an unfair method of distributing the available network resources.

An example of how static partitioning can be unfair is to consider the case where the population of C_1 users in the network is much greater than that of C_2 users, and the network has reserved more units for C_2 than C_1. Initially, this may seem like a reasonable thing to do, due to the fact that the C_2 users are higher-paying clients, and should expect a better performance. However, in this case a lot of C_1 users may be blocked from service, and at the same time there may be some units reserved for C_2 users that have remained unused for a considerable length of time. These extra units could instead have been assigned to those blocked C_1 users rather than remaining unused. Therefore, it might be efficient to employ an adaptive partitioning of the total network resources, and in accordance with the population of the different class of users.

The adaptive resource partitioning could also consider the time of day as another important factor, and in conjunction with the information on location and population of users. An example of such would be at a location where the network provides coverage on a campus area that contains a large number of staff members with C_2 subscriptions. The network should ideally consider the peak and off-peak times in its adaptive resource partitioning algorithm, thereby favoring to serve more C_2 users than usual at peak times, and continuing to serve all the classes in a general manner during the off-peak times.

3.7 The Denial of Immediate Service for New Requests

Users that are denied immediate service due to the unavailability of units (even after degrading all possible connections) to attend to their requests can be queued, with each class of users having their own separate queues. If the network was allowed to monitor the queue lengths of each queue (in each cell), then the network could use this information to aid in the adaptive assignment of the units reserved for each class of users. If we consider again the simple example of having only two classes of users in the network (C_1 and C_2), we could set up the queues for each class as shown in *Figure 3.3.*

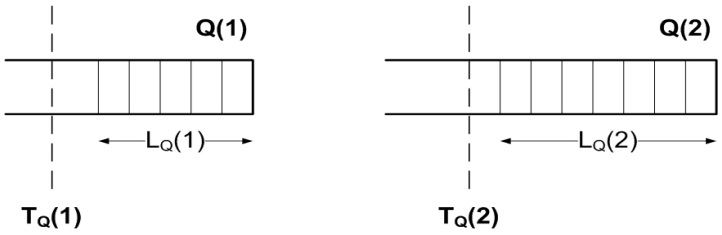

Fig. 3.3: The Queueing of Two Classes of Users

Let $L_Q(i)$ be the current queue lengths of the queues $Q(i)$, for each class i of users. The thresholds $T_Q(i)$ in each $Q(i)$ defines the level at which the network begins to make its decisions regarding the adaptive partitioning of the resource, once one of the queue lengths $L_Q(i)$ crosses its corresponding threshold $T_Q(i)$.

Consider the simple case where $N = 2$, as illustrated in *Figure 3.3.* The re-assignment of the dividing partition T_H could be triggered once either of the queue lengths have exceed its threshold. The network could then proceed to compute the difference in queue lengths, and thereby adjusting the dividing partition T_H if $L_Q(1) - L_Q(2) \geq$

K', or $L_Q(1) - L_Q(2) \leq K''$, where K' and K'' are pre-determined constants by the network. The difference in the queue lengths would continue to be monitored if any of the queue lengths are above their designated thresholds. Such as scheme could also prompt the network to change either or both T_1 and T_2 in *Figure 3.1*, for further optimum performance.

To help with understanding why the difference in queue lengths are computed, consider the simple example where the queue length $L_Q(2)$ has exceeded $T_Q(2)$. If the network computes a significant difference in the queue lengths such that there are relatively more C_2 users waiting to be served than C_1 users, then the number of units reserved for C_2 clients should be increased. The increase in the number of reserved units could be proportional to the difference in the queue lengths computed. If the network computes a minor difference in the queue lengths (e.g. almost the same number of users in both queues), even if $L_Q(1)$ and $L_Q(2)$ are large, then the network could decide not to alter T_H. In addition, the network should have limits on how T_H can be altered either way, for both classes of users.

3.8 Subscription Pricing

Since we are dealing with multiple classes of subscriptions, it is necessary to emphasize that the difference in prices of the subscriptions will play a very important factor for the selection of the level of priority by the users [29, 30]. To help explain this, consider the case where there is a small difference in the subscription fees between the two classes. This would imply that a lot more users (if not all!) would prefer to have a C_2 subscription since the difference in price is quite low. A relatively low price on the C_2 subscription would allow more users to benefit greatly from this offer, since they are willing to pay that little extra for that subscription if necessary. This notion is familiar to economists as *"Consumer Surplus"*, which is a measure of the amount that consumers benefit by being able to purchase a product/service for a price that is less than what they are willing to pay.

Looking back at *Figure 3.1*, a significant increase in the traffic of C_2 connections, as a result of a low C_2 subscription price, could prompt the network to tighten the threshold T_1 towards T_H as well as additionally increasing the threshold T_2, to allow for a satisfactory performance for the large population of C_2 users. However, the shift in thresholds could also lower the overall performance for the C_1 users. On the other hand, a large price difference in the subscriptions will have the opposite effect on the thresholds. Network operators try to overcome this problem by strictly controlling the subscription prices such that there is a significant difference between the tariffs for different classes. This effectively apportions the population of network users across the

different classes. It may seem that network operators try to exploit this notion for their own benefit by profiting from this opportunity (which may be true!). However, in actual fact, the controlling of the subscription prices would ultimately serve for the benefit of both the users and the network operators, and ensuring that not all the subscription prices are "affordable to everyone"!!

The authors in [29] have analyzed how a price-based allocation scheme can be used to guarantee a minimal QoS, and how to achieve a socially optimal bandwidth allocation. In [30], the authors explain how the service classes which require more resources should be assigned relatively higher prices, which will prevent their usage by those who can settle for less, but not deter those who actually need them.

3.9 Short-Term & Long-Term Performance Assurance

With an increase in the number of subscriptions for each class of users, it may become quite difficult to manage and meet the service demands of all the different classes of users, given the limited resources available to the network. One approach would be to allow the network to keep track of the history of the performance acquired for each user throughout their subscription lifetime. In the short-term, this would imply that some users may not be able to get served straight away and may be blocked (or queued), with the network keeping track of that occurrence for each user. The network could also keep track of the history of connection degradations for each user, as well as the number of times a user was allocated a maximum (or near maximum) number of units defined for his subscription profile.

The user's chance of getting served with a better performance (within the limits defined by the network) should continue to increase in the next attempts and will be assured of that by the network through the assignment of some type of service priority. The level of priority could be made in accordance with the user's performance history and subscription-type. Therefore, in the long term, the users should be assured a certain level of performance based on the type of subscription, with the service levels fluctuating in the short term, e.g. a minimum of 50% optimal service could be guaranteed for C_2 users while C_1 users are only guaranteed 20% optimal service, on the average.

4. A QUEUEING MODEL OF ADAPTIVE BANDWIDTH ALLOCATION FOR MULTIPLE CLASSES OF USERS

The framework presented in the previous Chapter will serve as the basis for developing the queueing model in this Chapter. Since, we are dealing with the analysis of future generation wireless networks such as 4G networks, the model will be constructed for the case of a two-layered hierarchy network (e.g. WLAN overlaid by a GPRS network). Both networks are assumed to share the same adaptive bandwidth allocation policy.

Handoff and new connections will generally be treated equally, i.e. the prioritizing of handoff connections will not be assumed. A zero-buffer queue will also be considered, whereby a new request is admitted into the system only when enough bandwidth units can be made available to that request. A new connection-request that is denied immediate service is assumed to be blocked and lost from the system.

Even though the type of terminals utilized by the users was taken into consideration in the previous Chapter, the queueing model that is developed in this Chapter assumes that the network does not distinguish between the different types of terminals, and treats all users of the same subscription class under the same policy. This assumption was made for the purpose of simplifying the model.

4.1 The General System Model

Given that the type of terminals will not be considered in the models developed in this Chapter, the subscript j in the variables $U_{i,j}$ and $X_{i,j,k}$ that are defined in equations *(3.1)* and *(3.4)*, respectively, will need to be dropped. Since the models will also deal with the case of two networks, a further subscript, w , will also need to be added to the variable $X_{i,k}$ to denote the network number, i.e. $w = 1$ for network 1, and $w = 2$ for network 2. Hence, the state $\mathbf{S}$ of the whole system is given as,

$$\mathbf{S} = \Big\{ \; X_{i,k,w} \; ; \quad 1 \leq i \leq N, \quad U_i^{min} \leq k \leq U_i^{max}, \quad w = \{1, 2\} \; \Big\} \tag{4.1}$$

Let B_w denote the total number of bandwidth units in network w. Hence, the state space Ω is the set of possible states, and is given by,

$$\Omega \; = \; \Big\{ \; \mathbf{S} \; : \; \sum_{i=1}^{N} \sum_{k=U_i^{min}}^{U_i^{max}} k \cdot X_{i,k,w} \; \leq \; B_w \quad , \quad \forall w \; \Big\} \tag{4.2}$$

The system can be analyzed as a continuous-time Markov process with state space given by Ω.

The arrivals of new class i connection-request into a cell in network w is assumed to be a Poisson process with rates λ_i^w. For our purpose, the Poisson process is considered to be a reasonably good approximation since we are dealing with a connection-level admission control and resource management.

The channel(or bandwidth unit)-holding times for class i connections in network w

is assumed to be exponentially distributed with means $(\mu_i^w)^{-1}$.

Because we are dealing with the case of a two-network system, there is a chance that a user's ongoing connection is transferred to the adjacent network in the two-layer hierarchy due to mobility. The rate of transfer of an ongoing class i connection from network 1 to network 2 is given by the parameter γ_i^2, and the transfer-rate of an ongoing class i connection from network 2 to network 1 is given by the parameter γ_i^1. The connection-transfer process is assumed to be a Poisson process.

4.2 The Complete Partitioning of the Network's Resources

4.2.1 The System Description

In this Section, a queueing model will be formulated and analyzed for the case where the total bandwidth units B_w in each network is completely partitioned, such that a fixed number of units $B_w(i)$ is reserved for all of class i users in the cell of network w. Hence, B_w is partitioned as follows

$$B_w \;=\; B_w(1) \;+\; B_w(2) \;+\; \cdots \;+\; B_w(i) \;+\; \cdots \;+\; B_w(N) \tag{4.3}$$

The model in this Section will also assume that a class i connection can only obtain either U_i^{min} or U_i^{max} units, and none in between, throughout the entire duration of the connection. In other words, all users can expect only two different levels of service.

Figure 4.1 shows an example of the system model for the case of two classes of subscriptions.

The complete partitioning of the total bandwidth units allows for the independent analysis of the various classes of connections in the system, as it was previously explained in *Section 3.6* . *Figure 4.1* also shows the example of how the traffic of one class of connections has no influence on the traffic of the other class. Based on that, the performance of each class of connections in the system could be analyzed independently. Hence, the two classes in *Figure 4.1* could be analyzed as two independent sub-systems.

Therefore, each sub-system could be analyzed as a Markov process with the system

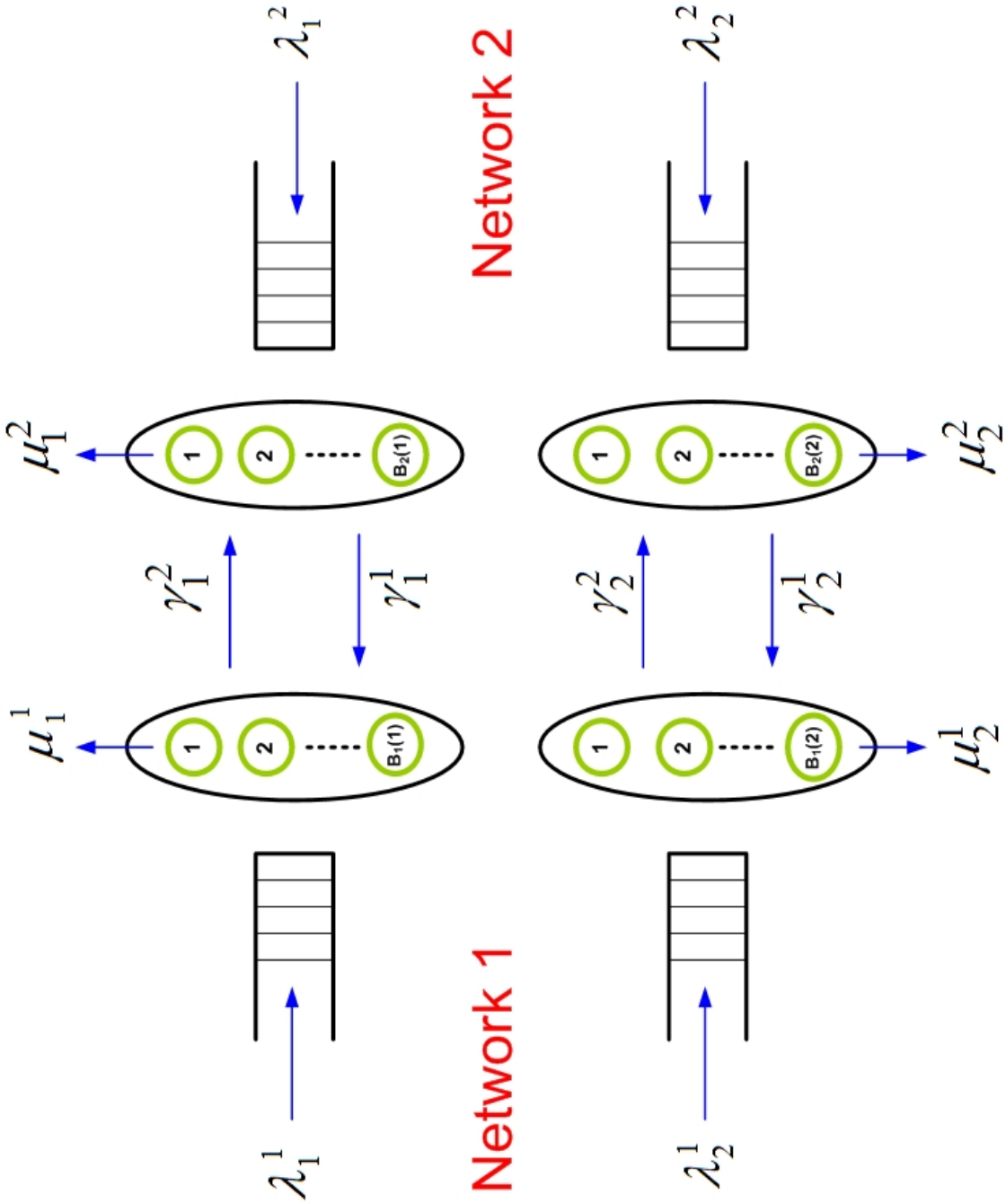

Fig. 4.1: The System Model with the Complete Partitioning of the Network's Resources

state vector $\mathbf{S}_p(i)$ for class i users, and given as

$$\mathbf{S}_p(i) \;=\; \Big\{\; X_{i,k,w} \;;\quad U_i^{min} \leq k \leq U_i^{max},\quad w = \{1,2\} \;\Big\} \tag{4.4}$$

Note that the variables $X_{i,k,w}$ have the following limits.

$$0 \;\leq\; X_{i,k,w} \;\leq\; \left\lfloor \frac{B_w(i)}{k} \right\rfloor \qquad ; \qquad \forall i \quad , \quad \forall k \quad , \quad \forall w = 1,2 \tag{4.5}$$

The state space $\Omega_p(i)$ of this Markov model is given as follows,

$$\Omega_p(i) \;=\; \Big\{\; \mathbf{S}_p(i) \;:\; \sum_{k=U_i^{min}}^{U_i^{max}} k \cdot X_{i,k,w} \;\leq\; B_w(i) \quad , \quad \forall w = 1,2 \Big\} \tag{4.6}$$

The size of the state space $\Omega_p(i)$ is given as follows,

$$|\Omega_p(i)| \;=\; \prod_{w=1}^{2} \left(\left\lfloor \frac{B_w(i)}{U_i^{min}} \right\rfloor \;+\; 1 \right) \tag{4.7}$$

The following assumptions regarding the system parameters $B_w(i)$, U_i^{min} , and U_i^{max}, were made. These assumptions were made for the purpose of simplifying the model, and it can also be argued that such assumptions might aid with the maximum utilization of the network's resource.

$$B_w(i) \; mod \; U_i^{min} \;=\; 0 \qquad and \qquad B_w(i) \; mod \; U_i^{max} \;=\; 0 \tag{4.8}$$

$$U_i^{min} \ mod \ \left(U_i^{max} - U_i^{min}\right) \ = \ 0 \tag{4.9}$$

$$\forall i \qquad \text{and} \qquad \forall w = 1, 2$$

The assumption in *(4.8)* implies that the partitioning of the total resources is made in such a way that is dependent on the network's pre-defined allocation units U_i^{max} and U_i^{min}. It further ensures that the fractions given in *(4.5)* and *(4.7)* are integers. The assumption in *(4.9)* ensures that no wastage of units occurs when a user's an existing connection is degraded. To illustrate, consider the example where $U_1^{min} = 2$ and $U_1^{max} = 5$ for class 1 users, and the system is in a state where the degrading of existing connections are necessary in order to accommodate a new class 1 request. A new request would cause the network to degrade one existing class 1 user with U_1^{max} units and thereby releasing 3 free units. But only 2 units are need for the new request, which means that 1 unit is wasted!

4.2.2 The System State Transitions

The set of possible system state transitions will next be described for the case of the system with states $\mathbf{S}_p(i)$ given in *(4.4)* . For simplicity of notation purposes, the subscript k in the variable $X_{i,k,w}$ will assume a value of 1 when $k = U_i^{min}$, and will assume a value of 2 when $k = U_i^{max}$. Hence, the state $\mathbf{S}_p(i)$ for this sub-system can be written as

$$\mathbf{S}_p(i) \;=\; (X_{i,1,1}\ ,\ X_{i,2,1}\ ,\ X_{i,1,2}\ ,\ X_{i,2,2}) \tag{4.10}$$

Consider first the set of possible system state transitions in network w, with this transition being equally true in both networks 1 and 2. If the network has enough free units to grant a class i request with U_i^{max}, then the system state transition can be described as shown in *Figure 4.2* .

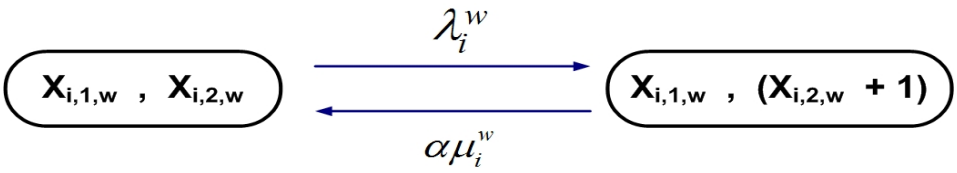

Fig. 4.2: The System State Transition for Allocating U_i^{max}

$$\alpha \;=\; \left(X_{i,1,w} \cdot U_i^{min} \;+\; (X_{i,2,w} + 1) \cdot U_i^{max}\right) \tag{4.11}$$

where α is the total number of units occupied by the users in the system state $\mathbf{S}_p(i)'$, given by the state transition $\mathbf{S}_p(i) \;\rightarrow\; \mathbf{S}_p(i)'$ in *Figure 4.2* .

Furthermore, the system state transition in *Figure 4.2* is only possible if

$$\left(X_{i,1,w} \cdot U_i^{min} \; + \; X_{i,2,w} \cdot U_i^{max}\right) \; \leq \; B_w(i) \; - \; U_i^{max} \tag{4.12}$$

Based on the properties given by *(4.8)* and *(4.9)* , the network will only begin to allocate U_i^{min} units to the new class i requests when the system has reached a state where all its reserved units for class i connections are utilized, prompting the need to degrade some randomly-chosen existing connections with U_i^{max} units, as described by the system state transition shown in *Figure 4.3* .

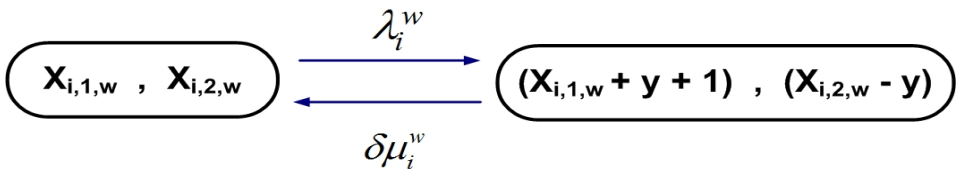

Fig. 4.3: The System State Transition for Allocating U_i^{min} by Connection Degradation

$$\delta \; = \; \left((X_{i,1,w} + y + 1) \cdot U_i^{min} \; + \; (X_{i,2,w} - y) \cdot U_i^{max} \right) \tag{4.13}$$

where δ is the total number of units occupied by the users in the system state $\mathbf{S}_p(i)'$, given by the state transition $\mathbf{S}_p(i) \; \rightarrow \; \mathbf{S}_p(i)'$ in *Figure 4.3* .

