THESIS / THÈSE

MASTER IN COMPUTER SCIENCE

Solutions to fast multistream proportional fair scheduling in HSDPA systems

Bodin, Morel

Award date:
2009

Solutions to Fast Multistream Proportional Fair Scheduling in HSDPA Systems

Morel Bodin

Licentiate Thesis
Faculty of Computer Science
University of Namur
August 2009

Abstract

With ever-increasing data rates and lowering latencies, the percieved quality and fairness of modern cellular telephony networks is largely determined by their packet schedulers.

In this document we examine the HSDPA system to locate the packet scheduler and its parameters, examine its influence on the system, and compare solutions both current and of historical value.

Keywords: Scheduling, MIMO, HSDPA, Fairness

Résumé

Avec des débits en augmentation et les diminution de temps de latences, la qualité perçue et l'équité des réseaux téléphoniques cellulaires modernes est en grande partie déterminée par leurs ordonnanceurs de paquets.

Dans ce document, nous examinerons le système HSDPA pour identifier le lieu de l'ordonnanceur de paquets, ses paramètres, ainsi que son influence sur le système.

Nous comparerons un petit nombre d'ordonnanceurs, tants historiques que plus récents.

Mots clés: Scheduling, MIMO, HSDPA, Equité

Acknowledgments

I wish express my wholehearted thanks to Professor Jean Fichefet — It is to his hard work and dedication that I owe the very opportunity afforded to me by the MIHD program originated by him, here at the University of Namur ten years ago.

I am particularly indebted to my thesis supervisor, Professor Laurent Schumacher for his guidance, encouragements, and patience.

Finally, my deepest gratitude goes to my wife Sandra, whose constant sacrifices availed me of the time I needed to see my studies to their conclusion. Her steadfast support has been invaluable from start to finish.

Abbreviations

16QAM	quadrature amplitude modulation with 16 symbols
3GPP	3rd Generation Partnership Project
3G	3rd generation
4G	4th generation
64QAM	quadrature amplitude modulation with 64 symbols
AMC	adaptive modulation and coding
ARQ	automatic repeat request
ATBFQ	adaptive token bank fair queueing
BLER	block error rate
CLM1	closed loop transmit diversity mode 1
CN	core network
CPICH	common pilot channel
CQI	channel quality indication
CRC	cyclic redundancy check
DCH	dedicated channel
DSCH	downlink shared channel
ECN	explicit congestion notification
FACH	forward access channel

FEC	forward error correction
GA	genetic algorithm
GSM	global system for mobile communication
HARQ	hybrid automatic repeat request
HGPS	hybrid genetic packet scheduler
HSDPA	high-speed downlink packet access
HS-DSCH	high speed downlink shared channel
HS-PDSCH	high speed physical downlink shared channel
MAC	media access control
Max CIR	maximum carrier to interference ratio
MCS	modulation and coding scheme
MIMO	multiple input multiple output
MINLP	mixed-integer nonlinear programing
OSI	Open Systems Interconnection
PF	proportional fair
PHY	physical layer
QoS	quality of service
QPSK	quadrature phase shift keying
RLC	radio link control
RNC	radio network controller
RR	round robin
RX	receive
SB	score-based
SINR	signal to interference plus noise ratio
SINR	signal to interference ratio

SISO	single input single output
SMP_{sel}	spatial multiplexing limited to selection transmit diversity
SMP_x	spatial multiplexing with full weight selection freedom
TCP	transmission control protocol
TTI	transmit time interval
TX	transmit
UE	user equipment
UMTS	universal mobile telecommunications system
UTRAN	UMTS terrestrial radio access network
VoIP	voice over IP
WCDMA	wideband code division multiple access

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cellular mobile telephony has come to pervade our modern society within the last quarter century. What started as a voice-only analogue telecommunications system in the early 1980s has blossomed into a full spectrum of digital services. From the humble beginnings of SMS as the first true mobile digital application available to consumers, cellular technology has made huge progress. Today, high-speed downlink packet access (HSDPA) for universal mobile telecommunications system (UMTS) provides mobile broadband to more than 85 million subscribers[16] across hundreds of UMTS networks worldwide.

New developments in the field appear at a steady, sustained pace. The packet scheduler occupies a key position in the radio resource allocation strategy of cellular networks, from where it exerts a fundamental influence the behavior of the system as a whole. With the current generation of HSDPA, the packet scheduler has the opportunity to select a different set of UEs every 2ms transmit time interval (TTI).

Although the solution to proportional fair scheduling has been considered NP-hard (see [13, 27, 28, 31]), there exist a number of algorithms which provide useful results within the tight time budget. Consequently fast packet scheduling algorithms remain the focus of a great deal of research. In this paper, a variety of solutions to the problem are examined: both historical ones as well as more recent developments, with a particular emphasis on the techniques used to mitigate the complexity of the fundamental problem.

This paper is structured as follows: following this introductory chapter, is a discussion of modern HSDPA infrastructure, and how the different components interrelate to provide the reliable high performance service that it has become. Following that is a discussion of four schedulers often cited in the literature for a number of interesting properties they posses despite their apparent simplicity. The fourth chapter presents a more formal approach to the scheduling problem, where every significant parameter of the problem space

finds its place. The fifth chapter discusses the growing trend of cross-layer design in packet schedulers, and presents one such scheduler for illustrative purposes. Following that, meta-heuristic solutions are examined, with a particular accent on genetic algorithms (GAs). Finally, conclusions are drawn and possible avenues of further research are suggested.

Chapter 2

UTRAN

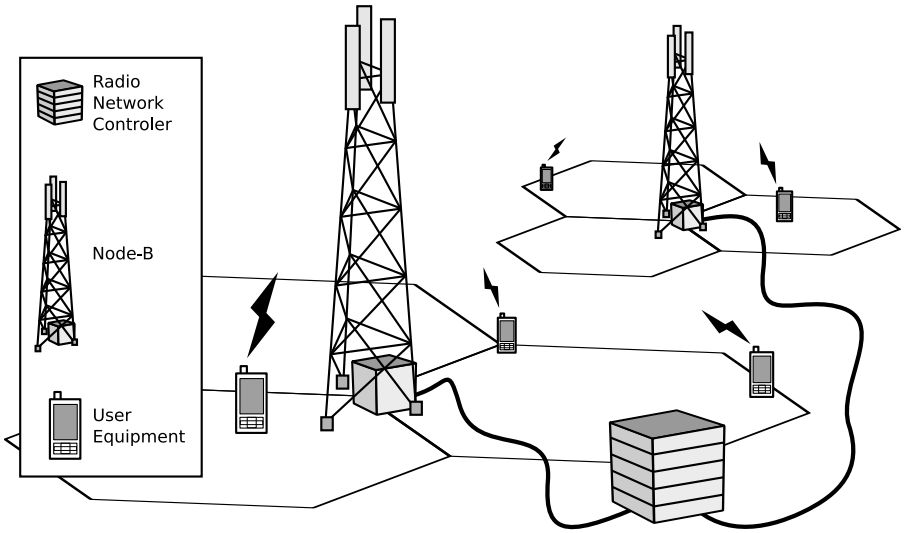

Figure 2.1: *UTRAN architecture*

2.1 Key elements and metrics

The UMTS terrestrial radio access network (UTRAN) (fig.2.1) can be schematized as consisting of three main components: radio network controllers (RNCs), Node Bs and UEs. Each RNC is further connected through a number of support nodes to the core network (CN) (not shown in fig.2.1). In the downlink direction packets destined for a UE transit from the core network to the serving RNC. The radio network controller then relays them either via wire or through a microwave link to the Node B of the cell in which the UE is currently located. From the Node B to the user equipment the packets are sent

over the air interface. Shown in fig. 2.1 are five UEs, two Node Bs, one RNC, and six individual cells.

2.1.1 Signal strength and CQI

An important measure of the quality of a radio signal from the perspective of a given actor is the signal to interference plus noise ratio (SINR). The lower the ratio, the more difficult it becomes to distinguish the signal. This can be expressed as the ratio between the power of the received signal and the sum of interference and noise. Holma et al. [19], on p. 124, give (2.1) for the SINR of HSDPA's high speed downlink shared channel (HS-DSCH).

$$SINR = SF_{16} \frac{P_{HS-DSCH}}{(1-\alpha) \cdot P_{own} + P_{other} + P_{noise}} \tag{2.1}$$

The presence of other transmitters in the vicinity of the receiver are sources of radio interference: from the perspective of a UE for example, these are typically its serving Node B (P_{own} in equation (2.1)), and the Node Bs of neighboring cells (P_{other}). With transmission frequencies ranging between 800MHz and 1900MHz the dominant source of noise is of thermal origin (P_{noise}), emanating from within the receiver itself.

The high speed physical downlink shared channel is specified with a fixed spreading factor (SF_{16}) of 16. The term α denotes the orthogonality factor: a value between zero and one described by Mehta et al. [30] as depending on "*the power delay profile of the multipath channel between the [UE] and its serving [Node B].*"

The power of the received signal, $P_{HS-DSCH}$ in equation (2.1) primarily depends on the transmission power, the distance separating the transmitter and the receiver and the degree of signal attenuation caused by the environment.

