UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament de Teoria del Senyal
i Comunicacions

Radio and Computing Resource Management in SDR Clouds

Tesi presentada per obtenir el títol de:

Doctor per la Universitat Politécnica de Catalunya

Ismael Gómez-Miguelez
ismael.gomez@tsc.upc.edu

Programa Doctorat:
Teoría del Senyal i Communicacions

Director: Dr. Antoni Gelonch
Departament de Teoría del Senyal i Communicacions
Universitat Politécnica de Catalunya
Parc Mediterrani de la Tecnologia
Av. Canal Olímpic S/N
08860 Castelldefels, Barcelona
e-mail: antoni@tsc.upc.edu

Barcelona, 6 de Novembre de 2013

To Andrea and my family.

Abstract

The aim of this thesis is defining and developing the concept of an efficient management of radio and computing resources in an SDR cloud. The SDR cloud breaks with today's cellular architecture. A set of distributed antennas are connected by optical fibre to data processing centres. The radio and computing infrastructure can be shared between different operators (virtualization), reducing costs and risks, while increasing the capacity and creating new business models and opportunities.

The data centre centralizes the management of all system resources: antennas, spectrum, computing, routing, etc. Specially relevant is the computing resource management (CRM), whose objective is dynamically providing sufficient computing resources for a real-time execution of signal processing algorithms. Current CRM techniques are not designed for wireless applications. We demonstrate that this imposes a limit on the wireless traffic a CRM entity is capable to support. Based on this, a distributed management is proposed, where multiple CRM entities manage a cluster of processors, whose optimal size is derived from the traffic density.

Radio resource management techniques (RRM) also need to be adapted to the characteristics of the new SDR cloud architecture. We introduce a linear cost model to measure the cost associated to the infrastructure resources consumed according to the pay-per-use model. Based on this model, we formulate the efficiency maximization power allocation problem (EMPA). The operational costs per transmitted bit achieved by EMPA are 6 times lower than with traditional power allocation methods. Analytical solutions are obtained for the single channel case, with and without channel state information at the transmitter. It is shown that the optimal transmission rate is an increasing function of the product of the channel gain with the operational costs divided by the power costs.

The EMPA solution for multiple channels has the form of water-filling, present in many power allocation problems. In order to be able to obtain insights about how the optimal solution behaves as a function of the problem parameters, a novel technique based on ordered statistics has been developed. This technique allows solving general water-filling problems based on the channel statistics rather than their realization. This approach has allowed designing a low complexity EMPA algorithm (2 to 4 orders of magnitude faster than state-of-the-art algorithms).

Using the ordered statistics technique, we have shown that the optimal transmission rate behaviour with respect to the average channel gains and cost parameters is equivalent to the single channel case and that the efficiency increases with the number of available channels. The results can be applied to design more efficient SDR clouds. As an example, we have derived the optimal ratio of number of antennas per user that maximizes the efficiency. As new users enter and leave the network, this ratio should be kept constant, enabling and disabling antennas dynamically. This approach exploits the dynamism and elasticity provided by the SDR cloud.

In summary, this dissertation aims at influencing towards a change in the communications system management model (typically RRM), considering the introduction of the new infrastructure model (SDR cloud), new business models (based on Cloud Computing) and a more conciliatory view of an efficient resource management, not only focused on the optimization of the spectrum usage.

Resumen

El objetivo de esta tesis es definir y desarrollar el concepto de gestión eficiente de los recursos de radio y computación en un SDR cloud. El SDR cloud rompe con la estructura del sistema celular actual. Un conjunto de antenas distribuidas se conectan a centros de procesamiento mediante enlaces de comunicación de fibra óptica. La infraestructura de radio y procesamiento puede ser compartida entre distintos operadores (virtualización), disminuyendo costes y riesgos, aumentando la capacidad y abriendo nuevos modelos y oportunidades de negocio.

La centralización de la gestión del sistema viene soportada por el centro de procesamiento, donde se realiza una gestión de todos los recursos del sistema: antenas, espectro, computación, enrutado, etc. Resulta de especial relevancia la gestión de los recursos de computación (CRM) cuyo objetivo es el de proveer, dinámicamente, de suficientes recursos de computación para la ejecución en tiempo real de algoritmos de procesado del señal. Las técnicas actuales de CRM no han sido diseñadas para aplicaciones de comunicaciones. Demostramos que esta característica impone un límite en el tráfico que un gestor CRM puede soportar. En base a ello, proponemos una gestión distribuida donde múltiples entidades CRM gestionan grupos de procesadores, cuyo tamaño óptimo se deriva de la densidad de tráfico.