The variable y in *Figure 4.3* represents the number of users that are needed to have their connections degraded, in order to attend to the request of a new class i user by granting it U_i^{min}, and assuming that $X_{i,2,w} \; \geq \; y$. Remember that the network selects from the same class as the new request for their connections to be degraded. The number

of users that are needed to be degraded is given by

$$y = \left(\frac{U_i^{min}}{U_i^{max} - U_i^{min}} \right) \tag{4.14}$$

Note that the equation *(4.14)* will always yield an integer, given that the property in *(4.9)* is satisfied.

The reverse system state transition in *Figure 4.3* also shows how y users of class i can be selected at random to have their connections upgraded as a result of the departure of a class i connection from the system, due to service completion.

The system state transition that describes the case of a class i connection being transferred between the two networks is given in *Figure 4.4* .

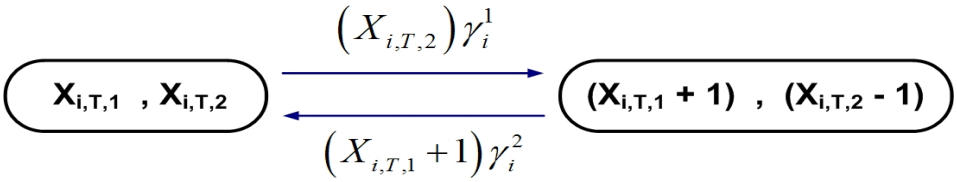

Fig. 4.4: The System State Transition for the Transfer of Connections Between Both Networks

$$where \;\; X_{i,T,1} = X_{i,1,1} + X_{i,2,1} \quad and \quad X_{i,T,2} = X_{i,1,2} + X_{i,2,2} \tag{4.15}$$

In *Figure 4.4* , all the variables that represent the number of class i users in each network were lumped together as given in *(4.15)*, since it is assumed that the transfer of a class i connection is independent of the number of units that it is receiving. Moreover,

a connection that has transferred to the new network may not necessarily receive the same number of units that was given by the previous network. In fact, it is assumed that from the point of view of the new network, the transferred connection would be treated as if it was a new connection request, and subject to the same bandwidth allocation policies given by the system state transitions in *Figures 4.2* and *4.3*.

An example of the overall state transition diagram for class 1 users in the system with the state vector $\mathbf{S}_p(1)$ is given in *Figure 4.5* . The example assumes that $U_1^{min} = 1$, $U_1^{max} = 2$, $B_1(1) = B_2(1) = 4$.

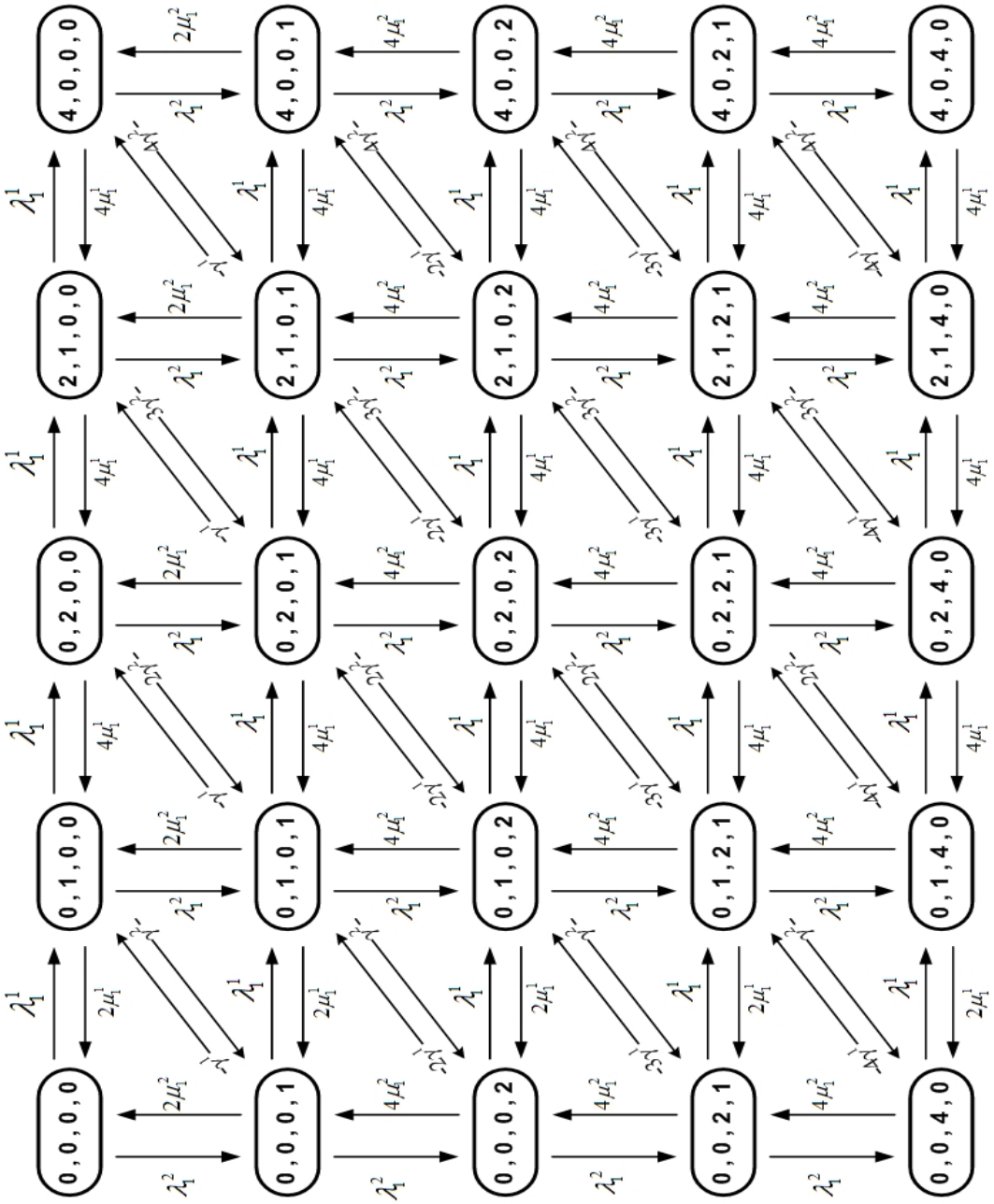

Fig. 4.5: An Example of the Overall State Transition Diagram for the System with the State Vector $\mathbf{S}_p(1)$

4.2.3 The Analysis of the Model

The Markov Process for this sub-system can be analyzed as a Quasi-Birth-Death process with a generator matrix given by $Q_p(i)$.

$$\mathbf{Q}_p(i) = \begin{bmatrix} q_{0,0} & q_{0,1} & & & \\ q_{1,0} & q_{1,1} & q_{1,2} & & \\ & q_{2,1} & q_{2,2} & q_{2,3} & \\ & & \ddots & \ddots & \ddots \\ & & & q_{t,t-1} & q_{t,t} \end{bmatrix} \tag{4.16}$$

$$where \quad t = \frac{B_1(i)}{U_i^{min}} \tag{4.17}$$

The matrices $q_{m,n}$, $\forall m, n$, are all square matrices of the order

$$\left(\frac{B_2(i)}{U_i^{min}} + 1 \right) \tag{4.18}$$

The index of the rows of the matrix $Q_p(i)$ represent the number of class i users in network 1, whereas the index of the rows in the matrices $q_{m,n}$ represent the number of class i users in network 2.

The matrices $q_{m,m+1}$, for $0 \leq m \leq t-1$, represent the arrival of a class i user into network 1, with the arrival being either a new request or a connection transfer from network 2. These matrices have the following form.

$$q_{m,m+1} = \begin{bmatrix} \lambda_i^1 & & & & \\ \gamma_i^1 & \lambda_i^1 & & & \\ & 2\gamma_i^1 & \lambda_i^1 & & \\ & & \ddots & \ddots & \\ & & & \phi\gamma_i^1 & \lambda_i^1 \end{bmatrix} \quad where\ \phi = \frac{B_2(i)}{U_i^{min}} \tag{4.19}$$

The matrices $q_{m,m-1}$, for $1 \leq m \leq t$, represent the departure of a class i user from network 1, with the departure being either due to a service completion or a connection transfer into network 2. These matrices have the following form

$$q_{m,m-1} = \begin{bmatrix} \alpha_1\mu_i^1 & m\gamma_i^2 & & & \\ & \alpha_1\mu_i^1 & m\gamma_i^2 & & \\ & & \alpha_1\mu_i^1 & m\gamma_i^2 & \\ & & & \ddots & \ddots \\ & & & & \alpha_1\mu_i^1 \end{bmatrix} \tag{4.20}$$

$$where \quad \alpha_1 = X_{i,1,1} \cdot U_i^{min} + X_{i,2,1} \cdot U_i^{max} \tag{4.21}$$

The matrices $q_{m,m}$, for $0 \leq m \leq t$, has the following form.

$$q_{m,m} = \begin{bmatrix} \theta_{m,0} & \lambda_i^2 & & & & \\ \alpha_2\mu_i^2 & \theta_{m,1} & \lambda_i^2 & & & \\ & \alpha_2\mu_i^2 & \theta_{m,2} & \lambda_i^2 & & \\ & & & \ddots & \ddots & \\ & & & & \alpha_2\mu_i^2 & \theta_{m,r} \end{bmatrix} \quad where\ r\ =\ \frac{B_2(i)}{U_i^{min}} \tag{4.22}$$

$$with \quad \alpha_2\ =\ X_{i,1,2} \cdot U_i^{min}\ +\ X_{i,2,2} \cdot U_i^{max} \tag{4.23}$$

$\theta_{m,x}$, for $0 \leq x \leq r$, is the negative of the sum of all the other elements that are in the same row as $\delta_{m,x}$ in the generator matrix $Q_p(i)$.

The steady-state distribution of the system with the generator matrix $Q_p(i)$ can be computed using the following,

$$0\ =\ \Phi\Big(\mathbf{S}_p(i)\Big) \cdot \mathbf{Q}_p(i) \qquad and \qquad \Phi\Big(\mathbf{S}_p(i)\Big) \cdot \mathbf{e}\ =\ 1 \tag{4.24}$$

$\Phi\Big(\mathbf{S}_p(i)\Big)$ is the steady-state probability vector of the system with states $\mathbf{S}_p(i)$, and contains the elements $p\,(x_{i,1,1}\ ,\ x_{i,2,1}\ ,\ x_{i,1,2}\ ,\ x_{i,2,2})$, i.e. the steady-state probabilities of the system. $\mathbf{e}$ is a column vector of 1. Matlab was used to solve for the steady-state probability distributions.

4.2.4 The Performance Metrics of the System

Three different performance metrics were defined for the purpose of analyzing the performance of the system. They may also be used to investigate the optimal network parameters, e.g. the different bandwidth unit allocations for the different classes of networks, as well as the partitioning of the network's total resources.

The first performance metric deals with the computation of the blocking probabilities of the class i users in network w. Note that the blocking probability incorporates both the blocking of new connection requests and the transfer of connections from the other network. Such a performance parameter has been traditionally used by many researchers, since the blocking of a connection-request from service has a considerable impact on the level of QoS offered by the system, as perceived by the users. The definition of the blocking probability is given by

$$P_b(i,w) \;=\; Pr\{Blocking\ of\ Class\ i\ connections\ in\ Network\ w\} \tag{4.25}$$

In order to be able to compute the above probability, the original definition of the system state vector $\mathbf{S}_p(i)$ given in equation *(4.4)* was re-written in a way such that the variables with the same class i index in each network w is grouped together to form a "super-variable".

$$\widehat{\mathbf{S}_p(i)} \;=\; \{\; X_{i,w}\;,\; \forall k \;\;:\;\; X_{i,w} \;=\; \sum_{k=U_i^{min}}^{U_i^{max}} X_{i,k,w}\;\} \;=\; (\; X_{i,1}\;,\; X_{i,2}\;) \tag{4.26}$$

The modified system state vector, $\widehat{\mathbf{S}_p(i)}$, does not alter in any way the definition of the system, and is used solely for the purpose of computing the blocking probabilities for class i subscribers in both networks, using the steady-state probability distribution $\Phi\left(\widehat{\mathbf{S}_p(i)}\right)$, which now contains the elements $p\Big(\ x_{i,1}\ ,\ x_{i,2}\ \Big)$.

$$P_b(i,1) = \{Blocking\ of\ Class\ i\ connections\ in\ Network\ 1\} \tag{4.27}$$

$$= \sum_{x_{i,2}=0}^{\frac{B_2(i)}{U_i^{min}}} p\left(\frac{B_1(i)}{U_i^{min}}\ ,\ x_{i,2}\right) \tag{4.28}$$

$$P_b(i,2) = \{Blocking\ of\ Class\ i\ connections\ in\ Network\ 2\} \tag{4.29}$$

$$= \sum_{x_{i,1}=0}^{\frac{B_1(i)}{U_i^{min}}} p\left(x_{i,1}\ ,\ \frac{B_2(i)}{U_i^{min}}\right) \tag{4.30}$$

The blocking probability $P_b(i,1)$ computes the sum of the probabilities $p\left(\frac{B_1(i)}{U_i^{min}}\ ,\ x_{i,2}\right)$, $\forall\ x_{i,2}$. These probabilities denote the state of the system where all if its units that are reserved for class i connections in network 1 have been utilized to the maximum. In other words, all the class i connections are receiving U_i^{min} units and cannot be degraded any further, implying that no more new requests of the same class can be admitted into network 1. A similar argument applies for the computation of $P_b(i,2)$.

The other performance metric that was defined for this system is the probability

that a class i connection would be granted U_i^{max} units by the network w, upon initial connection (for both new connections and those transferred from the other network). This probability is defined as follows.

$$P_{max}(i,w) \;=\; Pr\{Class\ i\ Obtaining\ U_i^{max}\ Upon\ Initial\ Connection\ into\ Network\ w\} \tag{4.31}$$

The steady-state probability distribution $\Phi\left(\widehat{\mathbf{S}_p(i)}\right)$ will again be used to compute the probabilities $P_{max}(i,w)$, and are given as follows.

$$P_{max}(i,1) \;=\; \sum_{x_{i,1}=0}^{\left(\frac{B_1(i)}{U_i^{max}}-1\right)} \; \sum_{x_{i,2}=0}^{\frac{B_2(i)}{U_i^{min}}} p\left(x_{i,1}\ ,\ x_{i,2}\right) \tag{4.32}$$

$$P_{max}(i,2) \;=\; \sum_{x_{i,1}=0}^{\frac{B_1(i)}{U_i^{min}}} \; \sum_{x_{i,2}=0}^{\left(\frac{B_2(i)}{U_i^{max}}-1\right)} p\left(x_{i,1}\ ,\ x_{i,2}\right) \tag{4.33}$$

If a further class i user is to be granted U_i^{max} units upon initial connection with network w, then the network should be in a state where the number of class i connections does not exceed $\left(\frac{B_w(i)}{U_i^{max}}-1\right)$. Hence, the following condition, $x_{i,w} \;\leq\; \left(\frac{B_2(i)}{U_i^{max}}-1\right)$, must be satisfied in order to ensure that at least another new class i request can be granted U_i^{max} units. This argument was used for computing $P_{max}(i,1)$ and $P_{max}(i,2)$.

The "Degrade Level" $E_w(i)$ of class i users in network w is another performance metric that was defined for this system. A similar definition and formulation was first presented by the authors in [34], who argued that such a performance measure could quantify the level of satisfaction from a user's perspective. $E_w(i)$ for class i users in network w is defined as the ratio of the average allocated bandwidth units to the desired maximum bandwidth units. According to the authors in [34], this assumes that the average allocated bandwidth units has a significant impact on the user's satisfaction.

$$E_w(i) \;=\; \sum_{\mathbf{s}} \; p\,(\mathbf{s}) \left(\frac{U_i^{max} \cdot x_{i,2,w} \;+\; U_i^{min} \cdot x_{i,1,w}}{U_i^{max} \cdot (x_{i,1,w} \;+\; x_{i,2,w})} \right) \tag{4.34}$$

$$\text{for all} \quad \mathbf{s} \in \Phi(\mathbf{S}_p(i)) \;, \quad \mathbf{s} \;=\; (x_{i,1,1} \;,\; x_{i,2,1} \;,\; x_{i,1,2} \;,\; x_{i,2,2})$$

The performance metric $E_w(i)$ actually measures the overall level of satisfaction of all the class i subscribers in network w, in terms of the number of bandwidth units that has been actually allocated to those users. In other words, a class i user is assumed to be satisfied at the most if he/she were allocated U_i^{max}, with that level of satisfaction dropping if he/she were instead allocated U_i^{min}.

4.2.5 Model Extension - Multiple Levels of Service

The model described in the previous sub-Sections can be extended to include the analysis of multiple levels of service, i.e. a class i can be assigned anywhere between U_i^{min} and U_i^{max} units. This extension will not affect the *overall* structure of the model in terms of the dimensions and structures of the QBD matrices. Hence, each class i user can be allocated U_i^L units, such that

$$U_i^{min} \leq U_i^L \leq U_i^{max} \tag{4.35}$$

For this case, a reasonable adaptive bandwidth allocation policy to adopt here would be to allow the network to assign the new class i user with $U_i^{L'}$ units, by degrading the existing connections that are utilizing $U_i^{L'+1}$ units, where $U_i^{min} \leq U_i^{L'} < U_i^{max}$. This would only occur if the network enters a stage where it has to degrade some existing connections to accommodate for more users. In other words, the network should first start to degrade only those connections with U_i^{max} units, with the new user receiving the next best level of service, i.e. $U_i^{max} - 1$ units. The same process is repeated for further incoming new requests until there are no more users with U_i^{max} units. The network would then proceed to degrade only those connections with service level $U_i^{max} - 1$ units, with the new user receiving the next best level of service, i.e. $U_i^{max} - 2$ units, and so on.

Furthermore, to allow for the efficient utilization of the network's resources, the

following properties for the network parameters $B_w(i)$ and U_i^L are assumed.

$$B_w(i) \; mod \; U_i^L \; = \; 0 \qquad ; \qquad \forall U_i^L \qquad U_i^{min} \; \leq \; U_i^L \; \leq \; U_i^{max} \tag{4.36}$$

To illustrate, consider the example of system where $B_1(1) = 6$, $U_1^{min} = 1$, and $U_1^{max} = 3$, for class 1 users in network 1. Hence, we have a system where these users can be assigned 3 different levels of service. *Figure 4.6* shows the state transition diagram for this example , where the system states are given by the vector $(Y_1 \; , \; Y_2 \; , \; Y_3)$, with Y_k being the number of class 1 users in network 1 with k bandwidth units.

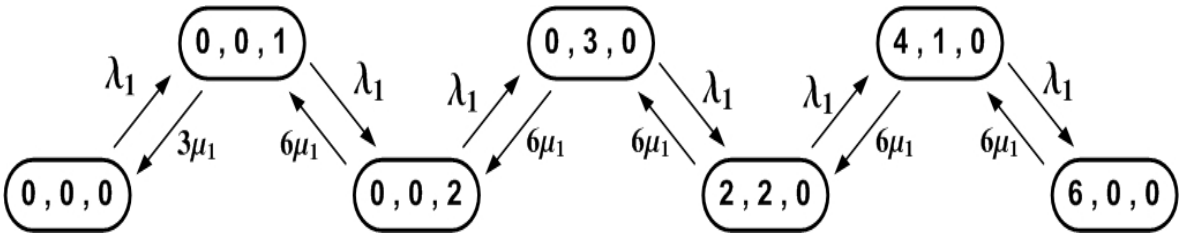

Fig. 4.6: An Example of the System State Transition with 3 Levels of Service

The properties in equations *(4.4)* to *(4.7)* also applies for the case of this extended model. The same matrix structures given by $Q_p(i)$ in *Section 4.2.3* can also be applied for the case of this extended model. Note that α_1 and α_2 from equations (4.23) and (4.23), respectively, will in this case be generally defined as

$$\alpha_w \; = \; \sum_{k=U_i^{min}}^{U_i^{max}} k \cdot X_{i,k,w} \qquad\qquad \forall w = 1, 2 \tag{4.37}$$

A similar approach to the one taken in *Section (4.2.4)* can be used to define the blocking probabilities for class i connections in network w. In fact, the equations *(4.27)* to *(4.30)* can also be used to compute the blocking probabilities for this extended model.