In urban environments, this attenuation is dominated by an effect referred to as Rayleigh fading (see fig. 2.2), and is caused by the existence of buildings blocking the line of sight separating the transmitter from the receiver: The geometry of these obstacles reflect and diffract the signal, the materials from which they are composed absorb or refract it. The greater the velocity of the UE, the more rapid the succession of peaks and troughs in received signal strength. In addition, a UE close to its serving Node B will experience less variations in its signal strength than will a UE further away [12].

Even in cases where there does exist a line of sight between the Node B and the UE, multipath interference can contribute to signal degradations in a situation referred to as Rician fading. This occurs when reflected radio waves

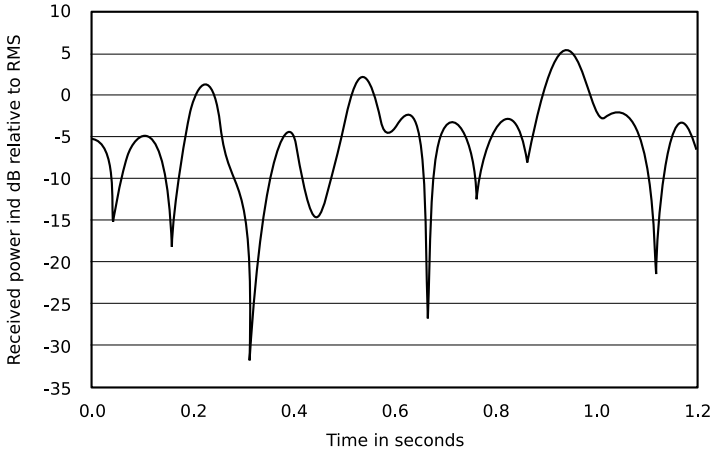

Figure 2.2: *Rayleigh fading as experienced by a UE at walking speed*

cancel the waves traveling along the direct route.

Among the innovations included in the specification of HSDPA was the requirement for UEs to periodically transmit channel quality indication (CQI) reports: estimates of the quality of the radio environment they currently experience, based on the signal to interference ratio (SINR) they observe with respect to the Node B's primary or secondary common pilot channel (CPICH). Owing to delays in reporting, inaccurate measurements, and data corruption during transmission, the CQI reports aren't always accurate. They nonetheless provide the invaluable feedback on which the vast majority of useful HSDPA schedulers base their decisions. (The notable exception to this is RR, discussed in 3.2.1.)

The CQI report is itself a synthetic value representing the index of a line in one of eleven tables defined by the 3rd Generation Partnership Project (3GPP) (see [5, pp. 52–61]). Each line in those tables fixes a number of transport parameters including transport block sizes, number of high speed physical downlink shared channel (HS-PDSCH) channels, modulation scheme, and reference power adjustment. The intent being to enable the UE to reliably and cheaply transmit to the Node B the highest order of transmission parameters which might still result in an acceptable error rate upon reception. Each UE category uses its own tables. Examples of UE categories and their corresponding maximum capabilities are listed in A.1.

2.1.2 Throughput, latency, and jitter

A number of fundamental properties characterizing computer networks also apply to cellular networks when these are used for the downlink transmission

of packet data. The metrics in question are primarily latency data rate. Variations in these values impact both the perceived fairness of the system and the quality of service. The choice of a scheduling strategy has a fundamental impact on these performance measures.

The throughput of the cellular system, as with any computer network is measured in bits per second. Services such as file transfer multimedia streaming generally benefit the most sensitive to high data rates. From the point of view of a UE accessing Internet services, the data rate bottleneck is often the air interface separating them from their serving Node B. In his paper, Bonald [12] underlines a notable exception to this:

> In practice, the data rate can indeed be limited by the wired network (e.g., the server) or by the mobile itself. Consider, for example, the widely used "stop-and-wait" error control protocol consisting in waiting for the acknowledgment of each packet before transmitting the next one. This introduces a minimum delay between the transmission of successive packets.

This Bonald's observation highlights what Shakkottai et al. [35] term an impedance mismatch between the wired and wireless world. They contend that TCP's explicit congestion notification (ECN) mechanism can help mitigate this problem affecting both network latency and throughput.

TCP matters aside, when examining the end-to-end transmission delays, it is the radio interface and the way it is managed, which sets HSDPA networks apart from wireline networks. That interface being inherently less reliable than copper or fiber, it is often the source of transmission errors requiring supplementary round trips between the UE and Node B. As will be seen in 2.3, there is often a trade off to be made between data rate and the degree of robustness of the signal in the face of transmission errors. This trade off inherently impacts latency as well.

Application	End-to-end latency
Scientific computing	1-10μs
Automated trading	100μs–1ms
Streaming media	100–250ms
voice over IP (VoIP)	<150ms
Online gaming	<200ms

Table 2.1: *Network latency: Acceptable upper bounds by application. (Source: [26, p. 2])*

Applications vary greatly when compared with respect to the demands they place on network latency (table 2.1). High performance scientific computing and high frequency automated trading place the most stringent requirements on network latencies, however they are at the moment quite unlikely to be encountered in use over cellular networks.

The combination of shortened TTI and partial transferal of responsibilities to the Node Bs has bestowed upon HSDPA a clear advantage over its predecessor in terms of network latency. Between Release '99 and Release '5, the estimated end-to-end delay has gone down from around 75ms to 35ms [19, fig. 7.30, p155]. Latencies experienced by a UE moving at 50kmh^{-1} can range between 70 and 120ms [34]. These values place streaming multimedia and VoIP within the reach of HSDPA, if strictly from the delay perspective.

In addition to throughput and network latency, the occurrence of network packets being dropped and large amounts of delay jitter can also prevent certain applications from functioning satisfactorily. In situations where the flow of packets is sufficiently great as to temporarily exceed the capacity of one of the links along the path. When this happens, the packets are queued just upstream of the bandwidth-constrained link, and transmitted when capacity becomes available. When this process occurs without ever completely filling the transmit queue, it induces a variable delay in the delivery of those packets which were buffered. Jitter is determined by calculating the standard deviation of measured network delay for representative traffic flows over a period of time. When the transmission queue in question overflows, packets are lost. Streaming media and online gaming are among the applications most sensitive to jitter.

2.1.3 Fairness

What can be considered fair is quite often subjective, and depends on the goals of the individual passing judgment. From the perspective of a cellular network subscriber, the criteria might be based on the dependability of the service subscribed to. A scheduling policy which leads to network starvation for UE in unfavorable environments would not be considered fair by that subscriber. The same scheduler may be considered fair by the network operator, particularly if subscribers pay according to traffic volumes: Why waste precious radio time by scheduling a UE whose volume will not maximize profits? A balance needs to be found between the two conflicting needs — in the scenario described, subscribers experiencing unreliable connectivity would quickly seek alternative service providers.

When evaluating the fairness of bandwidth allocation in computer net-

works, it is often Jain's fairness index (see [23]) which is used. The index maps quantitative bandwidth observations to an index in the range $[\frac{1}{n}..1]$, where n is the number of observations. An index of $\frac{1}{n}$ indicating that the system could not be less fair, an an index of 1 denoting perfect fairness. The equation provided by Jain et al. is:

$$f(x) = \frac{\left[\sum_{i=1}^{n} x_i\right]^2}{n \sum_{i-1}^{n} {x_i}^2} \quad \text{with} \quad x_i \geq 0, \tag{2.2}$$

in which x_i corresponds to the individual observation for user i. This equation can be used to compare the fairness of any allocation scheme. Network delay fairness can be calculated just as easily as throughput fairness by using observed delays for the values of x_i.

2.1.4 Quality of Service

The problem of fairness in computer networks is closely related to the perception of quality of service (QoS), and having an expression such as that given in (2.2) at our disposal does not necessarily remove all subjectivity from fairness measurements. The previous section the discussion centered around throughput fairness, however one could just as easily base the comparison on average delay instead. A situation where two subscribers experience equal throughput with unequal delays could be considered fair by a subscriber engaged in bulk file transfer, while a subscriber involved in a VoIP session would favor the opposite case. The discrepancies between the points of view of those subscribers will impact both the fairness and quality of service of the system as they experience it.

A distinction is to be made between QoS value judgments as in the case of the VoIP versus file transfer, transmission control protocol (TCP) traffic shaping with a view of enforcing QoS for non-realtime applications, and hard QoS involving service level guarantees as may be encountered in the medical or financial sectors. Throughout this paper, when mention is made regarding QoS, it will be in the sense of value judgments, unless otherwise noted. The added complexities imposed on fast fair scheduling by QoS provisions are outside of the scope of this document.

2.2 Origins of HSDPA

2.2.1 UMTS

The UMTS is currently the prevalent broadband mobile communications technology. It was codified and first standardized by the 3GPP in March 2000 [2], in a group of documents collectively referred to as Release '99. The 3GPP is a worldwide organization of some 370 actors from all sectors involved in cellular systems, including network operators, equipment manufacturers and standards organizations [21]. The partnership was was created in 1998 and has tasked itself with the maintenance and evolution of the global system for mobile communication (GSM) [1]. The then-new standard brought with it the dual promise of increased data rates and lowered network latency over the previous technologies.