Las técnicas actuales de gestión de recursos radio (RRM) también deben ser adaptadas a las características de la nueva arquitectura SDR cloud. Introducimos un modelo de coste lineal que caracteriza los costes asociados al consumo de recursos de la infraestructura según el modelo de pago-por-uso. A partir de este modelo, formulamos el problema de asignación de potencia de máxima eficiencia (EMPA). Mediante una asignación EMPA, los costes de operación por bit transmitido son del orden de 6 veces menores que con los métodos tradicionales. Se han obtenido soluciones analíticas para el caso de un solo canal, con y sin información del canal disponible en el transmisor, y se ha demostrado que la velocidad óptima de transmisión es una función creciente del producto de la ganancia del canal por los costes operativos dividido entre los costes de potencia.

La solución EMPA para varios canales satisface el modelo "water-filling", presente en muchos tipos de optimización de potencia. Con el objetivo de conocer cómo ésta se comporta en función de los parámetros del sistema, se ha desarrollado una técnica nueva basada en estadísticas ordenadas. Esta técnica permite solucionar el problema del water-filling basándose en la estadística del canal en vez de en su realización. Este planteamiento, después de profundos análisis matemáticos, ha permitido desarrollar un algoritmo de asignación de potencia de baja complejidad (2 a 4 ordenes de magnitud más rápido que el estado del arte).

Mediante esta técnica, se ha demostrado que la velocidad óptima de transmisión se comporta de forma equivalente al caso de un solo canal y que la eficiencia incrementa a medida que aumentan el numero de canales disponibles. Estos resultados pueden aplicarse a diseñar un SDR cloud de forma más eficiente. A modo de ejemplo, hemos obtenido el ratio óptimo de número de antenas por usuario que maximiza la eficiencia. A medida que los usuarios entran y salen de la red, este ratio debe mantenerse constante, a fin de mantener una eficiencia lo más alta posible, activando o desactivando antenas dinámicamente. De esta forma se explota completamente el dinamismo ofrecido por una arquitectura elástica como el SDR cloud.

En definitiva, este trabajo pretende incidir en un cambio del modelo de gestión de un sistema de comunicaciones (típicamente RRM) habida cuenta de la introducción de una nueva infraestructura (SDR cloud), nuevos modelos de negocio (basados en Cloud Computing) y una visión más integradora de la gestión eficiente de los recursos del sistema, no solo centrada en la optimización del uso del espectro.

Contents

List of Figures

List of Tables

Acronyms

ALOE	Abstraction Layer and Operating Environment
API	application program interface
ASIC	application-specific integrated circuit
AWGN	additive white Gaussian noise
BS	base station
CAPEX	capitalization expenditures
c.d.f.	cumulative density function
CNR	channel-to-noise ratio
CSI	channel state information
CORBA	Common Object Request Broker Architecture
DSP	digital signal processor
DFT	discrete Fourier transform
CR	constant rate
CRP	constant rate and power
CRM	computing resource management
EMPA	Efficiency Maximization Power Allocation
EM	efficiency maximization
EEM	energy efficiency maximization
FDD	frequency-division duplexing
FDMA	frequency-division multiplexing access
FIFO	first-in first-out
FIR	finite impulse response
FPGA	field programmable gate array
GPP	general purpose processor
GSM	Global System for Mobile Communications

GPRS	Global Packet Radio System
IC	interference cancellation
IDFT	inverse discrete Fourier transform
i.i.d.	independent and identically-distributed
ISI	inter-symbol interference
LOS	line of sight
LTE	Long Term Evolution
LUT	look-up table
MAC	multiply-accumulate
MBPTS	megabits per time slot
MCP	multi-cell processing
MD-MIMO	massively-scalable distributed MIMO
MIMO	multiple-input multiple-output
MOPTS	millions of operations per time slot
MP-SoC	Multi-Processor System On-Chip
MRT	maximum-ratio transmission
MM	margin maximization
MMSE	minimum mean square error
OS	operating system
OFDM	orthogonal frequency-division multiplexing
OFDMA	orthogonal frequency-division multiplexing access
OPEX	operational expenditures
P-HAL	Platform-Hardware Abstraction Layer
PA	power amplifier
p.d.f.	probability density function
PE	processing element
QoS	quality of service
RAT	radio access technologies
RAN	radio access network
RRM	radio resource management
RRH	remote radio head