Instead of using the previous definition of $P_{max}(i,k)$ given by the equation (4.31), an alternative definition could be used to further explore the performance of this extended model. For this case, the probability $P_k(i,w)$ can instead be defined as follows.

$$P_k(i,w) \;=\; Pr\left\{Class\ i\ User\ in\ Network\ w\ Obtaining\ \textbf{\textit{at Least}}\ k\ Units\ Upon\ Initial\ Connection\right\}$$

The following equations can be used to compute the probabilities $P_k(i,1)$ and $P_k(i,2)$ for networks 1 and 2, respectively.

$$P_k(i,1) \;=\; \sum_{x_{i,1}=0}^{\left(\frac{B_1(i)}{k}-1\right)} \; \sum_{x_{i,2}=0}^{\frac{B_2(i)}{U_i^{min}}} p\left(x_{i,1}\ ,\ x_{i,2}\right) \tag{4.38}$$

$$P_k(i,2) \;=\; \sum_{x_{i,1}=0}^{\frac{B_1(i)}{U_i^{min}}} \; \sum_{x_{i,2}=0}^{\left(\frac{B_2(i)}{k}-1\right)} p\left(x_{i,1}\ ,\ x_{i,2}\right) \tag{4.39}$$

$$\text{for} \quad U_i^{min} \;\leq\; k \;\leq\; U_i^{max}$$

The Degrade Level $E_w(i)$ for class i users in network w with multiple levels of service is defined as follows

$$E_w(i) \;=\; \sum_{\mathbf{s}} \; p\,(\mathbf{s}) \left(\frac{\sum_{k=U_i^{min}}^{U_i^{max}} k \cdot X_{i,k,w}}{U_i^{max} \cdot \sum_{k=U_i^{min}}^{U_i^{max}} X_{i,k,w}} \right) \tag{4.40}$$

$$\text{for all} \quad \mathbf{s} \in \Phi(\mathbf{S}_p(i)) \;, \quad \mathbf{s} \;=\; (x_{i,1,1} \;,\; x_{i,2,1} \;,\; x_{i,1,2} \;,\; x_{i,2,2})$$

4.2.6 Numerical Examples

In this Section, various numerical examples will be presented for the system described in *Section 4.2.2* . The results were obtained using the performance parameters defined in *Section 4.2.4* . The examples will consider the performance of the system for class 1 subscribers, and the following network parameters will be assumed, while keeping the remaining system parameters (i.e. arrival, service, and connection transfer rates) consistent.

$$B_1(1) = 6 \qquad B_2(1) = 6 \qquad U_1^{min} = 1 \qquad U_1^{max} = 2$$

The choice of network parameters may seem unreasonable but they were chosen for the purpose of showing certain properties in the behavior of the system.

The first set of graphs given in *Figures 4.7* and *4.8* shows the blocking probabilities in both networks for class 1 subscribers, corresponding to the varying of both the arrival rates in network 1 and network 2, respectively. A key point to note from these graphs is how the traffic in one network affects the blocking probabilities in both networks, due to the allowance of connection-transfers between the two networks in the system. This behavior can be similarly observed under varying service and connection-transfer rates.

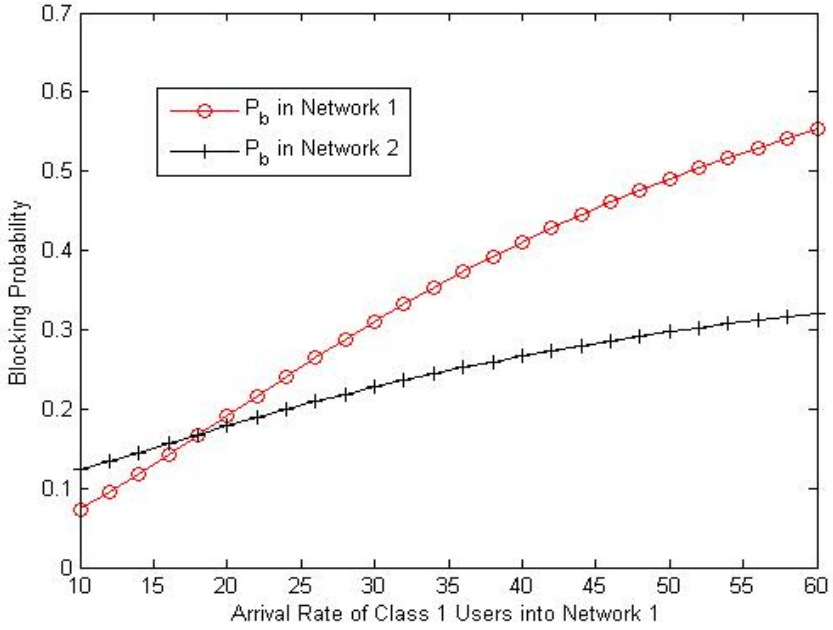

Fig. 4.7: A Graph Showing the Blocking Probabilities Corresponding to Varying Arrival Rates in Network 1

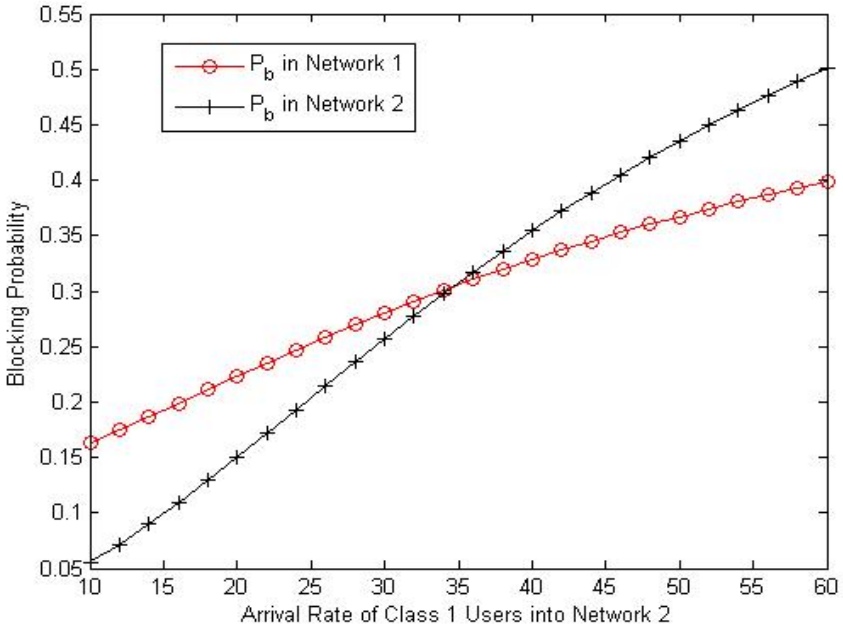

Fig. 4.8: A Graph Showing the Blocking Probabilities Corresponding to Varying Arrival Rates in Network 2

A similar behavior can be observed when looking at the probability of obtaining U_i^{max} units in both networks, corresponding to the varying of both the arrival rates in network 1 and network 2, and shown by the graphs in *Figures 4.9* and *4.10*, respectively.

The degrade levels for class 1 users in both networks, corresponding to the varying of both the arrival rates in network 1 and network 2, are shown by the graphs in *Figures 4.11* and *4.12*, respectively. Both graphs clearly show how the degrade levels, or levels of overall satisfaction, drops considerably as the traffic of class 1 subscribers in both networks increases. Moreover, the overall level of satisfaction in one network influences the overall level of satisfaction in the other network. The initial increase in the graph for the degrade level in network 1 given in *Figure 4.11* is due to the increase in the number of class 1 users that have been allocated their maximum number of bandwidth units under low traffic conditions. Hence, the increase is due to the increase in the ***overall*** level of satisfaction for all class 1 users in network 1. The same is true for the case of the graph for the degrade level in network 2 given in *Figure 4.12* .

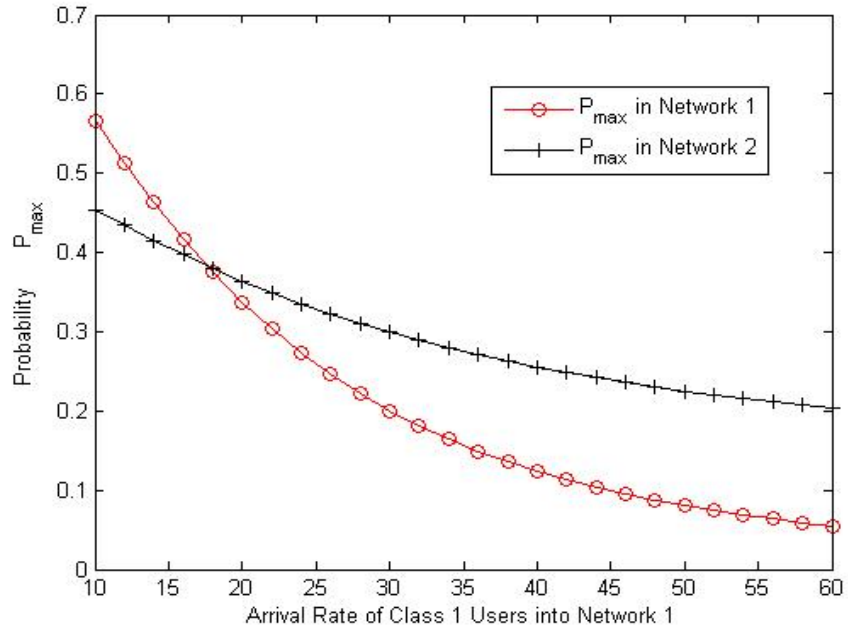

Fig. 4.9: A Graph Showing the Probabilities of Obtaining U_i^{max}, Corresponding to Varying Arrival Rates in Network 1

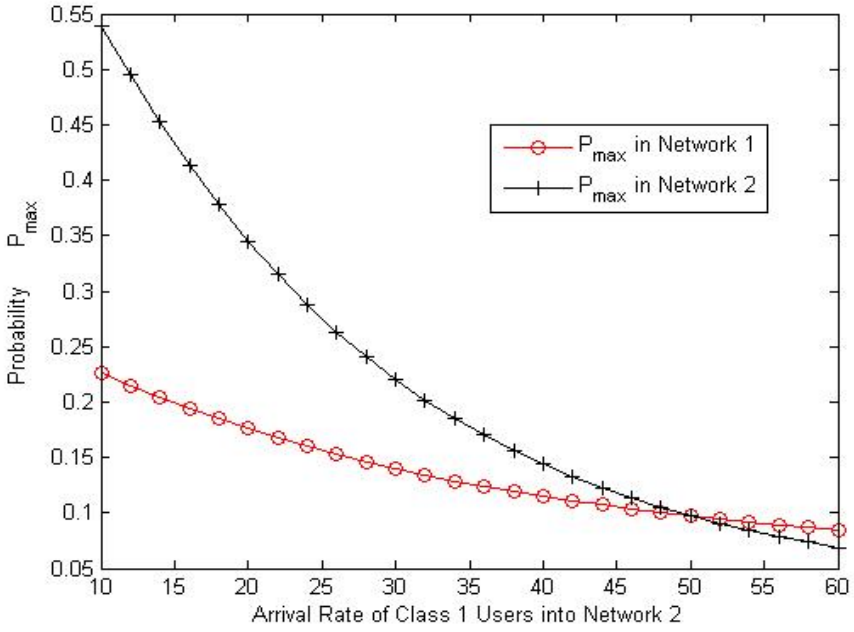

Fig. 4.10: A Graph Showing the Probabilities of Obtaining U_i^{max}, Corresponding to Varying Arrival Rates in Network 2

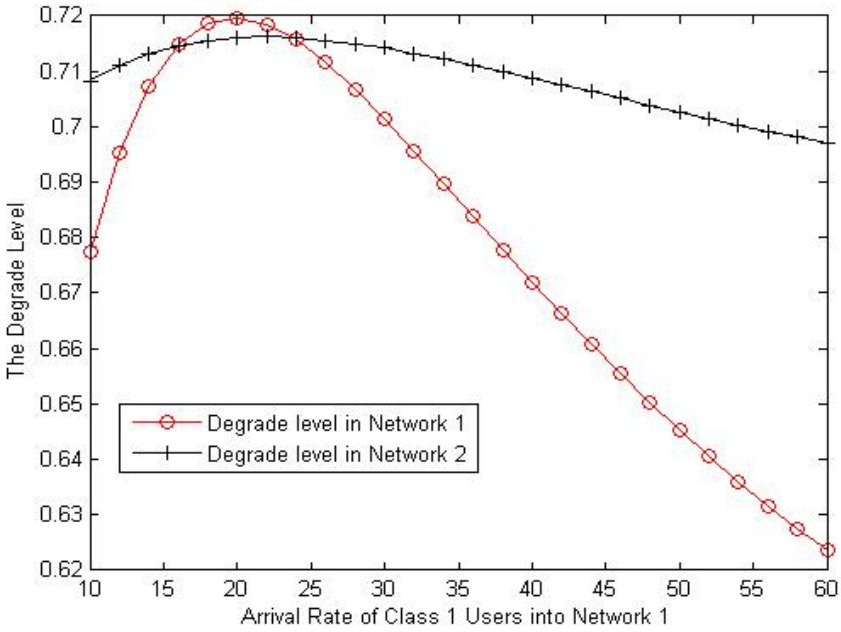

Fig. 4.11: A Graph Showing the Degrade Level for Class 1 Subscribers, Corresponding to Varying Arrival Rates in Network 1

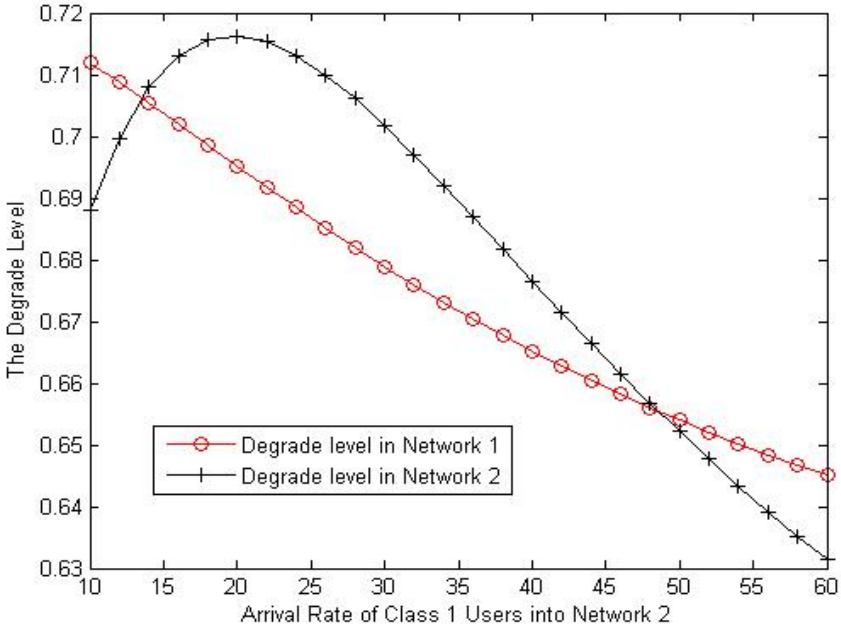

Fig. 4.12: A Graph Showing the Degrade Level for Class 1 Subscribers, Corresponding to Varying Arrival Rates in Network 2

4.3 The Complete Sharing of the Network's Resources

4.3.1 The System Description

A second model was developed which considered the case where the total network resource B_w, in each cell of the network w , is completely shared by all classes of subscribers in that network. This model will also assume that a class i connection can only obtain either U_i^{min} or U_i^{max} units, and none in between, throughout the entire duration of the connection. The case of two classes of subscriptions will only be considered in this model. Based on the definition of the system-state given in equation *(4.1)*, the state of the system for the model in this case is given by the vector $\mathbf{S}_h$, such that

$$\mathbf{S}_h = \Big\{ X_{i,k,w} \;\; : \;\; 1 \leq i \leq 2, \quad U_i^{min} \leq k \leq U_i^{max}, \quad w = \{1,2\} \Big\} \tag{4.41}$$

$$where \quad 0 \;\leq\; X_{i,k,w} \;\leq\; \left\lfloor \frac{B_w}{k} \right\rfloor \qquad \forall i,k,w \tag{4.42}$$

The state space Ω_h of this Markov Model is given as follows,

$$\Omega_h \;=\; \Big\{ \mathbf{S}_h \;:\; \sum_{i=1}^{2} \sum_{k=U_i^{min}}^{U_i^{max}} k \cdot X_{i,k,w} \;\leq\; B_w \;\;, \;\; \forall w = 1,2 \Big\} \tag{4.43}$$

The size of the state space is given as follows

$$|\Omega_h| \;=\; \prod_{w=1}^{2} \left(\sum_{t=0}^{\left\lfloor \frac{B_w}{U_2^{min}} \right\rfloor} \left[\left\lfloor \frac{B_w - tU_2^{min}}{U_1^{min}} \right\rfloor \;+\; 1 \right] \right) \tag{4.44}$$

The following assumptions regarding the system parameters B_w , U_i^{min} , and U_i^{max}, were made. These assumptions were made for the purpose of simplifying the model, and it can also be argued that such assumptions might aid with the maximum utilization of the network's resource.

$$B_w \; mod \; k \;=\; 0 \qquad ; \qquad \forall i, w, k \tag{4.45}$$

$$U_1^{min} \; mod \; \left(U_i^{max} - U_i^{min} \right) \;=\; 0 \qquad ; \qquad \forall i = 1, 2 \tag{4.46}$$

$$U_2^{min} \; mod \; \left(U_i^{max} - U_i^{min} \right) \;=\; 0 \qquad ; \qquad \forall i = 1, 2 \tag{4.47}$$

The assumptions given in equations *(4.46)* and *(4.47)* ensure that no bandwidth units are wasted when an existing connection undergoes degradation.

A diagram of the system model to be analyzed is given in *Figure 4.13* .

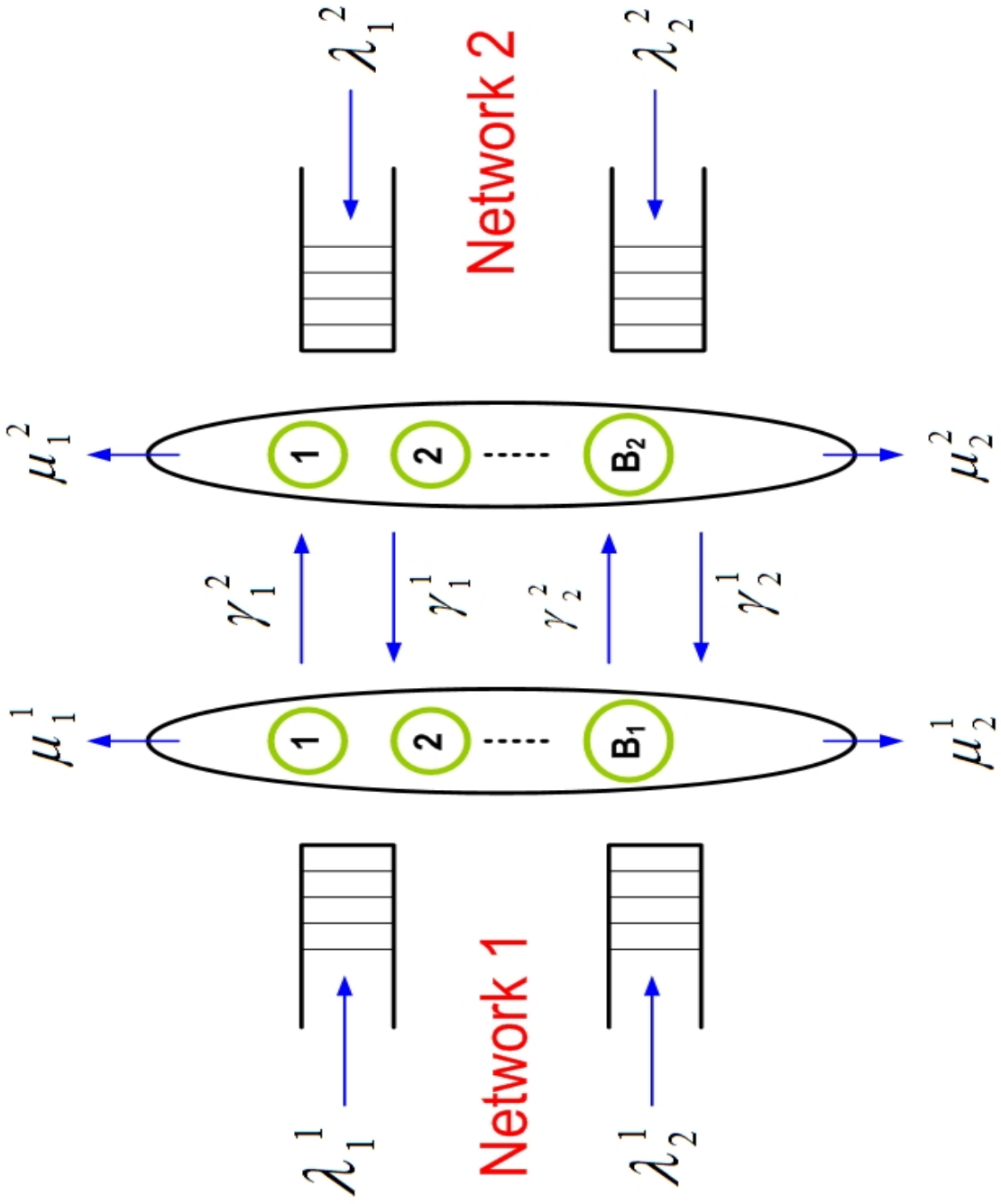

Fig. 4.13: The System Model with the Complete Sharing of the Network's Resource For Two Classes of Subscribers

4.3.2 The System State Transitions

The set of possible system state transitions will next be described for the case of the system with states $\mathbf{S}_h$ given in *(4.41)*. A similar method to the one taken in *Section 4.4.2* will be applied, where the subscript k in the variables $X_{i,k,w}$ used in the following set of state transition diagrams will assume a value of 1 when $k = U_i^{min}$, and a value of 2 when $k = U_i^{max}$. This modification will not affect the definition of the system, and is only done for the purpose of simplifying the notations (for clarity) presented in the set of possible system state transition diagrams for this sub-Section. Hence, the state $\mathbf{S}_h$ in the set of possible system state transition diagrams for this system can be written as

$$\mathbf{S}_h = (X_{1,1,1}\ ,\ X_{1,2,1}\ ,\ X_{2,1,1}\ ,\ X_{2,2,1}\ ,\ X_{1,1,2}\ ,\ X_{1,2,2}\ ,\ X_{2,1,2}\ ,\ X_{2,2,2}) \tag{4.48}$$

The system state transition shown in *Figure 4.14* describes the case where a network w has enough free units to grant a new class i request with U_i^{max}. This transition applies for both classes of new requests, and in both networks.