Release '99 UMTS provides for three modes of transmission for downlink packet data. These are the dedicated channel (DCH), the downlink shared channel (DSCH), and the forward access channel (FACH). The DCH is used for transmissions which require low latency and relatively low bandwidth. The FACH is considerably less flexible in its usage than either DCH or DSCH [18, p. 308]. The bulk of downlink packet data is carried by the DSCH, which is specified to provide a data rate of 384 kilobits per second. In the uplink direction, Release '99 allows for a data rate of 64kbps [36, p. 239].

2.2.2 Release 5

In the decade since its first specification, UMTS has seen the introduction and refinement of HSDPA, further improving performance. Recognizing shortcomings of Release '99 DSCH with regards to data rates and network latencies, particularly compared to those same metrics in wired networks, the 3GPP introduced a notable update to UMTS in March 2002[2]. The newly introduced standard, designated Release 5, contained specifications for HSDPA [3].

To differentiate services specified in Release '99 from those in Release 5, the first are commonly referred to as 3rd generation (3G) while the latter are designated 3.5G. In their book, Smith and Collins [36], on p. 239 note that architecturally HSDPA is quite similar to 3G. Setting 3.5G apart is its greater flexibility in the allocation of radio resources, combined with an increase in data rates for both uplink and downlink packet data.

While the peak theoretical physical layer data rate afforded by HSDPA can reach 28 Mbps [20], corresponding theoretical network latencies in HSDPA can be reduced to less than 70 milliseconds [20, p. 397]. These performance in-

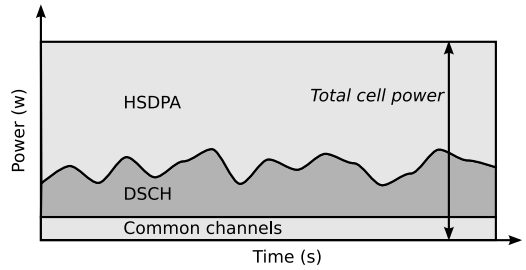

Figure 2.3: *Node B power allocation, showing portions of total power devolved to HSDPA, power-controlled dedicated channels, and common channel power budget reservation.*

dicators show that the current cellular mobile network is easily capable of supporting both low-latency protocols such as voice over IP as well as applications requiring relatively large channel capacities such as streaming video downloads. Although these figures represent peak data rates rather than average, they show a marked improvement, bringing wireless telecommunication a step closer to the performance seen in wired networks. As compared to the previous generation's data rates of hundreds of kilobits per second, and latencies twice as long as those typical of HSDPA, the current figures are a testament to the technological advances embodied in Release 5.

The gains in data rate and network latency observable in HSDPA systems systems can be attributed to the introduction of the HS-DSCH ushering in with it a shift in the distribution of responsibilities within the system. As its name implies, HS-DSCH is intended to supplement and improve upon Release '99 DSCH.

2.3 HS-DSCH

Whereas previously the majority of radio resource management functions related to packet data downlink were governed primarily by the RNCs, HS-DSCH is largely controlled by the Node Bs themselves: The innovations introduced at this level include link adaptation improvements and Node B based scheduling.

The HS-DSCH is a logical channel multiplexed over a number of HS-PDSCHs. Within a Node B's total transmit power budget in a shared carrier scenario, the HS-PDSCHs find their place in the surplus after allowance has been made for common signaling channels and the DSCH channels [32, p. 60]. The DSCH benefiting from fast power control, there frequently exists headroom within which HS-PDSCHs can be allocated (see fig. 2.3). In case of a dedicated chan-

nel, the HS-PDSCHs occupy the entire transmit power budget, save for the part devolved to common signaling channels. When the RNC has dedicated a portion of the total base station transmit power to HSDPA, the Node B can have the option of taking advantage of the power budget of DSCH when that channel is underutilized.

Key properties of HS-DSCH allow for considerably more opportunism in the allocation of resources in the face of varying channel conditions than does Release '99 DSCH. This opportunism translates directly to considerably better performance than UMTS Release '99 on average. Those properties are presented here.

2.3.1 Physical layer retransmission and HARQ

In early 3rd generation systems, transmission errors were detected through the use of cyclic redundancy check (CRC) data contained in each data packet. When a user equipment detects the occurrence of a transmission error, it signals this fact to the radio network controller. Although the actual method of making the RNC aware of the error depends on the specific protocol in place, the RNC invariably responds by sending the same packet once more to the UE. This means of error control, dubbed automatic repeat request (ARQ), introduces a high degree of latency: All retransmissions need to travel the relatively long path from the UE to the RNC, transiting through the Node B along the way. In addition to the latency issue, in situations where the channel is in an unfavorable state such repeated requests for data re-transmissions mobilize a significant portion of the available bandwidth.

In an effort to increase bandwidth efficiency and reduce latency in the presence of transmission errors, the designers of 3.5G systems have altered the error correction system in two important ways: Firstly, for as long as the UE remains within the Node B's cell and as long as the number of retransmissions remains sufficiently low, the responsibility for error correction and retransmission sits with the Node B rather than the RNC. This innovation by itself contributes greatly to reduce latency by moving the error detection and correction mechanism that much closer to the radio interface. Secondly, the relatively primitive error detection and correction of Release '99 was mechanism was upgraded to a system called hybrid automatic repeat request (HARQ).

HARQ is a probabilistic error detection and correction mechanism consisting of complementing the CRC bits within the data packets with a forward error correction (FEC) code. In HSDPA, turbo codes are used for this purpose. According to El Bahri et al. [15], turbo codes as they are used in HSDPA can allow the channel capacity to approach within 0.5 dB of the Shannon limit.

When the UE detects a transmission error, the erroneous packet is stored locally. The Node B progressively re-transmits those packets from its own HARQ buffer. This process continues until one of three events occurs:

- The UE has either received sufficient data to reconstruct the original packet.
- The UE is no longer within range of the Node B.
- The maximum number of re-tries is exceeded.

If the outcome of the error correction process does not result in the correct transmission of the data in question, responsibility in the matter is then transferred back from the Node B to the RNC. At that point, the RNC either instructs the same Node B to reschedule the packet, or it determines that the UE has moved out of the cell, and handover is arranged [19, pp. 36–39].

2.3.2 Adaptive modulation and coding

In Release '99 downlink packed data is transmitted over DSCH using a 4^{ary} quadrature phase shift keying (QPSK) (fig. 2.4). Although the modulation scheme is fixed, there exists a provision for fast power control and a variable spreading factor. The fast power control mechanism alters the downlink transmission power in lockstep with measured interference or fading, minimizing interference generated at times when channel conditions are good. For Release 5, this fast power control was abandoned in favor of other refinements, including adaptive modulation and coding (AMC) and HARQ.

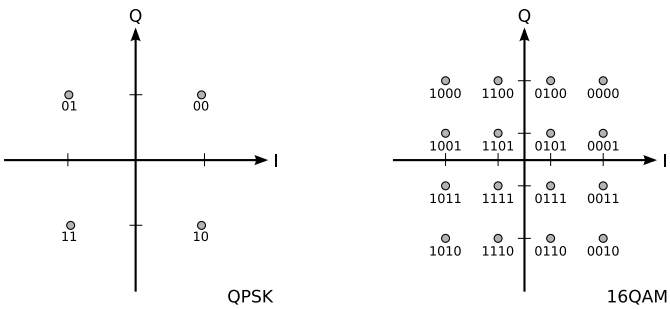

Figure 2.4: *Constellation diagrams with Gray coding for QPSK and 16QAM*

To make better use of the potential throughput offered by favorable channel conditions, HS-DSCH was specified to employ whichever of QPSK or

16QAM (fig. 2.4) channel conditions allow. Since the introduction of Release 7, quadrature amplitude modulation with 64 symbols (64QAM) is supported as well. Due to the short constellation point distance in 16QAM and 64QAM, they are far more complex to demodulate than QPSK. 16QAM doubles the attainable data rate over QPSK, while 64QAM triples it. When a Node B schedules a downlink transmission to a UE, it informs its choice of modulation scheme on the basis of the CQI sent to it by the UE in question. When the SINR is low, QPSK is used. When they channel conditions are good 16QAM can be used [19, 106–107]. Since the introduction of Release 7, and for as long as the UE has the capability, it follows that when channel conditions are such that even 16QAM under-utilizes the air interface, 64QAM can be selected. As the TTI remains fixed, switching between the three modulation schemes or varying the code rate means altering the transport block size, which in turn implies some variance in the system bit rate.

To further benefit from favorable conditions, the proportion of user data to error correction codes can be altered each TTI. As less errors occur when the channel is good, such variations in coding permit a much more efficient use of radio resources. The effective code rate can vary in this way between one quarter and three quarters, in increments of one quarter [18].