RF	radio frequency
RM	rate maximization
SCA	Software Communications Architecture
SDR	software-defined radio
SDMA	spatial-division multiplexing access
SINR	signal-to-interference-and-noise ratio
SNR	signal-to-noise ratio
SVD	singular-value decomposition
TDD	time-division duplexing
TDM	time-division multiplexing
TDMA	time-division multiplexing access
KKT	Karush-Kuhn-Tucker
UMTS	Universal Mobile Telecommunications System
WiMAX	Worldwide Interoperability for Microwave Access
ZF	zero-forcing

Chapter 1

Introduction

1.1 Modern Mobile Communications

During the last fifteen years, cellular networks have evolved from providers of ubiquitous coverage for voice-communication services to "anywhere-anytime on" serving access ports for high data rate Internet-based data services. A 3-orders of magnitude increase in the supported data rates has been achieved, from several kbps in 2G Global Packet Radio System (GPRS) up to tens of Mbps in the latest 4G Long Term Evolution (LTE) systems. However it is not enough at all, since the need for mobile data capacity is growing at an unprecedented extremely fast pace. Recent market studies, conducted by global organizations [1], wireless fora [2], telecom companies [3], and operators [4], have indicated that mobile data traffic is (at least) doubled every year. Projecting this rate at a decade, we get the so-called "1000x data challenge" or "capacity crunch" which should be efficiently dealt by future service providers. Based on shorter-term forecasts, the study of [3] predicts a 13x grow in mobile data for the 2012-2017 period (Fig. 1.1). This increase is justified by several trends, that is: the increase in the number of mobile devices, as by 2017 there will be 1.4 devices/holder; the increased penetration of machine-to-machine (M2M) devices, as billions of low data-rate devices with cellular connectivity are expected to be deployed and operate in the foreseen future; the increase in the usage of high-end portable devices like tablets and smartphones, as by 2017 each smartphone is expected to generate more than 2.7 GB per month (contrary to today's 350 MB/month figure), and the shift to data-hungry mobile video services, as currently half of the mobile traffic is video and in five years, it will have dominated the total load, possessing more or less the two thirds of it.

The cost to build, operate and upgrade today's radio access network (RAN) is becoming more and more expensive while the revenue is not growing at the same rate. In general, up to 80% capitalization expenditures (CAPEX) of a mobile operator is spent on the RAN [5]. The CAPEX is mainly spent at the stage of cell site constructions and consists of purchase and construction expenditures. Purchase expenditures include the purchases of BS and supplementary equipments, such as power and air conditioning equipments etc. Construction expenditures include network planning, site acquisition, civil works and so on. In consequence, more than half of CAPEX is not spent on productive wireless functionality.

State-of-the-art radio access solutions, such as the latest LTE releases, and their corresponding evolutionary paths (future LTE-Advanced releases) will not be able to fulfil the traffic demands while maintaining profits. From the radio point of view, interference is the major limitation inherent in such wireless networks and, therefore, a lot of research focuses on efficient interference mitigation techniques. multi-cell processing (MCP) and interference cancellation (IC) for dense multi-tier heterogeneous deployments, are regarded as key candi-

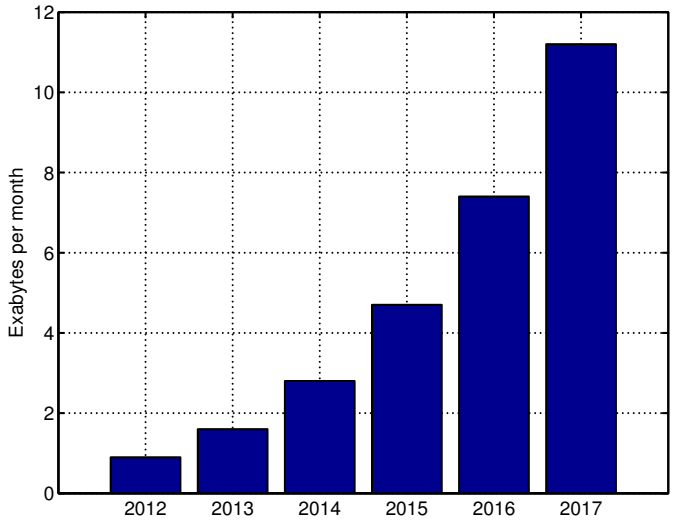

Figure 1.1: Expected growth of global mobile data traffic as predicted by Cisco VNI Mobile Forecast [3].