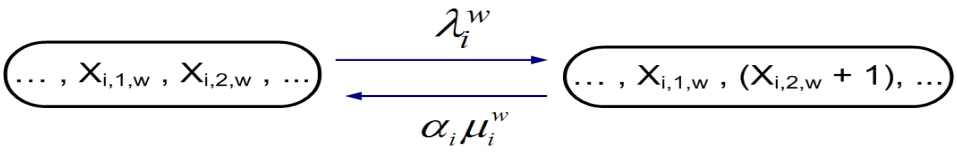

Fig. 4.14: The System State Transition for Allocating U_i^{max}

$$\alpha_i \;=\; \left(X_{i,1,w} \cdot U_i^{min} \;+\; (X_{i,2,w}+1) \cdot U_i^{max}\right) \tag{4.49}$$

α_i is the total number of units occupied by all class i users in network w, in the forward state. The system state transition in *Figure 4.14* is only possible if

$$\sum_{i=1}^{2} \sum_{k=U_i^{min}}^{U_i^{max}} k \cdot X_{i,k,w} \;\leq\; B_w \;-\; U_i^{max} \tag{4.50}$$

However, if the system is in a state where there are not enough free units to allocate a class i user U_i^{max} in network w, and instead there is enough to grant it U_i^{min} units, then the system state transition for this case is given in *Figure 4.15* . This transition also applies for both classes of new requests, and in both networks.

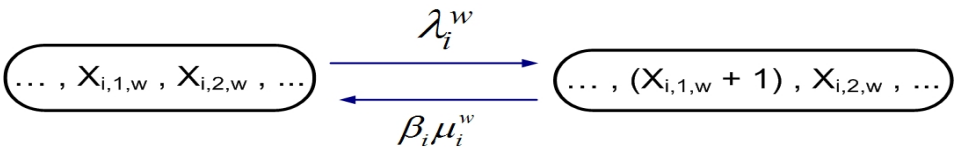

Fig. 4.15: The System State Transition for Allocating U_i^{min}

$$\beta_i \;=\; \left((X_{i,1,w}+1) \cdot U_i^{min} \;+\; X_{i,2,w} \cdot U_i^{max}\right) \tag{4.51}$$

β_i is the total number of units occupied by all class i users in network w, in the forward state. The system state transition in *Figure 4.15* is only possible if

$$\sum_{i=1}^{2} \sum_{k=U_i^{min}}^{U_i^{max}} k \cdot X_{i,k,w} \;\leq\; B_w \;-\; U_i^{min} \tag{4.52}$$

If the system is in a state where there are not enough free units to grant a new class i request in network w with U_i^{min}, then the system undergoes the process of degrading one or some of the existing connections in network w at random. However, since the total network resources in each network w is shared amongst all the classes of users, a different connection degradation policy to the one used in *Section 4.2* will need to be adopted.

It is proposed to have the network first degraded the existing class 1 connections before allowing for the degrading of the existing class 2 connections, regardless of the class of the new request. In Addition, it is also proposed to have the newly arriving class 2 connections to be always allocated U_2^{max} units whenever possible, even if it means having to degrade some of the existing class 1 connections. Otherwise, U_2^{min} will be assigned to the new class 2 users. On the other hand, the new class 1 requests should only be allowed U_1^{min}, if the network is in the state where connections need to be degraded.

This policy implies that the existing class 2 connections can only be degraded when all of class 1 connections have been degraded. Such a policy gives class 2 connections a higher priority in keeping their U_2^{max} units, which seems reasonable since they are assumed to have higher subscriptions. The connection degradation policy adopted for this model will be described by the following possible sets of system transition states.

Assuming that the network initiates the connection degradation process (due to the lack of bandwidth availability for the new class i request), and assuming that there are

enough existing class 1 users with U_1^{max} units that can be degraded, then the following system state transitions given in *Figures 4.16* to *4.18* are possible, and equally applies in both networks.

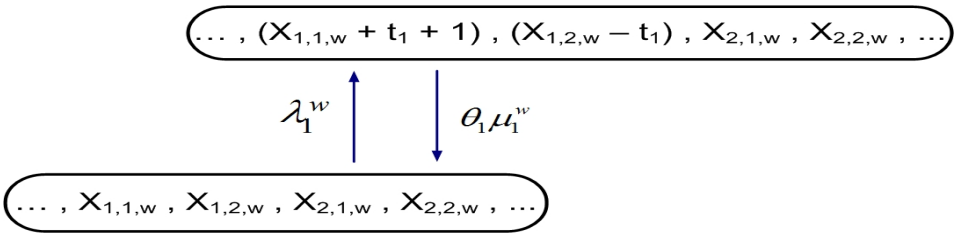

Fig. 4.16: The System State Transition for Allocating U_1^{min} to a New Class 1 Request By Degrading Existing Class 1 Connections

$$\theta_1 \;=\; \Big((X_{1,1,w} + t_1 + 1) \cdot U_1^{min} \;+\; (X_{1,2,w} - t_1) \cdot U_1^{max} \Big) \tag{4.53}$$

$$t_1 \;=\; \frac{U_1^{min}}{U_1^{max} \;-\; U_1^{min}} \tag{4.54}$$

The state transition given in *Figure 4.16* is for the case of the arrival of a new class 1 request, where t_1 in *(4.54)* is the number of class 1 connections with U_1^{max} that need to be degraded in order to accept the new class 1 request. θ_1 in *(4.53)* is the total number of units that are occupied by all of class 1 connections in network w, in the forward state. This transition is only possible if the condition in *(4.55)* can be satisfied, i.e. enough units can be accumulated by degrading the existing class 1 connections.

$$X_{1,2,w} \cdot \Big(U_1^{max} \;-\; U_1^{min} \Big) \;\geq\; U_1^{min} \qquad \textit{and} \qquad X_{1,2,w} \;\geq\; t_1 \tag{4.55}$$

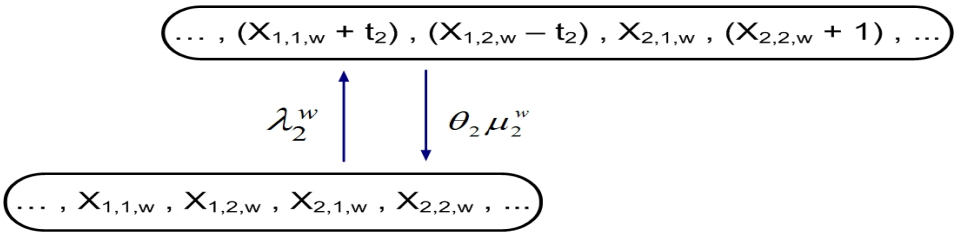

Fig. 4.17: The System State Transition for Allocating U_2^{max} to a New Class 2 Request By Degrading Existing Class 1 Connections

$$\theta_2 \;=\; \left(X_{2,1,w} \cdot U_2^{min} \;+\; (X_{2,2,w} + 1) \cdot U_2^{max}\right) \tag{4.56}$$

$$t_2 \;=\; \frac{U_2^{max}}{U_1^{max} \;-\; U_1^{min}} \tag{4.57}$$

The state transition given in *Figure 4.17* is for the case of the arrival of a new class 2 request, where t_2 in *(4.57)* is the number of class 1 connections with U_1^{max} that need to be degraded in order to accept the new class 2 request with U_2^{max}. θ_2 in *(4.56)* is the total number of units that are occupied by all of class 2 connections in network w, in the forward state. This transition is only possible if the condition in *(4.58)* can be satisfied, i.e. enough units can be accumulated to grant the new class 2 request with U_2^{max} units, by degrading the existing class 1 connections.

$$X_{1,2,w} \cdot \left(U_1^{max} \;-\; U_1^{min}\right) \;\geq\; U_2^{max} \qquad \textit{and} \qquad X_{1,2,w} \;\geq\; t_2 \tag{4.58}$$

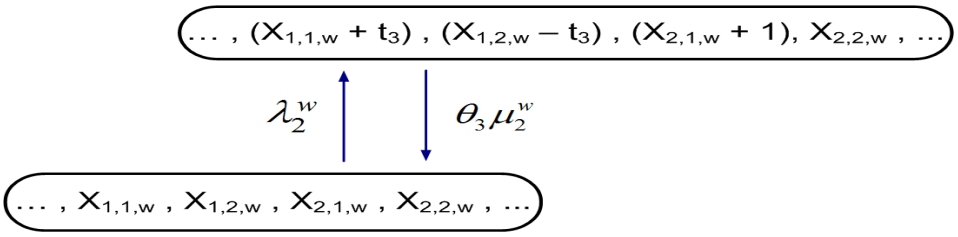

Fig. 4.18: The System State Transition for Allocating U_2^{min} to a New Class 2 Request By Degrading Existing Class 1 Connections

$$\theta_3 \;=\; \Big((X_{2,1,w} + 1) \cdot U_2^{min} \;+\; X_{2,2,w} \cdot U_2^{max} \Big) \tag{4.59}$$

$$t_3 \;=\; \frac{U_2^{min}}{U_1^{max} \;-\; U_1^{min}} \tag{4.60}$$

The state transition given in *Figure 4.18* is for the case of the arrival of a new class 2 request, where t_3 in *(4.60)* is the number of class 1 connections with U_1^{max} that need to be degraded in order to accept the new class 2 request with U_2^{min}, given that U_2^{max} cannot be allocated. θ_3 in *(4.59)* is the total number of units that are occupied by all of class 2 connections in network w, in the forward state. This transition is only possible if the condition in *(4.61)* can be satisfied, i.e. enough units can be accumulated to grant the new class 2 request with U_2^{min} units, by degrading the existing class 1 connections.

$$X_{1,2,w} \cdot \Big(U_1^{max} \;-\; U_1^{min} \Big) \;\geq\; U_2^{min} \qquad \textit{and} \qquad X_{1,2,w} \;\geq\; t_3 \tag{4.61}$$

The system may end up being in a state where the degrading of the existing class 1 connections with U_1^{max} units is not enough to meet the demands of the of the new

request (or $X_{1,2,w} = 0$). This would mean that the network should begin to select a certain number of class 2 connections with U_2^{max} units to be degraded, in addition to having all the class 1 connections degraded. This can be described by the system state transitions given in *Figures 4.19* and *4.20* for the case of class 1 and class 2 new requests, respectively, and equally applies in both networks.

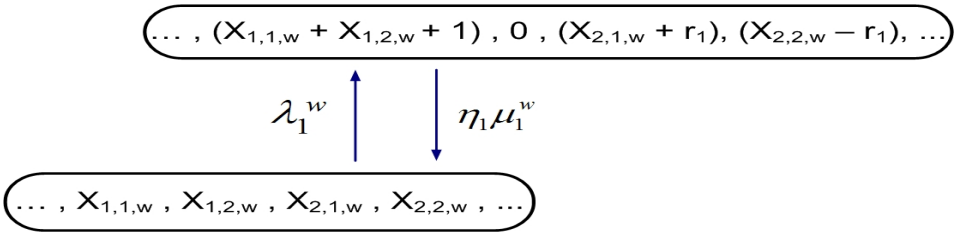

Fig. 4.19: The System State Transition for Allocating U_1^{min} to a New Class 1 Request By Degrading All the Existing Class 1 Connections and Some Class 2 Connections

$$\eta_1 = (X_{1,1,w} + X_{1,2,w} + 1) \cdot U_1^{min} \tag{4.62}$$

$$r_1 = \frac{U_1^{min}}{X_{1,2,w} \cdot (U_1^{max} - U_1^{min}) + (U_2^{max} - U_2^{min})} \tag{4.63}$$

The state transition given in *Figure 4.19* is for the case of the arrival of a new class 1 request, with r_1 in *(4.63)* being the number of class 2 connections with U_2^{max} that need to be degraded in order to accept the new class 1 request with U_1^{min}, in addition to degrading all of the existing class 1 connections. η_1 in *(4.62)* is the total number of units that are occupied by all of the class 1 connections in network w, in the forward state. This transition is only possible if the condition in *(4.64)* can be satisfied, i.e.

enough units can be accumulated to grant the new class 1 request with U_1^{min} units.

$$X_{1,2,w} \cdot (U_1^{max} - U_1^{min}) + X_{2,2,w} \cdot (U_2^{max} - U_2^{min}) \geq U_1^{min} \qquad and \qquad X_{2,2,w} \geq r_1 \tag{4.64}$$

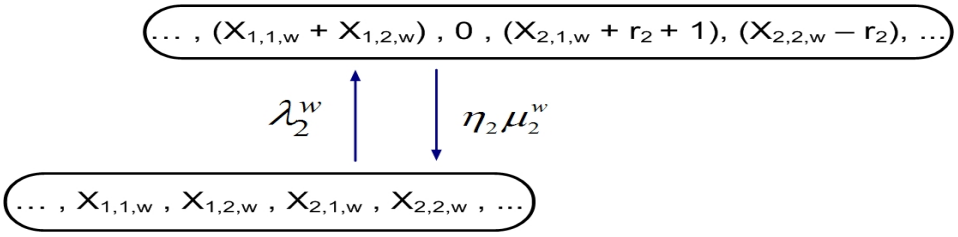

Fig. 4.20: The System State Transition for Allocating U_2^{min} to a New Class 2 Request By Degrading All the Existing Class 1 Connections and Some Class 2 Connections

$$\eta_2 = \left((X_{2,1,w} + r_2 + 1) \cdot U_2^{min} + (X_{2,2,w} - r_2) \cdot U_2^{max} \right) \tag{4.65}$$

$$r_2 = \frac{U_2^{min}}{X_{1,2,w} \cdot (U_1^{max} - U_1^{min}) + (U_2^{max} - U_2^{min})} \tag{4.66}$$

The state transition given in *Figure 4.20* is for the case of the arrival of a new class 2 request, with r_2 in *(4.66)* being the number of class 2 connections with U_2^{max} that need to be degraded in order to accept the new class 2 request with U_2^{min}, in addition to degrading all of the existing class 1 connections. η_2 in *(4.65)* is the total number of units that are occupied by all of the class 2 connections in network w, in the forward state. This transition is only possible if the condition in *(4.67)* can be satisfied, i.e.

enough units can be accumulated to grant the new class 2 request with U_2^{min} units.

$$X_{1,2,w} \cdot (U_1^{max} - U_1^{min}) + X_{2,2,w} \cdot (U_2^{max} - U_2^{min}) \geq U_2^{min} \qquad and \qquad X_{2,2,w} \geq r_2 \tag{4.67}$$

The system state transition that describes the case of a class i connection being transferred between the two networks is given in *Figure 4.21* .

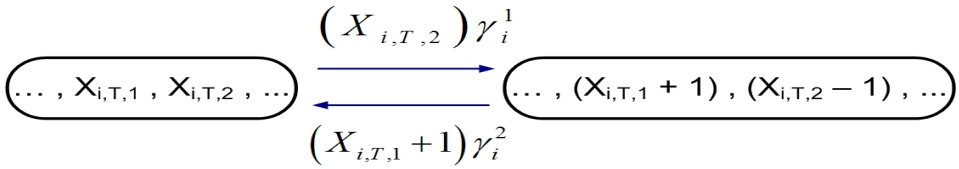

Fig. 4.21: The System State Transition for the Transfer of Connections Between Both Networks

$$where \quad X_{i,T,w} = \sum_{k=U_i^{min}}^{U_i^{max}} X_{i,k,w} \qquad \forall i, w \tag{4.68}$$

Following a similar approach to the one used in *Section 4.2.2* , the variables that represent the number of class i connections were lumped together in the manner given by *(4.68)* , since it is assumed that the transfer of a class i connection is independent of the number of units it is receiving. Moreover, a connection that has transferred to the new network may not necessarily receive the same number of units that was given by the previous network. In fact, it is assumed that from the point of view of the new network, the transferred connection would be treated as if it was a new connection request, and subject to the same bandwidth allocation policies given by the previous system state

transitions.

A summary of the bandwidth allocation algorithm for this model is given in *Figure 4.22* .

The overall system state transitions can be illustrated by a 4-dimensional state transition diagram and a clear construction of the diagram could not be made due to the complexity of the dimensions. A 2-D extraction of the overall system state transition diagram will instead be considered. This means that the state transitions for both classes of users in only one network will be illustrated, with the system exhibiting the same behavior in the other network. Hence, the 2-D state diagram will look at the behavior of the sub-system with the state vector $\mathbf{S}'_h = (X_{1,1,w}\ ,\ X_{1,2,w}\ ,\ X_{2,1,w}\ ,\ X_{2,2,w})$. The transitions showing the rate of connection transfers was omitted for clarity purposes, but remains to be included in the analysis. An example of the state transition diagram for this sub-system in network 1 is given in *Figure 4.23* , with the following numerical values used for the system parameters in this example.

$$B_1 = 6, \quad \text{with} \quad U_1^{min} = 1, \quad U_1^{max} = 2, \quad U_2^{min} = 2, \quad U_2^{max} = 3$$