2.3.3 Multicode transmission

In the downlink direction, HSDPA makes use of the code-division properties of wideband code division multiple access (WCDMA), with a fixed spreading factor of 16. The codes are assigned to cells in one of two ways: either the RNC assigns a fixed set of codes to the Node B once and for all, or it slowly matches the set of codes assigned to the cell with the prevailing usage patterns within that cell [32, p. 57]. During each TTI the Node B can freely draw from its assigned pool of codes, allocating a certain number for each user equipment it has scheduled to transmit data to. In this way, the bandwidth allocated to each user equipment can be readily adapted as the situation warrants. Among the codes which have been allocated to the Node B by the RNC, one is always kept in reserve for shared signaling purposes. To each of the remaining codes corresponds a HS-PDSCHs. Theoretically, HSDPA downlink data rates to a single UE can reach 14 Mbps when the use of 16QAM is combined with the allocation of the maximum of 15 HS-PDSCHs, and an effective code rate approaches one [20, p449]. By comparison, the same parameters using 64QAM affords a theoretical data rate of 21.1 Mbps.

2.4 MIMO enhancements

In addition to the introduction of 64QAM, Release 7 saw the debut of multiple input multiple output (MIMO) for UMTS [4]. In that release, there is provision for the use of multiple antennas both at Node Bs and the by user equipments.

Depending on their individual hardware characteristics, HSDPA user equipments are assigned a category by their manufacturer, which they report to their serving Node B. Each category can support a specific maximum number of multicodes, a specific set of modulation schemes and a maximum coding rate. These characteristics place an upper bound on the data rate the UE is capable of sustaining in the download direction. In the appendices, table A.1 gives an overview of these categories. Therein, the theoretical maximum data rates for MIMO systems are shown to be double those of their direct SISO counterparts.

Scenarios with either multiple acsurx antennas, multiple TX antennas, or both multiple TX and RX antennas present a number of advantages not available in single input single output (SISO). Such configurations also entail a corresponding increase in system complexity. Berger [10] highlights two of the most useful modes MIMO operation: closed loop transmit diversity mode 1 (CLM1) and spatial multiplexing with full weight selection freedom (SMP_x).

2.4.1 Transmit diversity

CLM1 is an example of transmit diversity involving involving airing the same signal from all antennas, possibly with a small phase variation between the two. Total transmit power is divided between the active antennas. The intended result is that due to differences in the propagation paths taken by each signal component, they reach at least one of the receive antennas in phase with each other. The effect is a much better signal reception than might have been possible with only a single TX antenna. When more than one RX antenna

In CLM1, the receiving side can respond with received phase differences the in the feedback response. The transmitter can then use this response to adjust the phase differences between each antenna. Improving the phase correlation at the reception end and maximizing array gain.

Berger [10], in Table 2.4 cites a theoretical gain for 2×2 CLM1 of 4.66 dB over 1×1 SISO. Berger [10], on p. 39 notes that either round robin (RR) or proportional fair (PF) can be applied to this mode of operation which is, in essence, a simple beamforming technique.

2.4.2 Spatial multiplexing

At its very simplest, spatial multiplexing is what occurs with two neighboring cells, each containing a Node B equipped with a single omnidirectional antenna, each transmitting to a single UE. In this case, the Node Bs can be viewed as the multiple output side of the MIMO equation, while the two UEs represent the multiple input end. Clearly, in such a case the total bandwidth of the system as a whole can be twice what it would be if there were either only a single Node B or a single UE.

If there are t TX antennas, and the total number of RX antennas, all UEs combined is r, then $\mathrm{SMP}_{\mathrm{x}}$ consists of transmitting $\min(t, r)$ distinct data streams at once, each from a different TX antenna. The target RX antennas can all belong to the same UE, or they can belong to distinct UEs. The primary objective being achieve the highest throughput.

Berger [10], in Table 2.4 cites a theoretical gain for $2\times2\,\mathrm{SMP}_{\mathrm{sel}}$ of 4.39 dB over 1×1 SISO. Berger further notes that given the lack of correlation between the fading conditions experienced at both RX antennas of a UE, there is often more to be gained by systematically scheduling distinct UEs simultaneously.

Chapter 3

Packet Scheduling in HSDPA

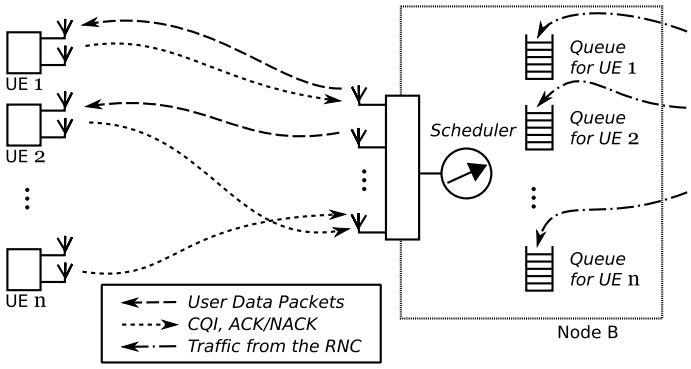

Figure 3.1: *Scheduler system Model*

3.1 Introduction

As was discussed in section 2.1.1, UEs experience widely fluctuating channel conditions between themselves their serving Node-B. Rayleigh fading was mentioned in 2.1.1 for its role as the primary mode of fluctuations in received signal strength at the UE in urban contexts.

The effects of AMC and HARQ are decreases in latency and increases in bandwidth. Although the HSDPA fast scheduler impacts those characteristics as well, its effects are on system throughput and quality of service are considerably broader than those of AMC and HARQ combined [8, 32]. At any given moment, there will exist a fixed number of user equipments within range of the Node-B. For each of those UEs, the Node-B maintains a queue containing data awaiting transmission. (See Figure 3.1 [7, 14, 35]). As time advances, each of these queues will vary independently in depth, as will the channel condi-

tions experienced by the corresponding UEs. Each TTI, the Node-B selects for transmission a quantity of data from a subset of the pending traffic queues.

As input upon which to base scheduling decisions, packet schedulers primarily use CQI values reported by UEs and the actual data transfer needs of each UE on the basis of the presence of queued data destined to them at the Node B (see fig.3.1). The scheduler will also take into account the number of actual HS-PDSCH codes at its disposal, as well as the proportion of the Node B's total transmit power budget which is available for use by HSDPA, as discussed in section 2.3.

3.2 Baseline Schedulers

To more fully understand the complexities of fast scheduling in HSDPA networks, we compare and contrast a small number of traditional solutions, with an eye towards their applicability in a MIMO context.

3.2.1 The RR Scheduler

If we were to use (2.2) to quantify the fairness of the round robin (RR) scheduler based on the air time it assigns to each UE, one would be hard pressed to find a scheduler with a higher index. Indeed, RR operates by simply transmitting download packets for each UE sequentially and equally, assigning the maximum number of codes and portion of available transmit power to that UE.

Computationally, it could hardly be simpler: the only input parameter taken into account by the RR scheduler is the presence of queued data. The problem lies in the fact that a great deal of bandwidth goes to waste in this manner. Due to this shortcoming, RR is unsuitable in most situations. Although it would certainly be possible to adapt the RR to multiple input multiple output applications, there really would be little point of pairing such an advanced transmission technology with such a wasteful scheduler.

3.2.2 The Max CIR Scheduler

The maximum carrier to interference ratio (Max CIR) scheduler (see [8, p. 43]) is for all intents and purposes the opposite of the RR scheduler: it transmits packets exclusively to UEs who report the most favorable channel conditions. In this manner, the scheduler makes very efficient use of the radio interface in terms of data rate, since those UEs with the best channel conditions can

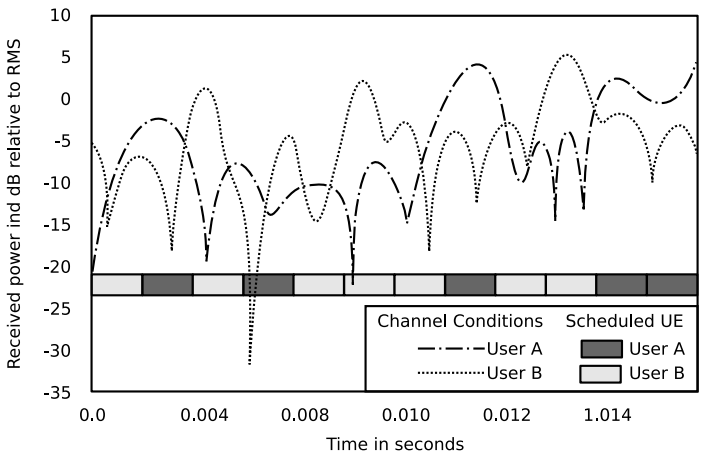

Figure 3.2: *Fading environments of two UEs moving at high speed, and the resulting Max CIR scheduler decisions*

support the highest data rates. The primary input parameters for Max CIR are transmit buffer occupancy and CQI. The Max CIR scheduler's behavior for two UEs in the presence of varying channel conditions is schematized in figure 3.2. Following the lead of both Bonald [12] and Berger [10], we can note the choice the scheduler makes of UE U from N_{queued} UEs, where UE u benefits from an immediate achievable data rate TP_u thus:

$$U = \underset{u \in [1,\ldots N_{queued}]}{\arg\max} \; TP_u. \tag{3.1}$$

Clearly, if the Max CIR scheduler only transmits to to those benefiting from a good channel, UEs with less favorable conditions will be left out. Indeed, a UE at the edge of a cell may never be scheduled at all. The advantages of the Max CIR scheduler are its extreme simplicity combined with its relatively high throughput. Its key failing is its patent lack of fairness in the worst case. In the ideal case, where the fading environment UE experienced by users averages out in the long run, Max CIR can be quite fair indeed. The problem, as Bonald [12] observes, is that:

> In practice, users do not experience the same fading. Fading is an extremely complex phenomenon caused by the interaction between the propagation environment and user mobility. While Rayleigh fading naturally arises from multipath reflections, the presence of a significant line-of-sight component results in Rician fading. The transmission data rate to SINR is also not linear, especially for high data rates, and depends on modulation and coding schemes.