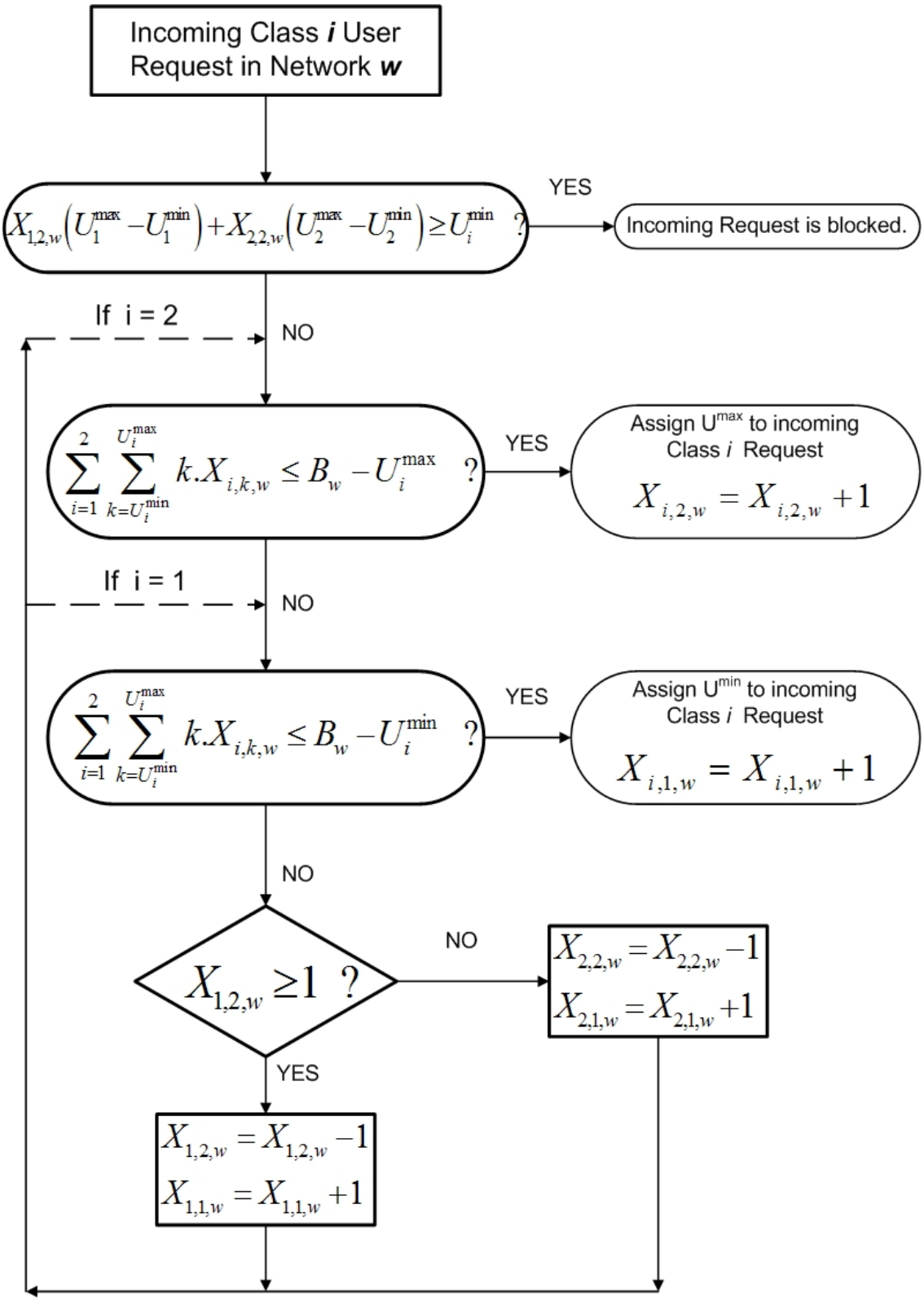

Fig. 4.22: The Bandwidth Allocation and Connection Degrading Algorithm

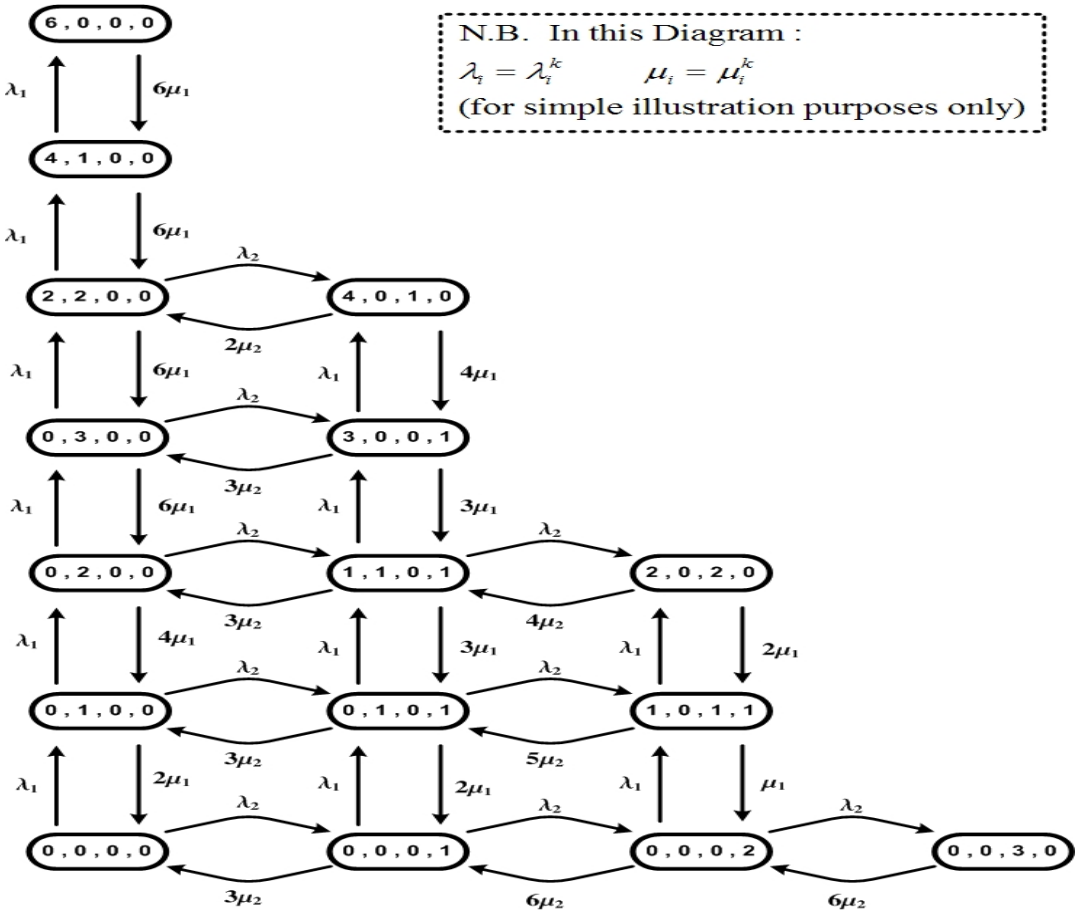

Fig. 4.23: The Example of the State Transition Diagram for Network 1 with Complete Sharing of Network Resource

4.3.3 The Analysis of the Model

The Markov Process for this model was also analyzed as a Quasi-Birth-Death process with a generator matrix given by Q_h.

$$\mathbf{Q}_h = \begin{bmatrix} q_{0,0} & q_{0,1} & & & \\ q_{1,0} & q_{1,1} & q_{1,2} & & \\ & q_{2,1} & q_{2,2} & q_{2,3} & \\ & & \ddots & \ddots & \ddots \\ & & & q_{f,f-1} & q_{f,f} \end{bmatrix} \tag{4.69}$$

$$where \quad f = \frac{B_1}{U_1^{min}} \tag{4.70}$$

The matrices $q_{m,n}$, $\forall m \neq n$, are rectangular matrices with varying dimensions, and $q_{m,m}$ are square matrices of varying dimensions and of the order

$$\left(\left\lfloor \frac{B_1 - mU_1^{min}}{U_2^{min}} \right\rfloor + 1 \right) \cdot \Delta_2 \tag{4.71}$$

$$where \Delta_2 = \left(\sum_{t=0}^{\frac{B_2}{U_2^{min}}} \left[\left\lfloor \frac{B_2 - tU_2^{min}}{U_1^{min}} \right\rfloor + 1 \right] \right) \tag{4.72}$$

m is the number of the number of class 1 subscribers in network 1 of the system, and is represented by the index of the rows in the QBD matrix Q_h. The matrices $q_{m,m-1}$, where $1 \leq m \leq f$, represents the departure of a class 1 subscriber from network 1 in

the system, such that

$$q_{m,m-1} = \begin{bmatrix} D_1(\alpha_1, m) & & & \\ & D_1(\alpha_1, m) & & \\ & & \ddots & \\ & & & D_1(\alpha_1, m) \end{bmatrix} \tag{4.73}$$

$$where, \quad \alpha_1 \; = \sum_{k=U_1^{min}}^{U_1^{max}} k \cdot X_{1,k,1} \quad ; \qquad 0 \leq \alpha_1 \leq B_1 \tag{4.74}$$

The dimensions of the matrices $q_{m,m-1}$ are given as

$$|q_{m,m-1}| \; = \; \Big(\Big(\; \left\lfloor \frac{B_1 - mU_1^{min}}{U_2^{min}} \right\rfloor \; + \; 1 \; \Big) \cdot \Delta_2 \Big) \times \Big(\Big(\; \left\lfloor \frac{B_1 - (m-1)U_1^{min}}{U_2^{min}} \right\rfloor \; + \; 1 \; \Big) \cdot \Delta_2 \Big) \tag{4.75}$$

The matrices $q_{m,m+1}$, where $0 \leq m \leq f-1$, represents the arrival of a class 1 subscriber into network 1 in the system, such that

$$q_{m,m+1} = \begin{bmatrix} E_1 & & & \\ & E_1 & & \\ & & \ddots & \\ & & & E_1 \end{bmatrix} \tag{4.76}$$

The dimensions of the matrices $q_{m,m+1}$ are given as

$$|q_{m,m+1}| \;=\; \left(\left(\;\left\lfloor\frac{B_1 - mU_1^{min}}{U_2^{min}}\right\rfloor\;+\;1\;\right)\cdot\Delta_2\right)\times\left(\left(\;\left\lfloor\frac{B_1-(m+1)U_1^{min}}{U_2^{min}}\right\rfloor\;+\;1\;\right)\cdot\Delta_2\right) \tag{4.77}$$

The matrices $q_{m,m}$, where $0 \leq m \leq f$, has the following form.

$$q_{m,m} = \begin{bmatrix} A_m(\alpha_1,\beta_1) & E_2 & & & & \\ D_2(\beta_1,1) & A_m(\alpha_1,\beta_1) & E_2 & & & \\ & D_2(\beta_1,2) & A_m(\alpha_1,\beta_1) & E_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & D_2\,(\beta_1,\eta) & A_m(\alpha_1,\beta_1) & \end{bmatrix} \tag{4.78}$$

$$where \quad \eta \;=\; \left\lfloor\frac{B_1 - mU_1^{min}}{U_2^{min}}\right\rfloor \tag{4.79}$$

The index of the rows in matrices $q_{m,n}$, $\forall m,n$, represent the number of class 2 subscribers in network 1, with the inner matrices being square matrices of the order Δ_2.

The matrices E_1 and E_2 represent the arrival of a class 1 and class 2 connection into network 1, respectively. The source of the arrivals could either be from new requests or the transfer of ongoing connections from network 2 to network 1.

$$E_1 = \begin{bmatrix} b_{0,0} & & & \\ b_{1,0} & b_{1,1} & & \\ & \ddots & \ddots & \\ & & b_{n,n-1} & b_{n,n} \end{bmatrix} \quad where \quad n = \frac{B_2}{U_1^{min}} \tag{4.80}$$

$$b_{x,x} = \begin{bmatrix} \lambda_1^1 & & & \\ & \lambda_1^1 & & \\ & & \ddots & \\ & & & \lambda_1^1 \end{bmatrix} \quad for \quad 0 \leq x \leq n \tag{4.81}$$

$$b_{x,x-1} = \begin{bmatrix} x\gamma_1^1 & & & \\ & x\gamma_1^1 & & \\ & & \ddots & \\ & & & x\gamma_1^1 \end{bmatrix} \quad for \quad 1 \leq x \leq n \tag{4.82}$$

$$E_2 = \begin{bmatrix} p_{0,0} & & & \\ & p_{1,1} & & \\ & & \ddots & \\ & & & p_{n,n} \end{bmatrix} \quad where \quad n = \frac{B_2}{U_1^{min}} \tag{4.83}$$

$$p_{y,y} = \begin{bmatrix} \lambda_2^1 & & & \\ \gamma_2^1 & \lambda_2^1 & & \\ & \ddots & \ddots & \\ & & \psi_2\gamma_2^1 & \lambda_2^1 \end{bmatrix} \quad for \quad 0 \leq y \leq n \tag{4.84}$$

$$where \quad \psi_2 = \left\lfloor \frac{B_2 - yU_1^{min}}{U_2^{min}} \right\rfloor \tag{4.85}$$

The index of the rows in the matrices E_1 and E_2 represent the number of class 1 users in network 2, while the index of the rows in matrices $b_{x,x}$, $b_{x,x-1}$, and $p_{y,y}$, represent the number of class 2 users in network 2. The matrices E_1 and E_2 are square matrices of the order Δ_2 , while the inner matrices $b_{x,x}$ and $p_{y,y}$ are square matrices of the order

$$\left(\left\lfloor \frac{B_2 - xU_1^{min}}{U_2^{min}} \right\rfloor + 1 \right) \quad for \quad b_{x,x} \tag{4.86}$$

$$\left(\left\lfloor \frac{B_2 - yU_1^{min}}{U_2^{min}} \right\rfloor + 1 \right) \quad for \quad p_{y,y} \tag{4.87}$$

The matrices $b_{x,x-1}$ are rectangular matrices with the following dimensions

$$\left(\left\lfloor \frac{B_2 - xU_1^{min}}{U_2^{min}} \right\rfloor + 1 \right) \times \left(\left\lfloor \frac{B_2 - (x-1)U_1^{min}}{U_2^{min}} \right\rfloor + 1 \right) \tag{4.88}$$

The matrices $D_1(\alpha_1, m)$ and $D_2(\beta_1, \theta)$ represent the departure of class 1 and class 2 connections from network 1, respectively. The users may have departed the network as

a result of service completion (or service completion within a cell as a result of handoff), or the transfer of its ongoing connection from network 1 to network 2.

$$D_1(\alpha_1, m) = \begin{bmatrix} d_{0,0} & d_{0,1} & & & \\ & d_{1,1} & d_{1,2} & & \\ & & \ddots & \ddots & \\ & & & & d_{z,z} \end{bmatrix} \quad where \quad z = \frac{B_2}{U_1^{min}} \tag{4.89}$$

$$d_{x,x} = \begin{bmatrix} \alpha_1\mu_1^1 & & & & \\ & \alpha_1\mu_1^1 & & & \\ & & \ddots & & \\ & & & & \alpha_1\mu_1^1 \end{bmatrix} \quad for \quad 0 \leq x \leq z \tag{4.90}$$

$$d_{x,x+1} = \begin{bmatrix} m\gamma_1^2 & & & & \\ & m\gamma_1^2 & & & \\ & & \ddots & & \\ & & & & m\gamma_1^2 \end{bmatrix} \quad for \quad 0 \leq x \leq z-1 \tag{4.91}$$

$$D_2(\beta_1, \theta) = \begin{bmatrix} g_{0,0} & & & \\ & g_{1,1} & & \\ & & \ddots & \\ & & & g_{z,z} \end{bmatrix} \quad where \quad z = \frac{B_2}{U_1^{min}} \tag{4.92}$$

$$g_{y,y} = \begin{bmatrix} \beta_1\mu_2^1 & \theta\gamma_2^2 & & \\ & \beta_1\mu_2^1 & \theta\gamma_2^2 & \\ & & \ddots & \ddots \\ & & & \beta_1\mu_2^1 \end{bmatrix} \quad for \quad 0 \leq y \leq z \tag{4.93}$$

$$where \quad \beta_1 = \sum_{k=U_1^{min}}^{U_1^{max}} k \cdot X_{2,k,1} \quad ; \qquad 0 \leq \beta_1 \leq B_1 \tag{4.94}$$

The matrices $d_{x,x}$ and $g_{y,y}$ are square matrices and of the same order given by equations (4.86) and (4.87), respectively. The matrices $d_{x,x+1}$ are rectangular matrices with the following dimensions

$$\left(\left\lfloor \frac{B_2 - xU_1^{min}}{U_2^{min}} \right\rfloor + 1 \right) \times \left(\left\lfloor \frac{B_2 - (x+1)U_1^{min}}{U_2^{min}} \right\rfloor + 1 \right) \tag{4.95}$$

The matrices $A_m(\alpha_1, \beta_1)$ has the following form.

$$A_m(\alpha_1, \beta_1) = \begin{bmatrix} a_{0,0} & a_{0,1} & & & \\ a_{1,0} & a_{1,1} & a_{1,2} & & \\ & a_{2,1} & a_{2,2} & a_{2,3} & \\ & & \ddots & \ddots & \ddots \\ & & & a_{z,z-1} & a_{z,z} \end{bmatrix} \quad where \quad z = \frac{B_2}{U_1^{min}} \tag{4.96}$$

$$a_{x,x+1} = \begin{bmatrix} \lambda_1^2 & & & \\ & \lambda_1^2 & & \\ & & \ddots & \\ & & & \lambda_1^2 \end{bmatrix} \quad for \quad 0 \leq x \leq z-1 \tag{4.97}$$

$$a_{x,x-1} = \begin{bmatrix} \varphi_1\mu_1^2 & & & \\ & \varphi_1\mu_1^2 & & \\ & & \ddots & \\ & & & \varphi_1\mu_1^2 \end{bmatrix} \quad for \quad 1 \leq x \leq z \tag{4.98}$$

$$where \quad \varphi_1 = \sum_{k=U_1^{min}}^{U_1^{max}} k \cdot X_{1,k,2} \quad ; \qquad 0 \leq \varphi_1 \leq B_2 \tag{4.99}$$

$$a_{x,x} = \begin{bmatrix} \Lambda_{m,x,0} & \lambda_2^2 & & & \\ \varphi_2\mu_2^2 & \Lambda_{m,x,1} & \lambda_2^2 & & \\ & \varphi_2\mu_2^2 & \Lambda_{m,x,2} & \lambda_2^2 & \\ & \ddots & \ddots & \ddots & \\ & & & \varphi_2\mu_2^2 & \Lambda_{m,x,R} \end{bmatrix} \quad with \;\; R = \left\lfloor \frac{B_2 - xU_1^{min}}{U_2^{min}} \right\rfloor \tag{4.100}$$

$$where \quad \varphi_2 = \sum_{k=U_1^{min}}^{U_1^{max}} k \cdot X_{2,k,2} \quad ; \qquad 0 \leq \varphi_2 \leq B_2 \tag{4.101}$$

$\Lambda_{m,x,z}$, for $0 \leq z \leq R$, is the negative of the sum of all the other elements that are in the same row as $\Lambda_{m,x,z}$ in the generator matrix $\mathbf{Q}_h$.

The steady-state distribution of the system with the generator matrix $\mathbf{Q}_h$ can be computed using the following,

$$0 = \Phi(\mathbf{S}_h) \cdot \mathbf{Q}_h \qquad \textit{and} \qquad \Phi(\mathbf{S}_h) \cdot \mathbf{e} = 1 \tag{4.102}$$

$\Phi(\mathbf{S}_h)$ is the steady-state probability vector of the system with the states $\mathbf{S}_h$, and contains the elements given by (4.103), i.e. the steady-state probabilities of the system. $\mathbf{e}$ is a column vector of 1.

$$p\Big(x_{i,k,w} \quad : \quad 1 \leq i \leq 2 \quad , \quad U_i^{min} \leq k \leq U_i^{max} \quad , \quad w = \{1,2\} \Big) \tag{4.103}$$

A product-form solution for the steady-state distribution could not be readily obtained due to the complexity of the model's structure, as well as its dimensions. One way of solving for the steady-state probability is through recursion techniques, such as the one proposed by the authors in [37]. However, Matlab was used to efficiently solve for the steady-state probability distribution.

4.3.4 The Performance Metrics of the System

The same performance metrics defined in *Section 4.2.4* will be used to analyzed the performance of the system with the steady-state probability distribution $\Phi(\mathbf{S}_h)$. A similar approach to the one made in *Section 4.2.4* will be used to re-write the system state vector $\mathbf{S}_h$ to the following

$$\widehat{\mathbf{S}_h} = \{ X_{i,w} , \forall i, w \ : \ X_{i,w} = \sum_{k=U_i^{min}}^{U_i^{max}} X_{i,k,w} \} = (X_{1,1} , X_{2,1} , X_{1,2} , X_{2,2}) \quad (4.104)$$

The modified system state vector, $\widehat{\mathbf{S}_h}$, does not alter in any way the definition of the system, and is used solely for the purpose of computing the blocking probabilities for class i subscribers in both networks, using the steady-state probability distribution $\Phi\left(\widehat{\mathbf{S}_h}\right)$, which now contains the elements $p\Big(x_{1,1} , x_{2,1} , x_{1,2} , x_{2,2} \Big)$.

$$P_b(1,1) = \{Blocking\ of\ Class\ 1\ connections\ in\ Network\ 1\} \quad (4.105)$$

$$= \sum_{x_{2,1}=0}^{\frac{B_1}{U_2^{min}}} \sum_{\forall x_{1,2}} \sum_{\forall x_{2,2}} p\left(\left\lfloor \frac{B_1 - x_{2,1}U_2^{min}}{U_1^{min}} \right\rfloor , x_{2,1} , x_{1,2} , x_{2,2} \right) \quad (4.106)$$

$$P_b(2,1) = \{Blocking\ of\ Class\ 2\ connections\ in\ Network\ 1\} \quad (4.107)$$

$$= \sum_{x_{1,1}=0}^{\frac{B_1}{U_1^{min}}} \sum_{\forall x_{1,2}} \sum_{\forall x_{2,2}} p\left(x_{1,1} , \left\lfloor \frac{B_1 - x_{1,1}U_1^{min}}{U_2^{min}} \right\rfloor , x_{1,2} , x_{2,2} \right) \quad (4.108)$$

$$P_b(1,2) = \{Blocking\ of\ Class\ 1\ connections\ in\ Network\ 2\} \quad (4.109)$$

$$= \sum_{\forall x_{1,1}} \sum_{\forall x_{2,1}} \sum_{x_{2,2}=0}^{\frac{B_2}{U_2^{min}}} p\left(x_{1,1}\ ,\ x_{2,1}\ ,\ \left\lfloor \frac{B_2 - x_{2,2}U_2^{min}}{U_1^{min}} \right\rfloor\ ,\ x_{2,2}\right) \quad (4.110)$$

$$P_b(2,2) = \{Blocking\ of\ Class\ 2\ connections\ in\ Network\ 2\} \quad (4.111)$$

$$= \sum_{\forall x_{1,1}} \sum_{\forall x_{2,1}} \sum_{x_{1,2}=0}^{\frac{B_2}{U_1^{min}}} p\left(x_{1,1}\ ,\ x_{2,1}\ ,\ x_{1,2}\ ,\ \left\lfloor \frac{B_2 - x_{1,2}U_1^{min}}{U_2^{min}} \right\rfloor\right) \quad (4.112)$$

The probability that a class i connection would be allocated U_i^{max} units by the network w upon initial connection (for both new connections, and those transferred from the other network), given by $P_{max}(i,k)$, is defined as follows. The steady-state probability distribution $\Phi\left(\widehat{\mathbf{S}_h}\right)$ was also used to compute the following probabilities.

$$P_{max}(1,1) = \sum_{x_{1,1}=0}^{\psi(1)} \sum_{x_{2,1}=0}^{\frac{B_1}{U_2^{max}}-1} \sum_{\forall x_{1,2}} \sum_{\forall x_{2,2}} p\left(x_{1,1}\ ,\ x_{2,1}\ ,\ x_{1,2}\ ,\ x_{2,2}\right) \quad (4.113)$$

$$P_{max}(1,2) = \sum_{\forall x_{1,1}} \sum_{\forall x_{2,1}} \sum_{x_{1,2}=0}^{\psi(2)} \sum_{x_{2,2}=0}^{\frac{B_2}{U_2^{max}}-1} p\left(x_{1,1}\ ,\ x_{2,1}\ ,\ x_{1,2}\ ,\ x_{2,2}\right) \quad (4.114)$$

$$where \quad \psi(w) \;=\; \left\lfloor \frac{B_w \;-\; x_{2,k} U_2^{max}}{U_1^{max}} \right\rfloor \;-\; 1 \tag{4.115}$$

$$P_{max}(2,1) \;=\; \sum_{x_{1,1}=0}^{\vartheta(1)} \; \sum_{x_{2,1}=0}^{\chi(1)} \; \sum_{\forall x_{1,2}} \; \sum_{\forall x_{2,2}} \pi\left(x_{1,1}\;,\; x_{2,1}\;,\; x_{1,2}\;,\; x_{2,2}\right) \tag{4.116}$$

$$P_{max}(2,2) \;=\; \sum_{\forall x_{1,1}} \; \sum_{\forall x_{2,1}} \; \sum_{x_{1,2}=0}^{\vartheta(2)} \; \sum_{x_{2,2}=0}^{\chi(2)} \pi\left(x_{1,1}\;,\; x_{2,1}\;,\; x_{1,2}\;,\; x_{2,2}\right) \tag{4.117}$$

$$where \quad \chi(w) \;=\; \max\left\{0\;,\; \left(\left\lfloor \frac{B_w \;-\; x_{1,w} U_1^{min}}{U_2^{max}} \right\rfloor \;-\; 1 \right) \right\} \tag{4.118}$$

$$and \quad \vartheta(w) \;=\; \left\lfloor \frac{B_w \;-\; U_2^{max}}{U_1^{min}} \right\rfloor \tag{4.119}$$

Finally, the Degrade Level $E_k(i)$ for class i subscribers in network k is defined as

$$E_w(i) \;=\; \sum_{\mathbf{s}} \; p\left(\mathbf{s}\right) \left(\frac{U_i^{max} \cdot x_{i,2,w} \;+\; U_i^{min} \cdot x_{i,1,w}}{U_i^{max} \cdot \left(x_{i,1,w} \;+\; x_{i,2,w}\right)} \right) \tag{4.120}$$

for all $\quad \mathbf{s} \in \Phi(\mathbf{S}_h)\;, \quad \mathbf{s} \;=\; \left(x_{1,1,1}\;,\; x_{1,2,1}\;,\; x_{2,1,1}\;,\; x_{2,2,1}\;,\; x_{1,1,2}\;,\; x_{1,2,2}\;,\; x_{2,1,2}\;,\; x_{2,2,2}\right)$

4.3.5 Numerical Examples

In this Section, various numerical results will be presented for the system of described in *Section 4.3.1* . The results were obtained using the performance parameters defined in *Section 4.3.4* .

The following values for the network parameters will be assumed throughout the analysis ,while keeping the remaining system parameters (i.e. arrival, service, and connection transfer rates) consistent. These values can be shown to satisfy the requirements given by *(4.45)* to *(4.47)* .

$$B_1 = 6 \qquad B_2 = 6 \qquad U_1^{min} = 1 \qquad U_1^{max} = 2 \qquad U_2^{min} = 2 \qquad U_2^{max} = 3$$

The choice of network parameters may seem unreasonable but they were chosen for the purpose of showing certain properties in the behavior of the system. The numerical values that were obtained in the examples may be adjusted by altering the values for the network parameters, with the behaviors concluded in this Section remaining the same.

The first set of graphs given in *Figures 4.24* to *4.27* shows the blocking probabilities for both classes of subscriptions in both networks, corresponding to the varying of the arrival rates of class 1 and class 2 subscribers in network 1 and network 2, respectively. What is interesting to observe in these graphs is that unlike the results of the previous system in *Section 4.2.6* , the increased traffic of one class of subscribers influences the blocking probabilities for both classes in both networks. Such a behavior is plausible since the network resources are completely shared amongst the users of all classes.

The graphs also show class 2 subscribers having a higher blocking probability than class 1 subscribers. This can be explained by observing that class 2 users require almost double the units of class 1 , in accordance with the parameters used to generate the results. Since the total bandwidth units are shared amongst all the classes of users, a class 1 user is more likely to find enough units to meet its request than class 2 users. Furthermore, when the system is at a stage where further class 2 requests are blocked, it might still be able to attend to some class 1 connection-requests. In other words, there may not be enough units to accept a new class 2 connection but the system might just have enough units for some new class 1 connections.

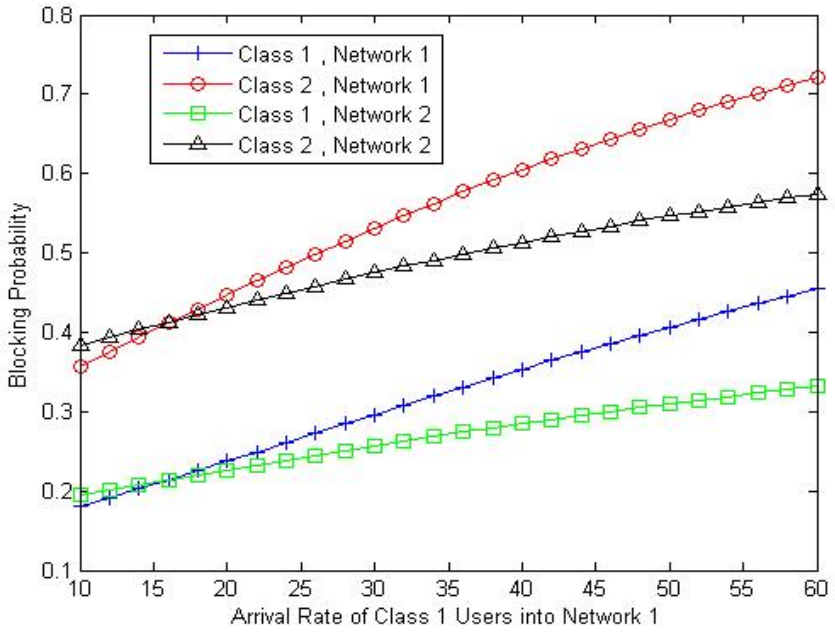

Fig. 4.24: A Graph Showing the Blocking Probabilities Corresponding to Varying Arrival Rates of Class 1 Users in Network 1

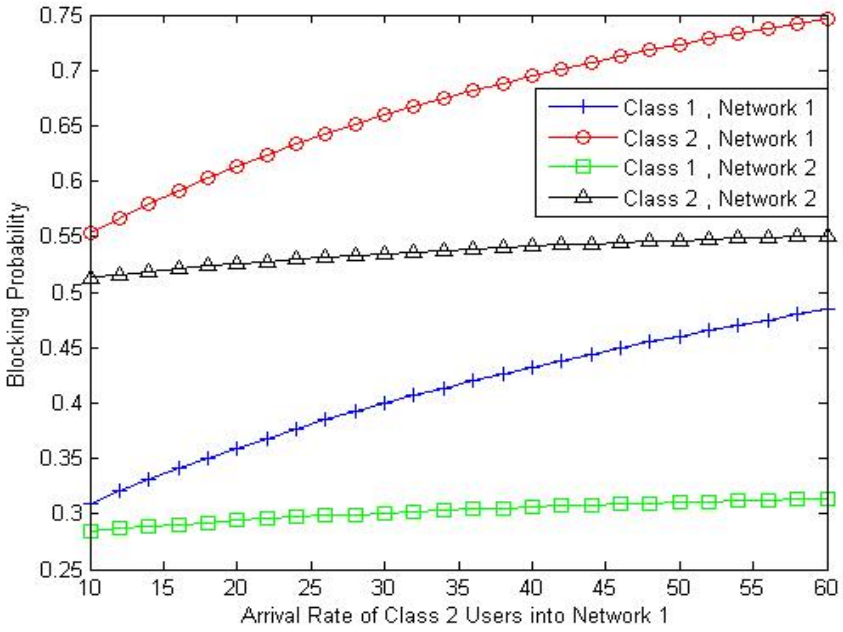

Fig. 4.25: A Graph Showing the Blocking Probabilities Corresponding to Varying the Arrival Rates of Class 2 Users in Network 1

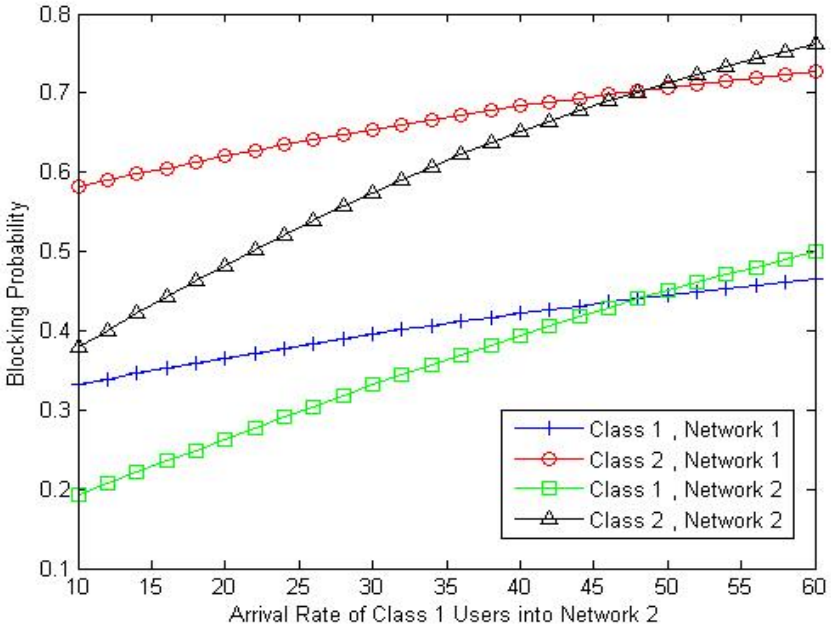

Fig. 4.26: A Graph Showing the Blocking Probabilities Corresponding to Varying the Arrival Rates of Class 1 Users in Network 2

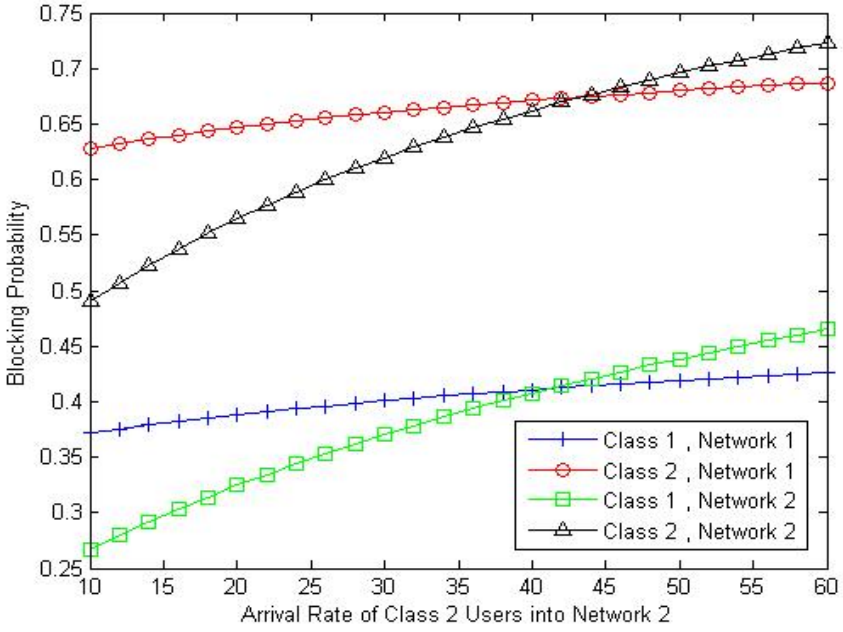

Fig. 4.27: A Graph Showing the Blocking Probabilities Corresponding to Varying the Arrival Rates of Class 2 Users in Network 2

The probability of obtaining U_i^{max} units for both classes of subscribers in both networks, corresponding to the varying of the arrival rates of class 1 and class 2 users in network 1 and network 2, are shown by the graphs in *Figures 4.28* to *4.31*, respectively.

The results from these graphs indicate that these probabilities start reducing as the traffic in the system increases. In addition, the probabilities of a new class 2 connection obtaining U_2^{max} units is almost always greater than those for class 1 subscriptions. Such a behavior is to be expected since the bandwidth allocation policy for this system always tries to grant class 2 users with the maximum number of units, even at the cost of degrading existing class 1 connections.

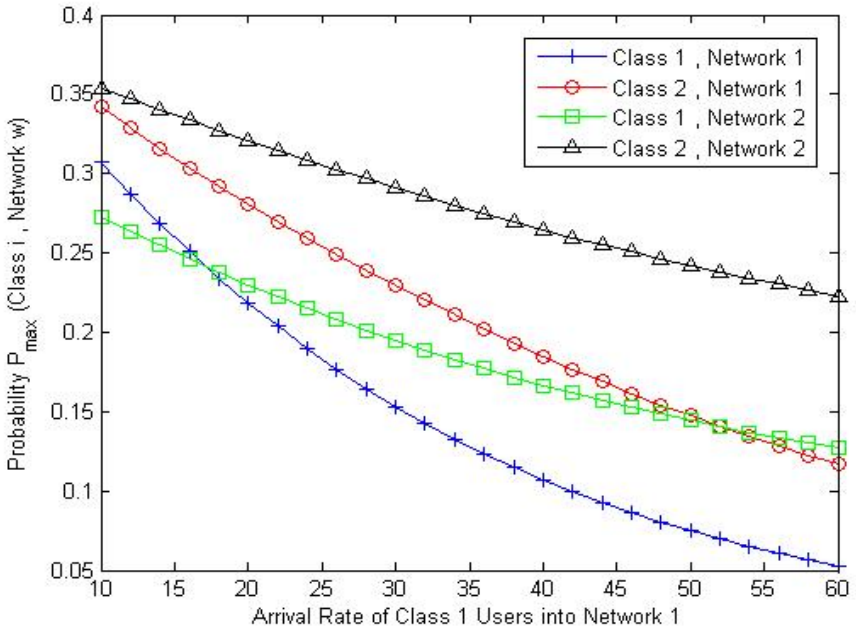

Fig. 4.28: A Graph Showing the Probabilities of Obtaining U_i^{max} Units Upon Initial Connection, Corresponding to Varying Arrival Rates of Class 1 Users in Network 1

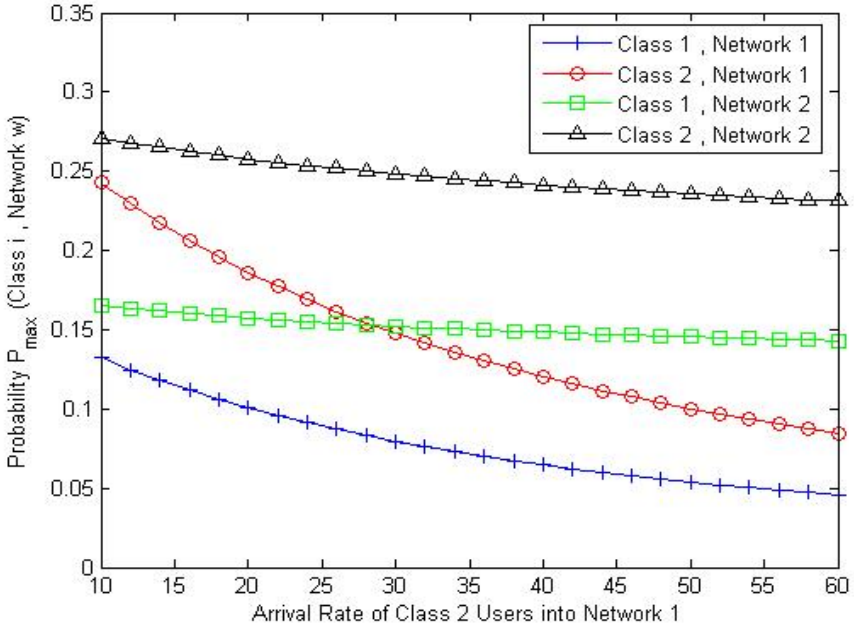

Fig. 4.29: A Graph Showing the Probabilities $P_{max}(i, w)$, Corresponding to Varying the Arrival Rates of Class 2 Users in Network 1

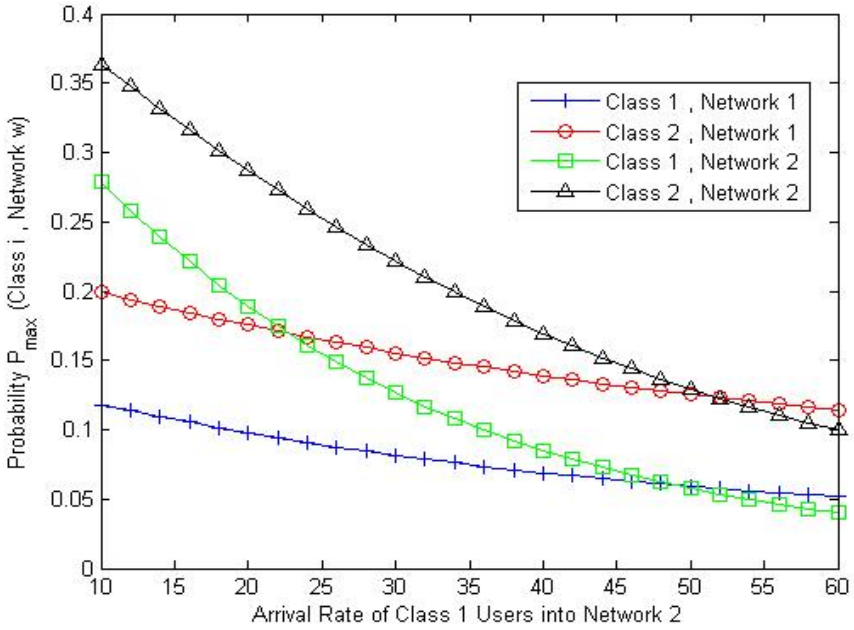

Fig. 4.30: A Graph Showing the Probabilities $P_{max}(i, w)$, Corresponding to Varying the Arrival Rates of Class 1 Users in Network 2

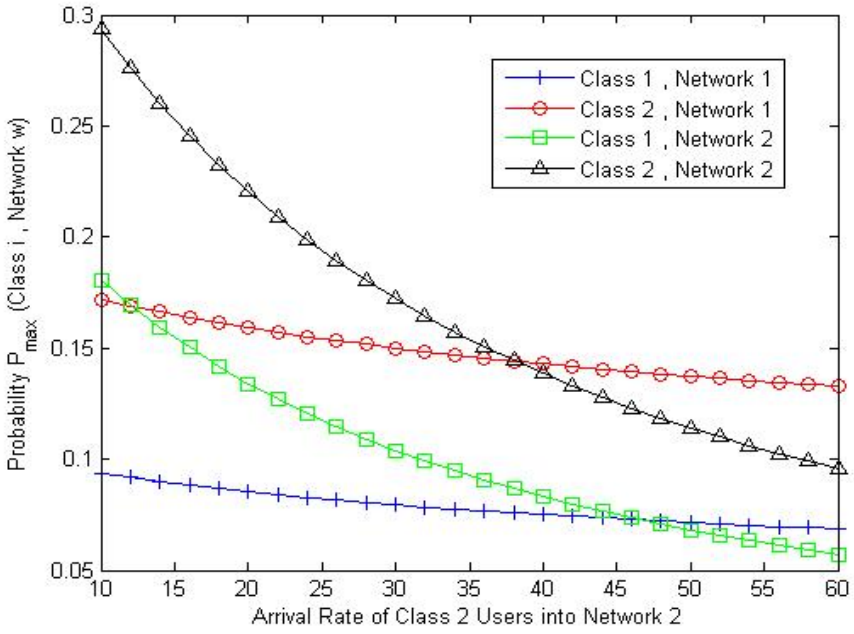

Fig. 4.31: A Graph Showing the Probabilities $P_{max}(i, w)$, Corresponding to Varying the Arrival Rates of Class 2 Users in Network 2

The degrade levels for both classes of subscribers in both networks, corresponding to the varying of the arrival rates of class 1 and class 2 users in network 1 and network 2, are shown by the graphs in *Figures 4.32* to *4.35*, respectively.