In environments in which Bonald's observation bears out, the Max CIR scheduler can't possibly lead to any fairness, be it long or short term.

In a spatial multiplexing scenario, one would expect the Max CIR scheduler to perform better both in terms of throughput if not fairness, than the default Max CIR scheduler. The case for Max CIR in a transmit diversity situation is less clear.

3.2.3 The PF Scheduler

The weaknesses of the RR and Max CIR schedulers are to some degree addressed by the well-known proportional fair (PF) scheduler. As with the Max CIR scheduler, the primary input parameters for PF are transmit buffer occupancy and CQI, for which it keeps a number of historical observations for each UE. Its strategy consists of keeping an average of the experienced throughput of each UE u (noted $\overline{TP}_u$) over a fixed time window. Each TTI, the scheduled UE is the one whose ratio of immediate achievable data rate (noted TP_u) to average throughput is the highest. Out of N_{queued} UEs, Berger [10] expresses this choice of scheduled UE U simply as

$$U = \underset{u \in [1,\ldots N_{queued}]}{\arg\max} \left\{ \frac{TP_u}{\overline{TP}_u} \right\} . \tag{3.2}$$

Berger [10], on pp. 36–38, analyzes the throughput gain afforded by the PF scheduler in an idealized environment where no single UE has access on average to better channel conditions than any other, and all UEs move at the same rate. (Ensuring by that token that their Rayleigh fading will be statistically equivalent.) Further, he uses the assumption that the averaging window is sufficiently long such that the average throughputs don't change in time. Finally, he observes that under certain conditions

$$TP_u \propto SINR_u .$$

In this simplified environment, Berger's calculations place the mean gain of the PF with ten queued users over the case where there is only a single queued user at 4.67 dB (see [10, p. 38]). Given exactly the same environment, one could reasonably expect the **(!** ((**!**)cir) scheduler to perform just as well, both from a fairness and a throughput perspective.

From the standpoint of its actual fairness, Bonald [12] contends that PF suffers from the same long term problem stemming from the assumption that fading is equally experienced by all UEs, as was the case for Max CIR. From

this, he concludes that in real-world situations the PF scheduler favors UEs near to the Node-B, displaying similar sharing characteristics to Max CIR. The discrepancy in analysis between Bonald and Berger could stem from the fact that Bonald uses an "exponentially smoothed average" where Berger uses throughput values "averaged over a certain time window", taken to mean a cumulative moving average.

The PF scheduler has been adapted to use in MIMO by Lee et al. [27]. Prior to that, Berger [10], in equation 2.18 had provided the following generalization for the PF scheduling strategy to 2×2 $\mathrm{SMP_x}$:

$$U = \underset{\substack{u_1 \in [1,\ldots N_{queued}] \\ u_2 \in [1,\ldots N_{queued}] \\ m_1 \in [1,2] \\ m_2 \in [1,2] \wedge m_2 \neq m_1}}{\arg\max} \left\{ \frac{TP_{u_1,m_1}}{\overline{TP}_{u_1}}, \frac{TP_{u_1,m_1}|_{coint}}{\overline{TP}_{u_1}} + \frac{TP_{u_1,m_2}|_{coint}}{\overline{TP}_{u_2}} \right\}, \tag{3.3}$$

yielding by the same token the expression of a multistream proportional fair scheduler.

3.2.4 The SB Scheduler

When fading conditions are experienced equally among all UEs in a cell, the PF scheduler displays excellent fairness and throughput. Bonald [12] proposed a scheduling scheme which is designed to work around the location-dependent differences in channel condition variations. The score-based (SB) scheduler is specified to use the throughput statistics of UEs as the inputs for the Node-Bs scheduling decisions. These statistics can be the transmission rate or the SINR: Bonald's paper used the former while recent literature uses the latter (see Bokhari et al. [11]). A history of observations of the signal to interference plus noise ratio of each UE is kept, with a window size W. The score $s_i(t_k)$ for UE i at any given time t_k can be calculated using (3.4). In that expression, X_l denotes an independent and identically-distributed random binary value, and $r_i(t_k)$ is the rate experienced by UE i at t_k.

$$s_i(t_k) = 1 + \sum_{l=1}^{W-1} 1_{\{r_i(t_k) < r_i(t_{k-1})\}} + \sum_{l=1}^{W-1} 1_{\{r_i(t_k) = r_i(t_{k-1})\}} X_l \tag{3.4}$$

Bonald claims that given a sufficiently large W, the scores of any given UE will be uniformly distributed over the positive integers less than or equal to W. At each time slot t, it is the UE whose score is the lowest that is scheduled to receive data.

The dual promise of fairness in the face of unequal fading environments and very modest complexity are surely what has prompted the choice of SB as the "current baseline scheduler" for some instances of 4th generation (4G) wireless research [11, p. 1996], from which we can deduce that it is quite suitable indeed for MIMO applications.

Chapter 4

Mixed-integer nonlinear programming

When viewed as a class of scheduling algorithms, the group presented in the previous section are all very close in terms of relative complexity and immediate applicability. It is possible to define a complete mixed-integer nonlinear programing (MINLP) describing the objective of the scheduler and its inputs in minute detail. Such a MINLP would certainly be quite computationally expensive, as surmised by Liu and Leung [28], Bu et al. [13], Lee et al. [27], and Nguyen and Han [31], who deem the exact solution to proportional fair scheduling to be NP-hard. On the other hand, a MINLP would have the advantage of at least theoretically finding the very best possible solution, where heuristic methods can only make very good approximations.

4.1 Presentation

As was seen in 3.2.3, the PF scheduler in its default configuration only selects one single UE each TTI, allocating all available multicodes to that UE, up to the limit defined by its CQI feedback.

Kim and Hong [25] propose to improve upon the results afforded by the PF scheduler taking advantage of multiuser diversity, by distributing the available multicodes among several UEs each TTI, rather than favoring only a single one. This means that the scheduler now needs to select the best combination of UEs, based on their current ratio of feasible rate to average throughput.

The method proposed in their paper consists of formulating a mixed-integer nonlinear programing (MINLP) problem with a view of maximizing the sum of data rates assigned to all UEs. In their MINLP, they make use of a parameter ζ representing the target "fairness" factor of the system, which can take a real value in $[0..1)$. Values closer to 0 causes the MINLP to maximize throughput. Values for ζ taken from other extreme of the interval cause the system to favor

long-term fairness.

Kim and Hong compare their MINLP scheduler with a slightly modified PF having a similar ζ parameter. Values for ζ near 1 cause the thus-modified PF to behave as described in 3.2.3. Conversely, values near 0 give rise to behavior identical to that of the Max CIR scheduler.

Throughout the range of values taken by ζ Kim and Hong note that their MINLP scheduler achieves a throughput gain of between seven and ten percent over the modified PF when all fifteen available multicodes are allocated. Intuitively, the MINLP scheduler should reach three times the throughput of the modified PF scheduler when both schedulers are set to allocate only five multicodes per user. This follows since PF would only allocate those five codes to a single user, while MINLP has the leisure of allocating three different sets of five codes to three separate UEs simultaneously. Their system nearly achieves the threefold increase by performing more than two times better than the modified PF.