As the traffic of class 1 subscribers in either network increases (*Figures 4.32* and *4.34*), the degrade level, or the overall level of satisfaction for that class of subscribers, starts to reduce at a very low rate after having increased to a maximum level. But the levels for the class 2 subscribers start reducing at a greater rate. Note that the results assume a constant traffic rate of class 2 subscribers. However, the overall levels of satisfaction for class 2 subscribers in both networks reduces considerably.

The results were different when analyzing the behavior of the performance for increased traffic rates of class 2 connections (*Figures 4.33* and *4.35*). These graphs show how the degrade levels in both networks continue to increase for class 2 users until it reaches a somewhat steady level, while the degrade levels for class 1 users continue to decrease. Even though both graphs seem to indicate that the degrade levels for class 2 connections will reach some steady level, it actually will start to decline very slowly when the arrival rates further increases for class 2 subscribers. This can be observed in the next set of graphs given in *Section 4.4* .

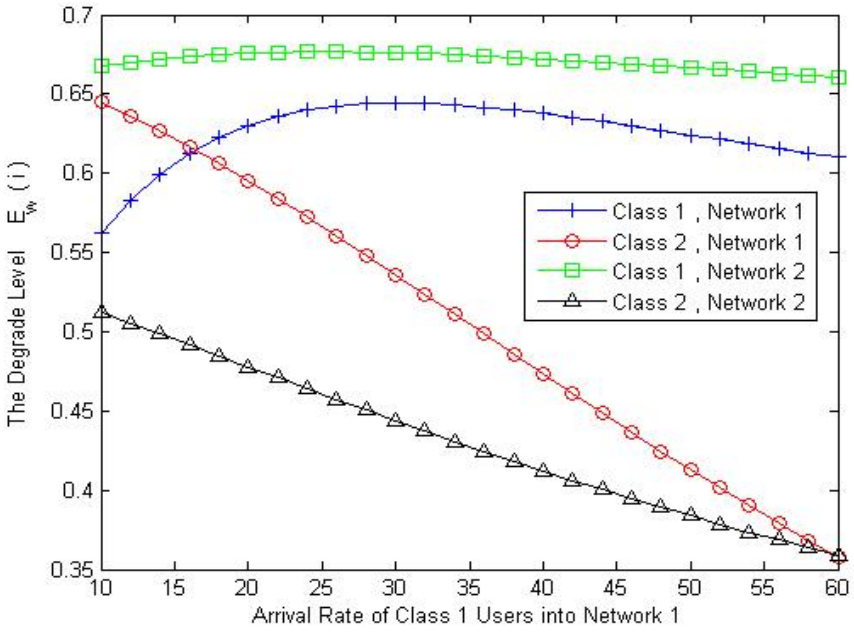

Fig. 4.32: A Graph Showing the Degrade Levels, Corresponding to Varying Arrival Rates of Class 1 Users in Network 1

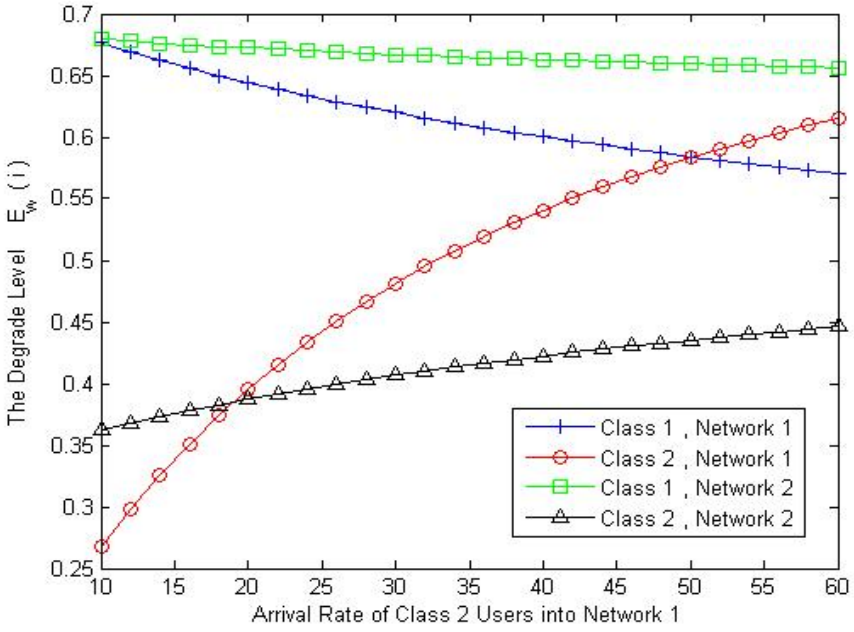

Fig. 4.33: A Graph Showing the Degrade Levels, Corresponding to Varying the Arrival Rates of Class 2 Users in Network 1

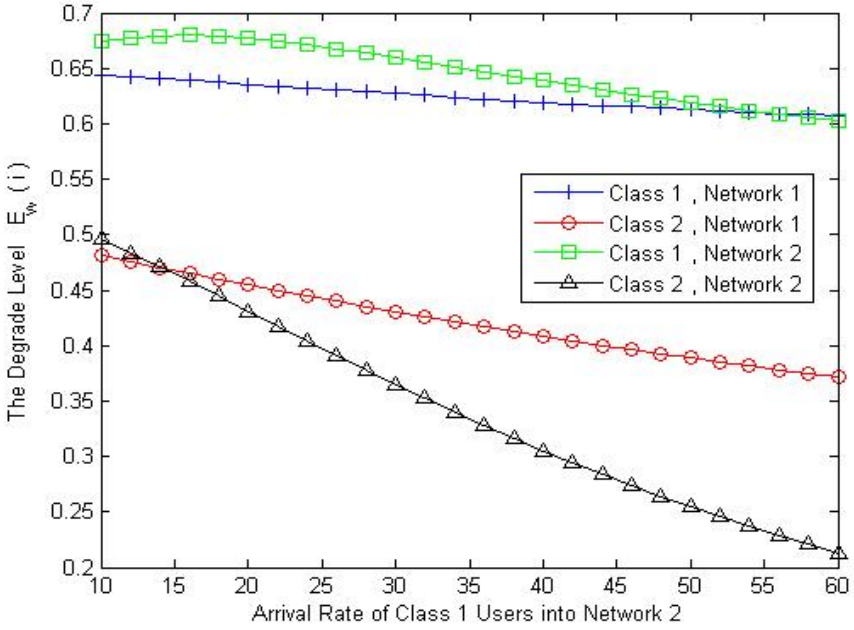

Fig. 4.34: A Graph Showing the Degrade Levels, Corresponding to Varying the Arrival Rates of Class 1 Users in Network 2

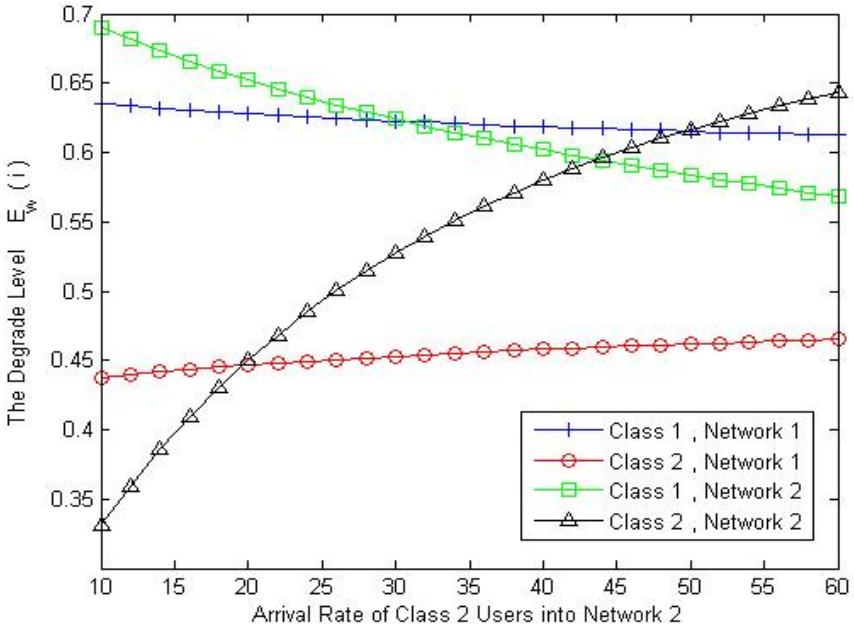

Fig. 4.35: A Graph Showing the Degrade Levels, Corresponding to Varying the Arrival Rates of Class 2 Users in Network 2

The relationship between the three system performance parameters $P_b(i, w)$, $P_{max}(i, w)$, and $E_w(i)$, are shown in the following graphs for each class of users in each network, given in *Figures 4.36* to *4.39*, respectively.

The results given by these sets of graphs focuses on illustrating how the system performance parameters behave for each class of subscribers in each network, while only varying their corresponding traffic rates, and assuming that all other rates remain constant. The graphs show the same behaviors explained earlier in the Section. In addition, the graphs clearly show how the degrade levels reaches an almost steady level for class 2 users under heavy network traffic, while it slowly reduces for class 1 users. This implies that a constant satisfaction level could be maintained for class 2 users under heavy traffic conditions.

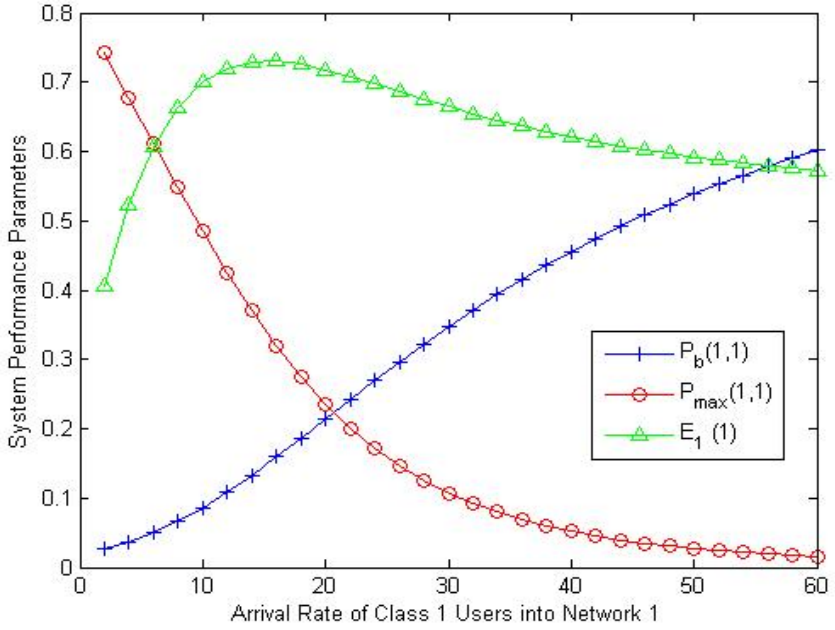

Fig. 4.36: A Graph Showing the System Performance Measures for Class 1 Users in Network 1

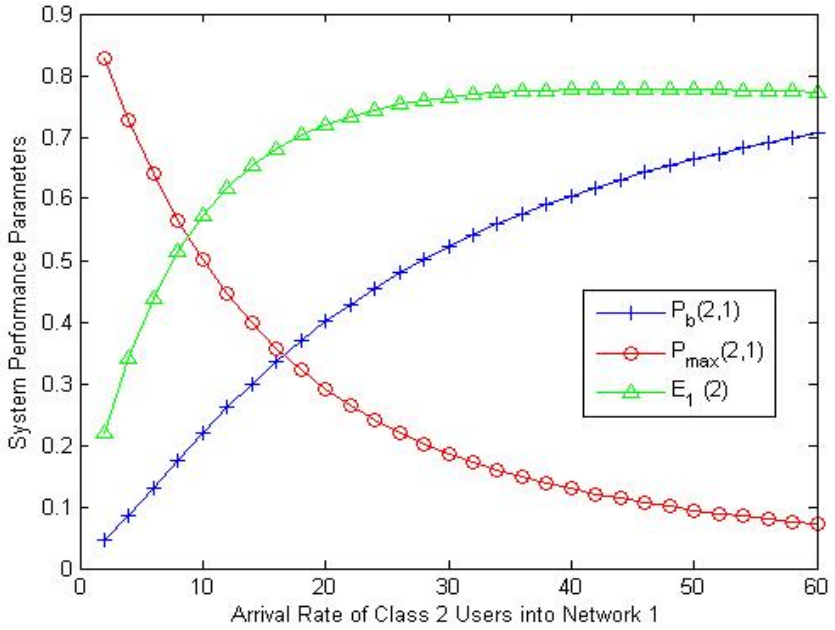

Fig. 4.37: A Graph Showing the System Performance Measures for Class 2 Users in Network 1

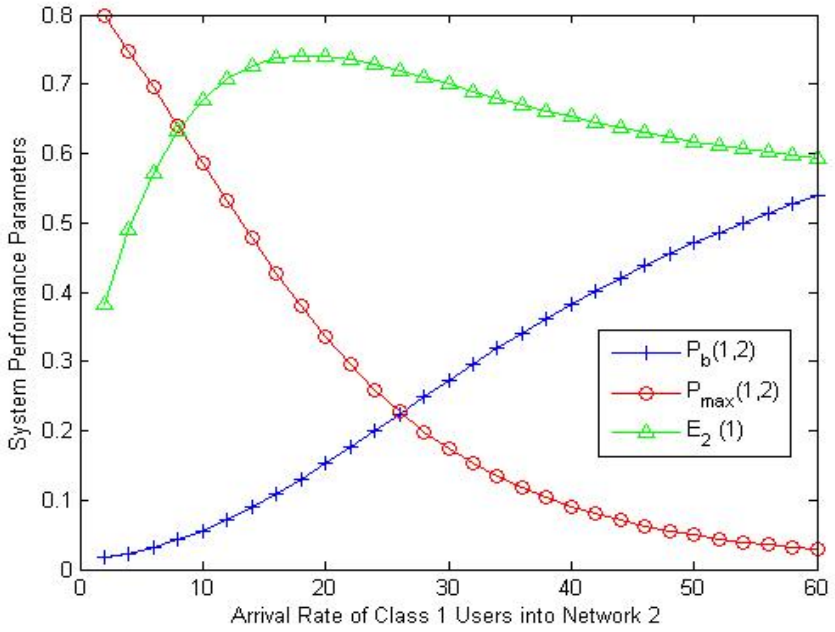

Fig. 4.38: A Graph Showing the System Performance Measures for Class 1 Users in Network 2

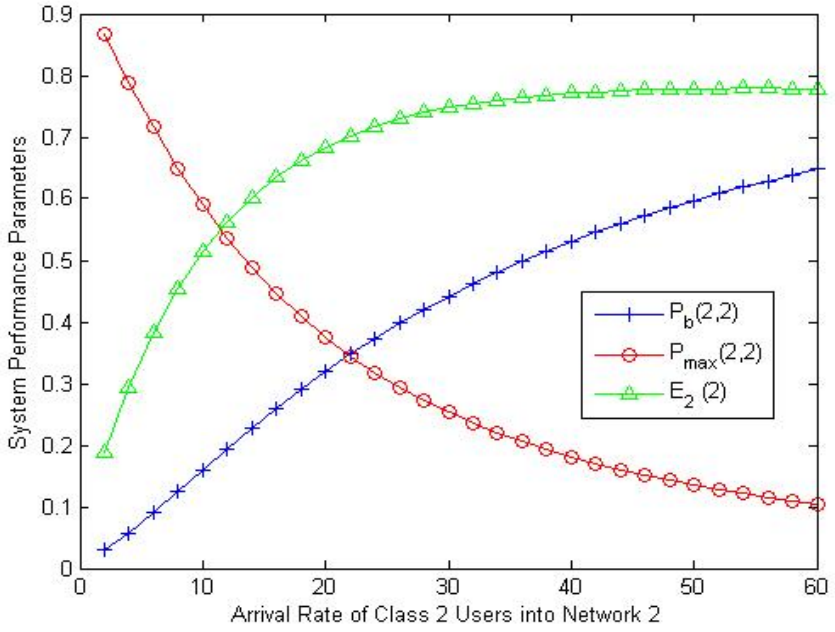

Fig. 4.39: A Graph Showing the System Performance Measures for Class 2 Users in Network 2

4.4 Comparison Between the Two Systems

In this Section, the performance of the two different systems described in *Sections 4.2* and *4.3* was compared in terms of the system performance measures defined for each of the models. The main differences between the two systems was in the allocation of the total resource B_w in each network (i.e. completely partitioned or shared for each class of subscribers), along with the bandwidth-unit allocation and connection degradation policies adopted in each of the two systems. Numerical examples will only be given for the case of the system performance in network 1, since the behavior for the different classes of users in one network was found to be similar to the ones in the other network. The system parameters used in the computation of the performance measures were kept consistent throughout the analysis.

Figures 4.40 and *4.41* compares the blocking probabilities for both class 1 and class 2 users in network 1 , respectively. Both graphs show an overall lower blocking probability in the system where the resources are completely shared amongst all the users. However, both systems seem to behave in a similar manner when the traffic is low, in terms of the blocking probabilities. In fact, the system with the resources completely partitioned is seen to perform slightly better than the other system when the incoming traffic is very low (*see Figure 4.40*) .

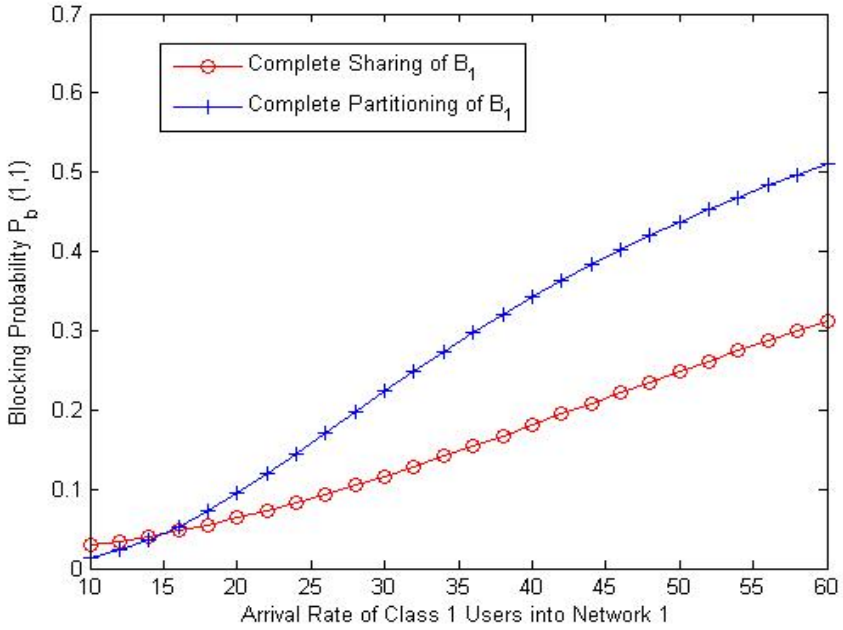

Fig. 4.40: A Graph Comparing the Blocking Probabilities for Class 1 Users in Network 1

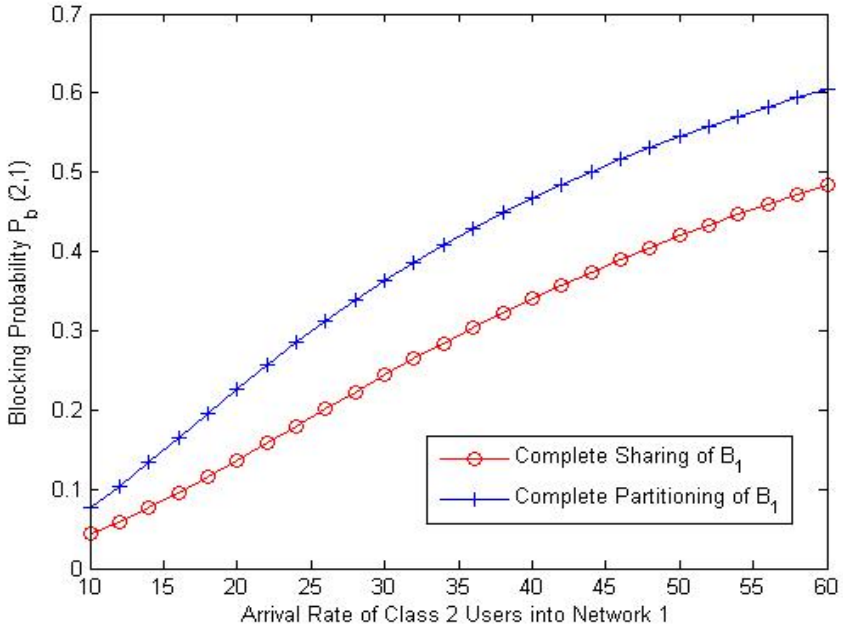

Fig. 4.41: A Graph Comparing the Blocking Probabilities for Class 2 Users in Network 1

Figures 4.42 and *4.43* compare the probabilities of obtaining U_i^{max} units upon initial connection, for both class 1 and class 2 users in network 1 , respectively. Both graphs also show that the system where the resources are completely shared performs better on the overall. The graphs also show how the performance of both systems become indifferent under heavy traffic conditions.