4.2 Formulation

The MINLP solution put forth by Kim and Hong made no mention of the availability of CQI feedback. Consequently, the SINR and power calculations featured prominently in the MINLP constraints. The formulation below is simplified compared to theirs by the replacement of those calculations with certain assumptions regarding the CQI values and their usefulness, at the cost of decreasing the accuracy of the MINLP as originally proposed.

$$T = \max_{n,m}\{\tau - \rho\} \tag{4.1a}$$

subject to

$$a_{ij} \in \{0,1\} \quad \forall i,j \tag{4.1b}$$

$$\sum_{j-1}^{J} a_{ij} = 1 \quad \forall i \tag{4.1c}$$

$$n_i \leq N_{i,\max} \quad \forall i \tag{4.1d}$$

$$\sum_{i-1}^{L} n_i \leq N_{\max} \tag{4.1e}$$

$$\sum_{i-1}^{L} P_i \leq P_{\max} \tag{4.1f}$$

$$\sum_{j-1}^{J} a_{ij} P_{i,j}(n_i) = P_i \quad \forall i \tag{4.1g}$$

where

$$\tau = \sum_{i=1}^{L}\sum_{j=1}^{J} \frac{a_{ij} n_i r_{ij}(t)}{\bar{r}_i(t)} \tag{4.1h}$$

$$\rho = \beta\left(\sum_{i=1}^{L} \frac{P_i}{P_{\max}} + \sum_{i=1}^{L} \frac{n_i}{N_{\max}}\right) \tag{4.1i}$$

$$r_{ij} = \frac{W}{g} R_c^{(j)} \log_2 M_j \tag{4.1j}$$

$$\sigma = \begin{cases} 1 & \text{when UE } i \text{ is served with } n_i r_{ij}(t) \\ 0 & \text{otherwise} \end{cases} \tag{4.1k}$$

$$\bar{r}_i = \begin{cases} \bar{r}_i(t+1) = (1-\zeta)\,\bar{r}_i(t) + \zeta\, n_i r_{ij}(t) & \text{when } \sigma = 1 \\ \bar{r}_i(t+1) = (1-\zeta)\,\bar{r}_i(t) & \text{otherwise} \end{cases} \tag{4.1l}$$

In the objective function (4.1a), τ is an expression of data rate based on a choice of UEs, and ratio of requested data rate with a set number multicodes to average data rate. As various combinations parameters in τ will yield the same data rate, ρ is a term serving to enforce power and multicode frugality such that the expression $\tau - \rho$ reaches a maximum when maximum throughput coincides with minimum power and multicodes for a given set of parameters.

In (4.1h), L denotes the number of UEs within the cell, while J the number of different modulation and coding schemes (MCSs) available to those UEs.

The term a_{ij} (4.1b) denotes the choice of transmission to UE i using MCS j, while (4.1c) insures that only one MCS is active at a time for any given user.

The number of multicodes allocated to UE i is noted n_i, while (4.1d) guarantees that the number of multicodes allocated to UE i does not exceed $N_{i,max}$, the number of multicodes UE i can handle at a time. The constraint (4.1e) insures that the total number of codes assigned does not exceed the number of codes available for use by the Node B (N_{max}).

The term $r_{ij}(t)$ denotes the achieved data rate at the time t, as given by (4.1j), with chip rate W, spreading factor g, code rate $R_c^{(j)}$ for MCS j, and number of points in the modulation scheme M_j.

The denominator $\bar{r}_i(t)$ in (4.1h) is the average of past achieved data rates for UE i, as calculated by (4.1k) and (4.1l). The ζ in (4.1l) is the value discussed in 4.1.

In (4.1a), ρ, as expressed (4.1i) P_i/P_{max} is the ratio of transmit power to total available power, in watts, when transmitting to UE i. Similarly, n_i/N_{max} is the ratio of multicodes allocated to multicodes available, when transmitting to UE i. The term β is a small constant scaling factor, described by Kim and Hong [25], p. 227 thus:

> ...ρ should have a minor effect on the value of T. Therefore, we use β as a small constant. To decide the value of β, we performed a simulation that shows the system throughput versus the value of β. The result notes that when $\beta < 1$, the system throughput converges on a maximum value. In our analysis, we set β as 10^{-3} ...

The value of $P_{i,j}(n_i)$ in constraint (4.1g) is the power in watts required to reach a sufficiently low block error rate (BLER) when a UE i transmits with n_i multicodes using MCS j, $P_{i,j}(n_i)$. In Kim and Hong's thorough analysis of the problem, the values of measured interference power imputable to neighboring base stations enters into consideration in the calculation of the value $P_{i,j}(n_i)$, along with representation of the orthogonality factor, power of thermal noise,

and multiple access interference. Rather than measure the individual factors and calculate $P_{i,j}(n_i)$ every TTI, the CQI feedback index, combined with the tables in [5], can provide a useful approximation.

4.3 Discussion

The solution put forward by Kim and Hong takes into account each important aspect of the proportional fair problem. On a purely theoretical basis, it is sound and complete. There is however very little hope that an implementation based on the resolution of the MINLP would complete in less than 2ms for any reasonable number of UEs. Indeed, in their paper Kim and Hong do not make mention of the running time of their simulations. Extending the formulation to include either transmit diversity or spatial multiplexing would only compound the problem.

The MINLP formulation serves as an excellent reference model describing the system in its entirety, and can certainly find use as a standard in offline quantification of more practical schedulers.

Chapter 5

Cross-Layer scheduling

5.1 The OSI model in HSDPA

For over a quarter century, the Open Systems Interconnection (OSI) layered model of computer networks [22], spanning the seven layers from the physical layer to the application, has served as a point of reference. Designers of network hardware, protocols, and applications use the model to guide their design decisions. It is a useful aid in reducing complexity by serving as a modularity guideline. The model serves as a conceptual basis helping to limit the scope of the responsibilities of the artifact being designed, encouraging reliance on products in other layers to perform their assigned duties, so that the system as a whole can perform meaningful tasks.

From the point of view of the classical Open Systems Interconnection Reference Model, The HSDPA architecture concerns itself primarily with the first two layers: The physical layer (PHY), and the data link layer. In the context of HSDPA, the data link layer is further subdivided into the media access control (MAC), directly above the PHY layer; and the radio link control (RLC) above that. The fast scheduling aspect of HSDPA finds its place mainly in the MAC layer [see 19, p. 24].

5.2 Imperfections of the model

It has become increasingly clear for several years now that it is quite difficult indeed to design a fast scheduler which fully realizes the benefits of MIMO systems without regard for both the PHY layer and the MAC layer. Ajib and Haccoun [7] observed:

> With the mixture of different traffic requirements and changing

> channel conditions, it becomes necessary to design the MAC layer to be adaptive to the traffic profile and channel characteristics.

As traffic profiles are most commonly associated with applications and network usage scenarios, their observation implies taking into consideration the layers above the OSI data link layer. From their initial observation, they therefore concluded that a cross-layer fast scheduler design is substantive in making efficient use of channel resources, particularly when QoS guarantees are to be reckoned with.

In a multiple input multiple output context, the need for a cross-layer scheduler seems all the more unavoidable. For example, the approach proposed in Aniba and Aïssa [9] also spans both the MAC layer and the PHY layer. The MIMO aspect of their work imposes a cross-layer approach from the outset, for reasons differing markedly from those put forth in [7].

> Under multiple antennas at the transmitter and [mobile station], the scheduler needs not only to select the set of users to transmit to, an issue pertaining to the MAC layer, but also the antenna(s) over which the data associated to each user would have to be transmitted, which is basically a PHY issue.

In the related discipline of wireless sensor networks, the shortcomings of the layered OSI model have been felt as well (see Mahalik [29], p. 87). Among the problems common to both HSDPA and wireless sensor networks is the observation that the wireline heritage of the OSI model imposed unreasonable constraints in the wireless domain. Notably, in wireline networks it was safe to assume that link capacity does not fluctuate. As we have seen in preceding sections, that is far from being the case in the HSDPA world. The strong influence congestion control and power and packet scheduling exert upon each other underlines the case for more comprehensive solutions.

There are also strong arguments against breaking with tradition and departing from the OSI model, chief of which is the decrease in system modularity. One of the principal benefits of modularity is the reduction of side effects when the modules are truly self contained. When that requirement is relaxed, it becomes more difficult to sort out functional dependencies within the system, leading to an increase in errors. Subsequent modifications to a system with highly relaxed layering can become quite difficult as a consequence.

5.3 ATBFQ

Bokhari et al. [11] introduce the adaptive token bank fair queueing (ATBFQ) scheduler, claiming comparable throughput to that of SB, at the same reducing queuing delays and dropped packets without favoring certain users over others. They further claim enhanced data rates for UEs at the edge of their cell.

The ATBFQ scheduler is a token bucket mechanism which functions in two phases. In the eligibility phase, the scheduler places each of the UEs which have queued data awaiting transmission in a list L, which is subsequently sorted in order of decreasing priority index calculated for each UE in L. The priority index P_i is given as the ratio of the token balance E_i of the UE to the token generation rate r_i. The token borrowing budget of the UE i at the head of L is then established based on the quantity of tokens it has borrowed from the bank, and the quantity it has contributed, closing the user selection phase.

The resource allocation phase begins: provided that the token balance of i does not exceed what is available at the bank, i may be allocated radio resources. The available radio resources for transmission to i are ranked, and the one (noted j) which would procure the best SINR is selected for i. The coding rate and modulation scheme for j is then configured according to what i can sustain, subsequently both the i's balance and the bank's balance, and allocated are adjusted accordingly. For as long as it has queued data and tokens to its credit, i can receive data. When either is no longer the case, it i is classified as non active, and the scheduler starts over again.

On performance grounds, Bokhari et al. compare their scheduler to SB, against which they claim better fairness. They based that statement on the distribution of users obtaining at least a given number of bytes per frame per sector. By that measure, the slope of the ATBFQ curve is steeper when scheduling for eight active users. Although the situation reverses itself when 20 users are considered, the authors expound in all cases, the number of dropped packets and the queuing delays are less with their algorithm than with SB, furthermore, they highlight the better performance observed with ATBFQ for edge UE.

The ATBFQ scheduler falls in the class of heuristics rather than exact solutions, and as such certainly has the run-time performance advantage over any direct implementation of a comparable MINLP.

Chapter 6

Meta-heuristic Scheduling

6.1 Genetic algorithms: A brief introduction

Described in 1975 by Holland [17] GAs, consist of a directed random search through solution space. The term algorithm is a slight misnomer in this context, as there generally exists no guarantee that such a process will systematically yield the true optimum solution. As such, these solutions belong to the class of heuristics.

GAs depend critically on the definition of virtual a genome representing the solution space, a recombination operator, a fitness function against which candidate solutions can be compared and the definition of a stop condition. The search is initialized with a 'parent' population. That population is often seeded at random, but can contain a number of genomes expressing possible solutions obtained by any conceivable means. At each iteration of the process the parent population is recombined in a manner reminiscent of biological reproduction, optionally including random mutations to produce a new population. Each individual in the new population is then assigned a score using the fitness function. The parent population for the next iteration is selected from those individuals who have scored the highest. The process continues in this way until a stop condition is met. When that happens, the most fit individual or individuals are presented as candidate solutions to the problem in question.

The genomes used for a given problem can take any number of forms, depending on what is appropriate for representing that problem's solution space. At the simplest, one can use an array of a fixed number of bits, although the use of trees and graphs has been encountered for this purpose. The recombination operator is usually defined to take as input two randomly selected individuals from the parent population, and produce as output a single offspring genome, inheriting material from either or both parents. Operators for random muta-

tion are sometimes included, as these can help prevent the solution from converging on local optimums. The stop condition condition is usually one the following events:

- An individual is found whose fitness score meets or exceeds a predefined threshold.
- A set number of iterations have been made through the process.
- The algorithm has run for a fixed amount of time determined in advance.

Each step in the described process usually needs to be empirically fine-tuned to promote the speed and efficiency of the algorithm. Population size and mutation probabilities are further parameters for adjusting the performance of the heuristic.

There are a number of interesting advantages to genetic approaches to numerical optimization problems, not least of which is the high degree of parallelism which can be incorporated into the solution. The general algorithm imposes no fixed limit on the number of concurrent populations evolving towards a solution to the same goal. Genetic algorithms are by this token in an excellent position to take advantage of recent trends in multicore computer hardware manufacturing.

A second advantage of note is the fact that the heuristic makes no fixed assumptions regarding the smoothness or granularity of the solution space, increasing their likelihood of finding global optimums rather than getting stuck with locally good solutions.

6.2 The HGPS scheduler

Abedi and Vadgama [6] utilize GA techniques to construct their hybrid genetic packet scheduler (HGPS) scheduler. For a Node-B having C_t channelization codes to distribute at time t, the genome they describe consists of an array sets $\eta = \{UE, Po, Oct, MCS\}$, where UE denotes the UE, Po the channelization code power allocation, Oct the number of bytes to be transmitted, and MCS the modulation and coding scheme to be applied to the transmission.

With n active users in the cell at time t, and m channelization codes available to the Node B, the solution genome is represented as $\{\eta_{1,t}, \eta_{2,t}, \ldots \eta_{m,t}\}$. The population is initialized by first running Max CIR, and placing the parameters it returns in $\eta_{1,t}$. It does likewise for RR and $\eta_{2,t}$. At the implementer's option, other packet schedulers can fill out the genome. The default behavior

fills the rest of the genome with random values. The inclusion of RR and Max CIR serve as limits, increasing tremendously the likelihood that the solution will be at least as fair as RR, and at least as bandwidth effective as Max CIR. The fitness function utilized is the weighted sum of five distinct metrics:

1. The number of octets delivered to the UE.
2. The ratio of the UE's queued octets to the number of octets received by the Node B to that point for that UE.
3. Packet delay as measured relative to the oldest waiting octet in the buffer.
4. Total expected throughput for all scheduled users.
5. And a fairness function, calculated as the inverse of the variance in throughputs.

The authors provide standard GA crossover functions, but make no mention of a stop condition. The authors compare the performance of their HGPS against a "fifo weighted Max C/I" which they neither define nor provide references to, making it very difficult indeed to extract any meaningful information from the comparisons they make and the conclusions they draw.

6.3 Conclusions regarding meta-heuristics

On such meagre data, no true conclusion can be drawn. The example given by Abedi and Vadgama [6] does provide us with an example of which fitness functions one could select, as well as possibilities regarding genome configuration.

It is worth noting that meta-heuristics in general, as well as GAs in particular, have certainly had occasions to show their merit elsewhere, so the time may yet come when a workable GA HSDPA scheduler sees the light.

Of recent note in the meta-heuristics family are particle swarm methods [24] which seem to carry some promise: They are presented as having the advantage of being less susceptible to myriad hidden or arbitrary parameters as is the case with GAs: To illustrate, in the HGPS solution, these were to be found in the crossover operators, the weights in the fitness function, the choice of fitness function, the number of iterations for which the algorithm is to run, and the mutation probabilities to name only those.

Chapter 7

Conclusions

With regards to the algorithms studied in this paper, it is clear that the architecture and decision model of the packet scheduler in an HSDPA system plays a key role in enhancing system throughput specifically in multistream contexts. The true difficulty lies in their objective comparison. The discrepancy in analysis observed for even the simpler schedulers such as was the case with PF between Bonald and Berger highlights the sensitivity of the schedulers to both the exact definitions for the mathematical primitives in use, as well as having the same starting assumptions before meaningful comparisons can be made. Adding to that difficulty, each group of researchers uses a different means of quantifying fairness, in spite of the existence of a widely published, easy to use fairness index.

These differences in methodology make direct quantitative comparisons between solutions all but impossible without re-testing them all in a single environment. To that end, there exist great opportunities for projects such as the FUNDP UTRAN Testbed which Peteghem [33] presents. Such projects have the potential to shed some quantitative light into these areas. Running real-time tests on such equipment as Vanpeteghem describes with the proposed schedulers could give a much clearer picture of how they truly compare on all fronts.

More and more stringent demands being placed on wireless networks. The large differences existing between their performance profiles and those of their wireline counterparts have already prompted a significant body of research in cross-layer architectures. These approaches could in the long run prove costly to maintain due to the introduction of side effects from one layer to the next. If we find ourselves continually breaking our own rules in order to obtain the performance we need, perhaps it is time for new rules: It may be interesting to fundamentally re-think the layered approach to networks, in favor of some other paradigm. In that vein, Mahalik [29], p. 94 suggests a possible avenue of

investigation:

> We advocate keeping some degree of modularity in the design of cross-layer solutions. This could be achieved by relying on functional entities - as opposed to layers in the classical design philosophy - that implement particular functions. This would also have the positive consequence of limiting the duplication of functions that often characterizes a layered design. This functional redundancy is, in fact, one of the causes for poor system performance.

From the standpoint of mathematical system models, the MINLP solution shines for its completeness with respect to the domain it was originally designed to describe. It would be very interesting to complete that model to include provisions for transmit diversity, spatial multiplexing, and beamforming. Having such a model would allow in depth analysis both from the theoretical standpoint and for serving as a quality benchmark in non-realtime numerical analysis of scheduling solutions.

Bibliography

[1] 3rd Generation Partnership Project. About 3GPP. http://www.3gpp.org/About-3GPP, . [Online; accessed 22-August-2009].

[2] 3rd Generation Partnership Project. 3GPP - Releases. http://www.3gpp.org/releases, . [Online; accessed 24-August-2009].

[3] 3rd Generation Partnership Project. Overview of 3GPP Release 5 summary of all Release 5 features. http://www.3gpp.org/ftp/Information/WORK_PLAN/Description_Releases/Rel5_features_v_2003_09_09.zip, . [Online; accessed 24-August-2009].

[4] 3rd Generation Partnership Project. Overview of 3GPP Release 7 v0.9.6 (2009-06). http://www.3gpp.org/ftp/Information/WORK_PLAN/Description_Releases/Rel-07_description_20090608.zip, . [Online; accessed 24-August-2009].

[5] 3rd Generation Partnership Project; Technical Specification Group Radio Access Network. Physical layer procedures (fdd) (release 8) (3GPP TS 25.214 version 8.6.0). http://www.3gpp.org/ftp/Specs/archive/25_series/25.214/25214-860.zip. [Online; accessed 24-August-2009].

[6] S. Abedi and S. Vadgama. A genetic approach for downlink packet scheduling in HSDPA system. *Soft Comput.*, 9(2):116–127, 2005. ISSN 1432-7643. doi: 10.1007/s00500-003-0353-4.

[7] Wessam Ajib and David Haccoun. An overview of scheduling algorithms in MIMO-based fourth-generation wireless systems. *IEEE network*, 19(5): 43–48, 2005.