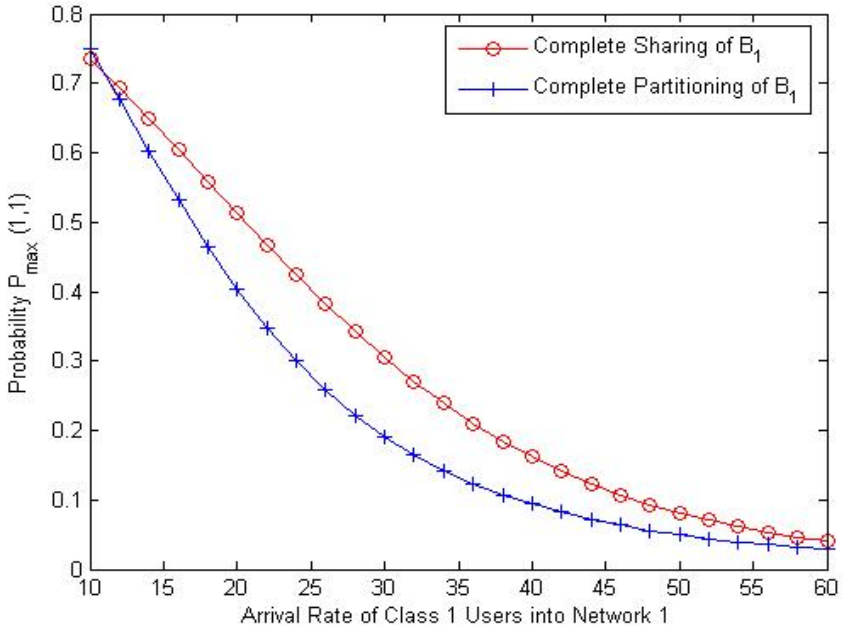

Fig. 4.42: A Graph Comparing the Probabilities of Obtaining Maximum Level of Service Upon Initial Connection, for Class 1 Users in Network 1

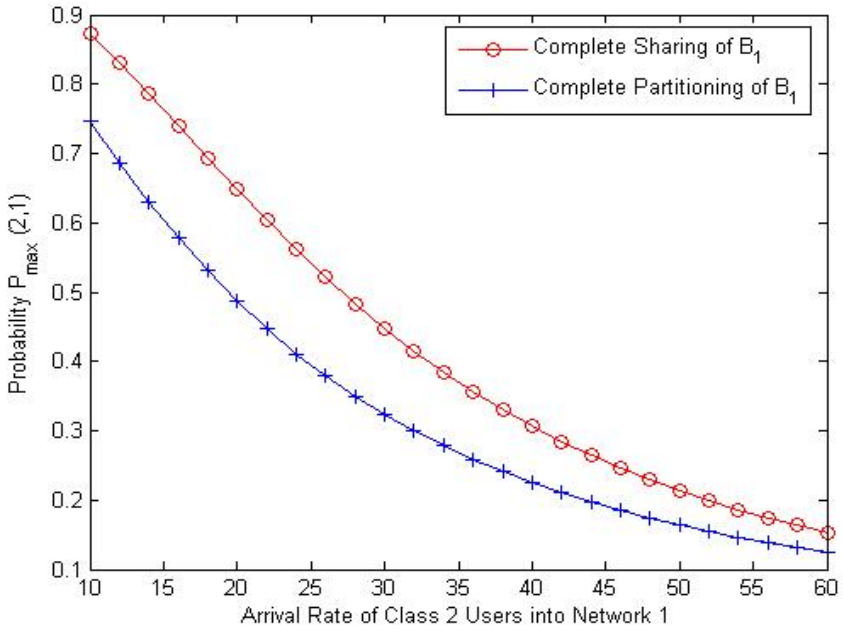

Fig. 4.43: A Graph Comparing the Probabilities of Obtaining Maximum Level of Service Upon Initial Connection, for Class 2 Users in Network 1

Figures 4.44 and *4.45* compares the degrade levels for both class 1 and class 2 users in network 1 , respectively. Both graphs generally show higher levels of overall satisfaction for the system where the resource is completely shared, and under heavy traffic conditions. In addition, *Figure 4.44* shows how an increase in the traffic rates for class 1 subscribers causes a considerable drop in the overall levels of satisfaction, with that level slowly declining for the case of class 2 subscribers shown in *Figure 4.45* . The reason behind the differences between the two behaviors is due to the values U_i^{min} and U_i^{max} that were used for generating the numerical results. A class 2 connection that is utilizing 2 bandwidth units instead of 3 corresponds to a 33% drop in QoS (and satisfaction), and a class 1 connection that is utilizing 1 bandwidth unit instead of 2 corresponds to a 50% drop in QoS. Hence, class 1 subcribers are expected to be a lot less satisfied with their minimum allowable bandwidth, when compared with class 2 subscribers. Therefore, the rate of decline of the overall level of satisfaction is expected to be higher for class 1 connections, as the number of those connections in the system increases, which eventually forces all those connections to utilize their minimum allowable bandwidth.

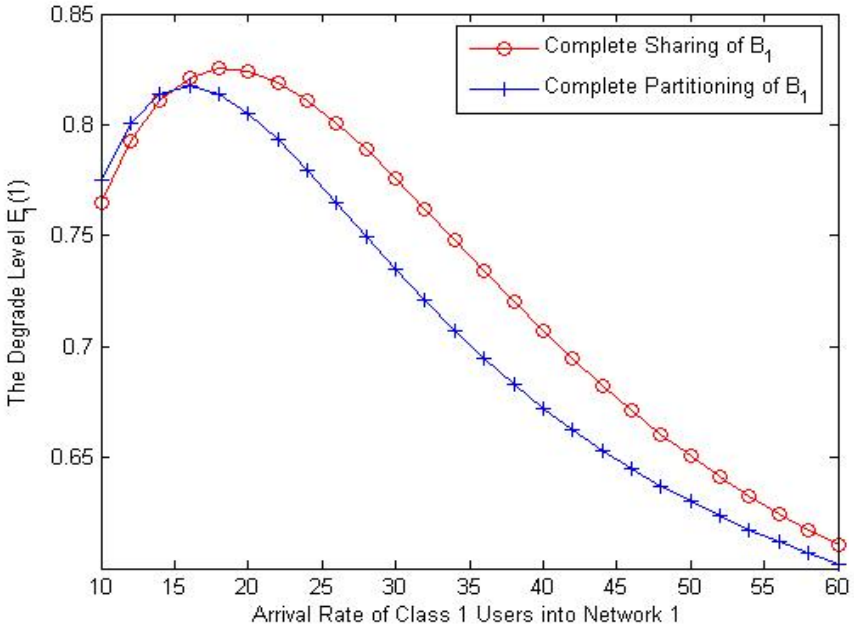

Fig. 4.44: A Graph Comparing the Degrade Levels for Class 1 Users in Network 1

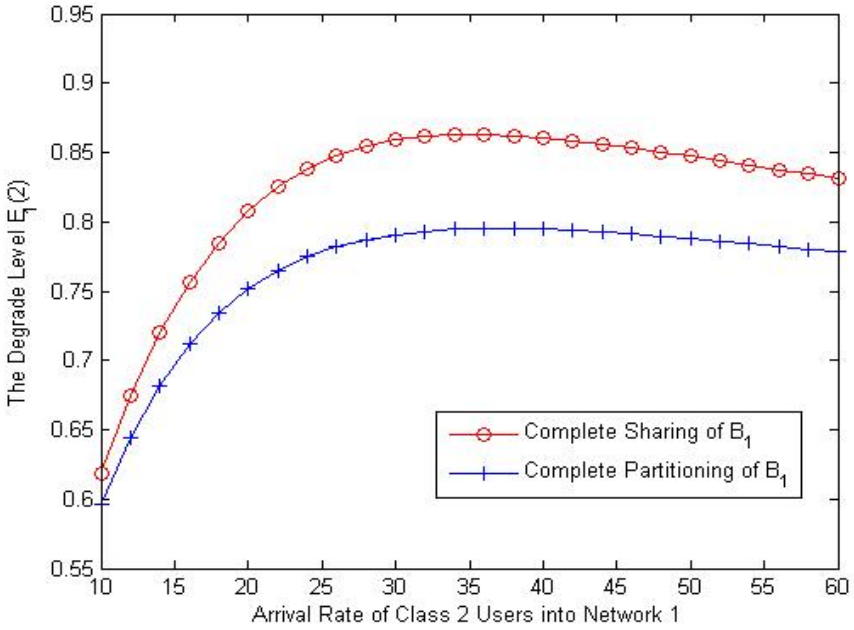

Fig. 4.45: A Graph Comparing the Degrade Levels for Class 2 Users in Network 1

5. CONCLUDING REMARKS

5.1 Conclusions and Comments

The results obtained from the numerical examples seem to indicate that an overall improvement is achieved if the complete sharing of the network's resources was allowed, as opposed to having the resource partitioned for each class of subscriptions. In addition, the performance in one network seems to influence the performance in the other network. However, the performance for one class of users has no impact on the performance of the other classes, for the case where the resources are completely partitioned in the system. This can be seen as an advantage of employing such a system, since it can give the network operators the ability to exercise a better control on their resources.

The examples for case of the system where the resources are completely shared have also shown how the performance of one class of subscribers can have a negative impact on the performance of the other class. In some cases, the performance for both classes can be at its worst under heavy network traffic conditions. The network-providers might have to look into ways of controlling the distribution of the different classes of subscriptions to the network-users. In other words, it could serve in the best interest of

everyone if the population of class 2 users were kept at a certain level, in order to avoid a high traffic of class 2 requests. One way of achieving this is by assigning appropriate costs for each subscription class, as explained in *Section 3.8* .

Even though having a higher subscription-class is beneficial in terms of the amount of bandwidth that is received on the average, it does however increase the chances of having those connections blocked from service, when compared with the lower subscriptions. This can be observed in the results given by the examples in *Section 4.3.5* . One way of reducing the blocking probabilities for those higher classes would be to again control the amount of traffic for each of the subscriptions by assigning appropriate prices for the subscriptions.

5.2 Summary of Contributions

The work presented in this thesis focused on developing an adaptive bandwidth allocation policy for subscription-based connections, unlike much of the previous work which looked at policies that were service-based. The previous work that is discussed in *Section 1.2* assumed that the user-demands are homogeneous in the sense that all users expect an equal level of QoS, relative to the user's running application. However, future wireless networks are projected to cater for a variety of user-requirements that is partly controlled by the level of service that the users can subscribe to. In addition, these policies were adopted for a system with two integrated wireless networks.

The systems described in *Section 4* had a particular adaptive bandwidth allocation

policy that was developed specifically for each of the different systems. In addition to analyzing each of the two systems individually, the performance of both systems were compared.

5.3 Proposal for Future Work

One of the areas that could be explored in a later work is to consider a hybrid of the systems described in *Section 4* . This could involve setting-up a system where part of the total network resources are shared amongst all classes of users, with the remaining resources partitioned for each class of subscribers. These partitioned units could be considered as guard channels for the different classes of requests.

In terms of partitioning the resources, another extension to the model could include having the partitioning scheme being adaptive, as described in the framework given in *Section 3* . An example would be to have the partitioning of the network resources being adaptive towards the traffic in the system. The partitioning scheme could further incorporate the idea of allowing some units to be borrowed from those that have been reserved for other classes of users.

The system with the complete sharing of resources could also be extended to consider multiple subscription classes. It was not readily possible to do this extension with the current model due to the complexity of the structure. Some preliminary work was done which considered extending the model to the case of three classes of subscriptions, and it was found that such an extension suffers from an exploding state-space. Moreover, there

was some difficulty with choosing a suitable adaptive bandwidth allocation policy for the case of three classes of subscriptions, in terms of assigning various levels of priority for each of the three classes.

The bandwidth allocation policies could also be further explored in order to find a policy that achieves a certain level of fairness amongst the various classes of subscriptions. This might mean having to define what is really meant by a "fair treatment" for the case of a system with multiple classes of subscribers and multiple networks.

Other performance metrics could also be defined to help with better understanding the performance of a system with heterogeneous wireless networks, and heterogeneous users.

REFERENCES

[1] K. Tachikawa, "A Perspective on the Evolution of Mobile Communications", IEEE Communications Magazine, October 2003.

[2] K. Pahlavan., P. Krishnamurthy, et al, "Handoff in Hybrid Mobile Data Networks", IEEE Personal Communications Volume: 7, Issue: 2, April 2000, pp.34-47.

[3] M. Birchler, P. P. Smyth et al. "Future of Mobile and Wireless Communications", BT Technology Journal,Vol. 21, No.3, July 2003.

[4] T. Otsu, I. Okajima, N. Umeda and Y. Yamao, "Network Architecture for Mobile Communications Systems Beyond IMT-2000", IEEE Personal Communications, October 2001.

[5] N. Nakajima and Y. Yamao, "Development for 4th Generation Mobile Communications", Wireless Communications and Mobile Computing, issue 1, 2001.

[6] Suk Yu Hui and Kai Hau Yeung, "Challenges in the Migration to 4G Mobile Systems", IEEE Communications Magazine, December 2003.

[7] M. Stemm and R. H. Katz, "Vertical Handoffs in Wireless Overlay Networks", Mobile Networks and Applications 3, (1998), pp. 335-350

[8] Kurt Aretz, Martin Haardt, et al., "The Future of Wireless Communications Beyond the Third Generation", Elsevier Computer Networks 37 (2001), pp. 83-92

[9] Th. Zahariadis, "Evolution of the Wireless PAN and LAN Standards", Elsevier Computer Standards & Interfaces No. 26, 2004.

[10] A. Duda and C. J. Sreenan, "Challenges for Quality of Service in Next Generation Mobile Networks", Information Technology and Telecommunications Conference 2003.

[11] J. Tourrilhes and C. Carter "P-Handoff: A Protocol for Fine Grained Peer-to-Peer Vertical Handoff", Personal, Indoor and Mobile Radio Communications, 2002. The 13th IEEE International Symposium on , Volume: 2 , 15-18 Sept. 2002.

[12] M. Frodigh et al., "Future-Generation Wireless Networks", IEEE Personal Communications, October 2001.

[13] H. Hsieh and R. Sivakumar, "Internetworking WWANs and WLANs in Next Generation Wireless Data Networks", In Proceedings of 3G Wireless and Beyond, San Francisco, CA USA, May 2002.

[14] M. Buddhikot, G. Chandranmenon et al., "Integration of 802.11 and Third-Generation Wireless Data Networks", In Proceedings of the IEEE INFOCOM, 2003.

[15] A. Salkintzis, C. Fors, and R. Pazhyannur, "WLAN-GPRS Integration for Next-Generation Mobile Data Networks", IEEE Wireless Communications, October 2002.

[16] T. Zhang and P. Agrawal, "IP-Based Base Stations And Soft Handoff in All-IP Wireless Networks", IEEE Personal Communications, October 2001.

[17] M.Salamah, F. Tansu and N. Khalil, "Buffering Requirements for Lossless Vertical Handoffs in Wireless Overlay Networks", Vehicular Technology Conference, 2003. VTC 2003-Spring. The 57th IEEE Semiannual , Volume: 3 , 22-25 April 2003.

[18] Qian Zhang et al. "Efficient Mobility Management for Vertical Handoff between WWAN and WLAN", IEEE Communications Magazine, November 2003.

[19] Mohr, W.; Konhauser, W. , "Access network evolution beyond third generation mobile communications", IEEE Communications Magazine, December 2000.

[20] R. Becher, M. Dillinger and M. Haardt, "Broad-Band Wireless Access and Future Communication Networks", Proceedings of the IEEE, vol. 89, No.1, January 2001.

[21] Berezdivin, R.; Breinig, R.; Topp, R. "Next-generation wireless communications concepts and technologies", IEEE Communications Magazine, March 2002.

[22] H. Wang, R. Katz and J. Giese, "Policy-Enabled Handoffs across Heterogeneous Wireless Networks", WMCSA 1999.

[23] Majlesi, A.; Khalaj, B.H.; "An adaptive fuzzy logic based handoff algorithm for hybrid networks", Signal Processing, 2002 6th International Conference Volume: 2 , Aug. 2002, pp. 1223-1228.

[24] E. Yanmaz, O. K. Tonguz; "Dynamic Load Balancing and Sharing Performance of Integrated Wireless Networks", IEEE Journal on Selected Areas in Communications, June 2004, vol. 22, pp. 862-871.

[25] Bing, B., Subramanian, R.; "An integrated multiple access technique with traffic load balancing for multimedia personal communication systems", 1997 IEEE 6th International Conference on Universal Personal Communications Record, Volume 2, Oct. 1997, pp. 756 - 760.

[26] N. Nasser, H. Hassanein; "Multi-class Bandwidth Allocation Policy for

3G Wireless Networks", Proc. 28th IEEE International Conference on Local Computer Networks (LCN '03).

[27] H. Hlavacs, G. Haring et al.; "Modelling Resource Management for Multi-class Traffic in Mobile Cellular Networks", Proc. 35th Hawaii International Conference on Local System Sciences 2002.

[28] Y. Guo, H. Chaskar; "Class-Based Quality of Service over Air Interfaces in 4G Mobile Networks", IEEE Communications Magazine, March 2002.

[29] P. Marbach; "Analysis of a Static Pricing Scheme for Priority Services", IEEE/ACM Transactions on Networking, April 2004, vol. 12, pp. 312-325.

[30] Ariel Orda and Nahum Shimkin; "Incentive Pricing in Multi-Class Communication Networks", Proceeding of the IEEE INFOCOM'97, Kobe, Japan, April 1997.

[31] Chi-chao Chao and Wai Chen; "Connection Admission Control for Mobile Multiple-Class Personal Communications Networks", IEEE Journal on Selected Areas in Communications, Vol. 15, No. 8, October 1997, pp. 1618 - 1626.

[32] Shun-Ping Chung and Jin-Chang Lee; "Performance Analysis and Overflowed Traffic Characterization in MultiService Hierarchical Wireless Net-

works", IEEE Transactions on Wireless Communications, Vol. 4, No. 3, May 2005, pp. 904 - 918.

[33] K. Mitchell and K. Sohraby; "An Analysis of the Effects of Mobility on Bandwidth Allocation Strategies in Multi-Class Cellular Wireless Networks", IEEE INFOCOM 2001.

[34] H. Chen, Q. Zeng, and D. P. Agrawal; "Evaluation of a New Adaptive Resource Management Scheme for Multi-Class Wireless and Mobile Networks", IEEE 58th Vehicular Technology Conference, VTC-2003, Vol. 4, 6-9 October 2003, pp. 2187 - 2191.

[35] S. K. Das, S. K. Sen, K. Basu, and H. Lin; "A Framework for Bandwidth Degradation and Call Admission Control Schemes for Multiclass Traffic in Next-Generation Wireless Networks", IEEE Journal on Selected Areas in Communications, Vol. 21, No. 10, December 2003, pp. 1790 - 1801.

[36] C. Chou and K. G. Chin; "Analysis of Adaptive Bandwidth Allocation in Wireless Networks with Multilevel Degradable Quality of Service", IEEE Transactions on Mobile Computing, Vol. 3, No. 1, January-March 2004.

[37] D. P. Gaver, P. A. Jacobs, and G. Latouche; "Finite Birth-and-Death Models in Randomly Changing Environments", Adv. Applied Prob., Vol. 16, 1984, pp. 715 - 731.