[8] Bader Al-Manthari, Hossam Hassanein, and Nidal Nasser. Packet scheduling in 3.5G high-speed downlink packet access networks: Breadth

and depth. *IEEE Network*, 21(1):41–46, January 2007. doi: 10.1109/MNET.2007.314537.

[9] Ghassane Aniba and Sonia Aïssa. Adaptive scheduling for MIMO wireless networks: cross-layer approach and application to HSDPA. *IEEE Transactions on Wireless Communications*, 6(1):259–268, 2007.

[10] Lars Torsten Berger. *Performance of Multi-Antenna Enhanced HSDPA*. PhD thesis, Aalborg University, Aalborg, Denmark, May 2005.

[11] Feroz A. Bokhari, William K. Wong, and Halim Yanikomeroglu. Adaptive Token Bank Fair Queuing Scheduling Algorithm in the Downlink of 4G Wireless Multicarrier Networks. In *Vehicular Technology Conference, 2005. VTC Spring 2008. IEEE*, 2008.

[12] Thomas Bonald. A score-based opportunistic scheduler for fading radio channels. In *The Fifth European Wireless Conference*, Barcelona, Spain, February 2004.

[13] Tian Bu, Li Li, and Ramachandran Ramjee. Generalized proportional fair scheduling in third generation wireless data networks. In *INFOCOM*, 2006.

[14] Graciela Corral-Briones, Alexis A Dowhuszko, Jyri Hämäläinen, and Risto Wichman. Downlink multiuser scheduling algorithms with HSDPA closed-loop feedback information. In *2005 IEEE 61st Vehicular Technology Conference, 2005. VTC 2005-Spring*, volume 2, 2005.

[15] M. Wissem El Bahri, Fathi Raouefi, Hatem Boujemâ, and Mohamed Siala. Performance of HARQ I schemes using turbo codes. In *12th IEEE International Conference on Electronics, Circuits and Systems, 2005. ICECS 2005*, pages 1–4, 2005.

[16] UMTS Forum. Annual report 2008 - 2009. `http://www.umts-forum.org/component/option,com_docman/task,doc_download/gid,1901/Itemid,12/`. [Online; accessed 28-August-2009].

[17] John Henry Holland. *Adaptation in natural and artificial systems*. University of Michigan Press, 1975.

[18] Harri Holma, Antti Toskala, et al. *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*. John Wiley & Sons, Inc., New York, NY, United States, 3 edition, 2004.

[19] Harri Holma, Antti Toskala, et al. *HSDPA/HSUPA for UMTS: High speed radio access for mobile communications*. John Wiley & Sons, Ltd., Chichester, West Sussex, England, 2006.

[20] Harri Holma, Antti Toskala, et al. *WCDMA for UMTS: HSPA Evolution and LTE*. John Wiley and Sons Ltd, 4 edition, 2007.

[21] European Telecommunications Standards Institute. 3GPP Membership. `http://webapp.etsi.org/3gppmembership/Results.asp?Member=ALL_PARTNERS`. [Online; accessed 24-August-2009].

[22] ITU-T. Information technology - open systems interconnection - basic reference model. `http://www.itu.int/rec/T-REC-X.200-199407-I/en`. [Online; accessed 25-August-2009].

[23] Rajendra K Jain, Dah-Ming W Chiu, and William R Hawe. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. *DEC Research Report TR-301*, September 1984.

[24] James Kennedy and Russel Eberhart. Particle swarm optimization. In *IEEE International Conference on Neural Networks, 1995. Proceedings.*, volume 4, 1995.

[25] Sangbum Kim and Daehyoung Hong. Multi-user fair scheduling in the downlink of CDMA packet systems. *IEEE Communications and Letters*, 11(3), March 2007. doi: 10.1109/LCOMM.2007.061703.

[26] Ramana Rao Kompella, Kirill Levchenko, Alex C. Snoeren, and George Varghese. Every Microsecond Counts: Tracking Fine-Grain Latencies with a Lossy Difference Aggregator. In *Proceedings of the ACM SIGCOMM Conference SIGCOMM '09*, Barcelona, Spain, August 2009.

[27] Suk-Bok Lee, Ioannis Pefkianakis, Adam Meyerson, Shugong Xu, and Songwu Lu. Proportional fair frequency-domain packet scheduling for 3GPP LTE uplink. *IEEE INFOCOM 2009 mini-symposium*, 2009.

[28] Erwu Liu and Kin K Leung. Proportional fair scheduling: Analytical insight under rayleigh fading environment. In *IEEE Wireless Communications & Networking Conference*, 2008.

[29] Nataigour P. Mahalik, editor. *Sensor Networks and Configuration: Fundamentals, Standards, Platforms, and Applications*. Springer, Berlin, Germany, 2007.

[30] Neelesh B. Mehta, Andreas F. Molisch, and Larry J. Greenstein. Orthogonality factor in WCDMA downlinks in urban macrocellular environments. In *IEEE Global Telecommunications Conference, 2005*, volume 6, 2005.

[31] T.D. Nguyen and Y. Han. A proportional fairness algorithm with QoS provision in downlink OFDMA systems. *IEEE Communications Letters*, 10(11):760–762, 2006.

[32] Stefan Parkvall, Eva Englund, Peter Malm, Tomas Hedberg, Magnus Persson, and Janne Peisa. WCDMA evolved-high-speed packet-data services. *Ericsson Review*, 2:56–65, 2003.

[33] Huges Van Peteghem. *Building a Testbed Emulating Cellular Networks: Design, Implementation, Cross-Validation and Exploitation of a Real-Time Framework to Evaluate QoS and QoE in the UTRAN*. PhD thesis, Faculty of Computer Science, Namur, Belgium, March 2007.

[34] Haakon Riiser, Pål Halvorsen, Carsten Griwodz, and Bjørn Hestnes. Performance measurements and evaluation of video streaming in HSDPA networks with 16QAM modulation. In *2008 IEEE International Conference on Multimedia and Expo*, pages 489–492, 2008.

[35] Sanjay Shakkottai, Theodore S. Rappaport, and Peter C. Karlsson. Cross-layer design for wireless networks. *IEEE Communications magazine*, 41 (10):74–80, 2003.

[36] Clint Smith and Daniel Collins. *3G wireless networks*. McGraw-Hill Osborne Media, 2 edition, 2007.

[37] Wikipedia. High-speed downlink packet access. `http://en.wikipedia.org/w/index.php?title=High-Speed_Downlink_Packet_Access`, 2009. [Online; accessed 25-August-2009].

Appendix A

Tables and Graphs

UE Category	HS-DSCH codes	Highest order modulation	coding rate	MIMO Capability	data rate [Mbits/s]
1	5	16QAM	0.76		1.2
2	5	16QAM	0.76		1.2
3	5	16QAM	0.76		1.8
4	5	16QAM	0.76		1.8
5	5	16QAM	0.76		3.6
6	5	16QAM	0.76		3.6
7	10	16QAM	0.75		7.2
8	10	16QAM	0.76		7.2
9	15	16QAM	0.70		10.1
10	15	16QAM	0.97		14.0
11	5	QPSK	0.76		0.9
12	5	QPSK	0.76		1.8
13	15	64QAM	0.82		17.6
14	15	64QAM	0.98		21.1
15	15	16QAM	0.83	2×2	23.4
16	15	16QAM	*ca.* 1	2×2	28.0
20	15	64QAM	0.98		42.2

Table A.1: *Capabilities of HSDPA user equipment (UE) by category. Source: Wikipedia [37], except for the lines containing UE categories 15 and 16, which were found in Holma et al. [20], table 15.1, p.449.*

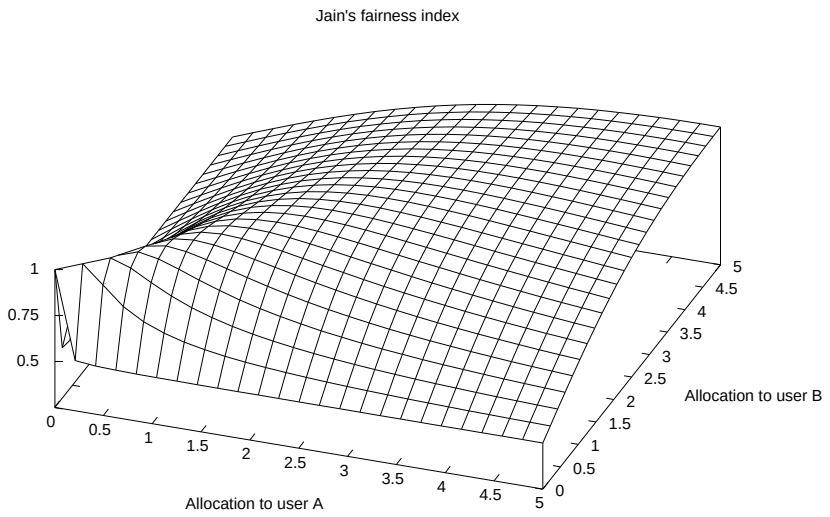

Figure A.1: *Values taken by Jain's fairness index when calculated for two users which are each allocated a value in the interval [0.005 .. 5]*