date technologies for next generation radio access. MCP [6] is based on the key concept of transforming all the interference into useful signal via cooperation to enhance system capacity, while IC [7] avoids interference among neighboring nodes through orthogonalization in various domains such as frequency, time, power and code.

The applicability of these technologies is limited by the restraints of the existing cellular-based architecture. The cellular architecture conceived for stand-alone working units with limited processing and inter-communication capabilities is intrinsically not suitable for coordinated and cooperative systems. This implies heavy protocols and imposes stringent constraints to be able to support cooperation, hence the actual benefits of the proposed technologies prove negligible compared to their theoretically predicted potential. The existing cellular-based architecture is the actual limiting factor not only for applying existing technologies, but also for inventing new ones that would potentially further improve system capacity. The cellular structure is susceptible to interference, inflexible, non-scalable and expensive.

1.2 Cell-less Network Architecture

We envisage a cell-less and massively-scalable distributed MIMO (MD-MIMO) network architecture, where the concept of massive multiple-input multiple-output (MIMO) system is extended to support cell-free network operation. In such a system, cell boundaries collapse and a very large number of low-cost infrastructure units, such as access points or radio-heads, are deployed practically everywhere. These units are connected through a backhaul network to a super centralized cloud-based infrastructure, where all communications processing is performed. We envision a radical shift to a user-centric cell-less architecture based on a cloud-empowered extremely dense network of low complexity infrastructure. Ideally, such a network will result in system performance levels independent of the number of served terminals. Enhanced by a universal per-user resources reuse, such an approach fulfills the vision of optimal capacity scaling. This leads to a fully scalable system, since an increase in the number of access terminals does not lead to a deterioration in the performance of active users and devices, contrary to conventional multiple-access cellular practices, as long as enough spatial resources are available. The fundamental ingredients of the envisioned system concept are:

- Number of infrastructure access nodes (such as access points or radio remote units) at least of the same order of magnitude as the number of users and connected devices;
- Global and centralized data communications processing;
- User-centric (instead of cellular-centric) system design and optimization thanks to the largely redundant spatial degrees of freedom, which allow universal resources reuse per user, independent of the number of access terminals;
- Flat system architecture, assuming no hierarchical levels and entities.

Throughout this thesis, we will assume the concept of cell-less, super-centralized and computationally powerful system, where all the communications processing as well as the computing and communications resources management operations reside in the cloud. Such a concept revolutionizes current thinking and practices of wireless network design and operation in many ways:

- System planning/retuning phases are no longer needed in their current form, since antenna units may be deployed everywhere without the necessity of planning;
- Cell boundaries disappear and each terminal is not attached to a single-cell but is rather served by a dynamically changing set of distributed transmission points;
- Bandwidth is not split among cells, but is fully available to every user at each time instant, catering for fully scalable network performance;
- Efficient multi-user access to system resources is realized through massively-scalable virtual MIMO multiplexing and resources optimization techniques. Thus system-wise instead of cellular/cluster-wise allocation decisions are taken on a small time-scale (ideally every transmission frame), exploiting global system knowledge;
- Communications processing, radio resources management and computing resources management procedures are co-located in the cloud and are thus jointly designed and optimized;
- The traditional spectral versus energy efficiency trade-off barrier is broken thanks to the massive spatial degrees of freedom, allowing both factors to increase simultaneously;
- Radio transceiver and access mechanisms are designed and optimized taking into account their processing cost, allowing for the introduction and characterization of computing resources-limited system capacity regions (similarly to existing backhaul-limited capacity approaches)
- The whole pool of available resources, spectrum, energy, antennas, backhaul network, computation, etc., is fully shared among the different users in the system providing a wide resource sharing and management scenario and introducing the base for new business models in the wireless arena.
- Real-time management of a large-scale distributed computing system: centralized control of distributed management units OR centralized management supported by distributed low-level managers.
- Efficient framework for dynamic deployment and management of waveforms.

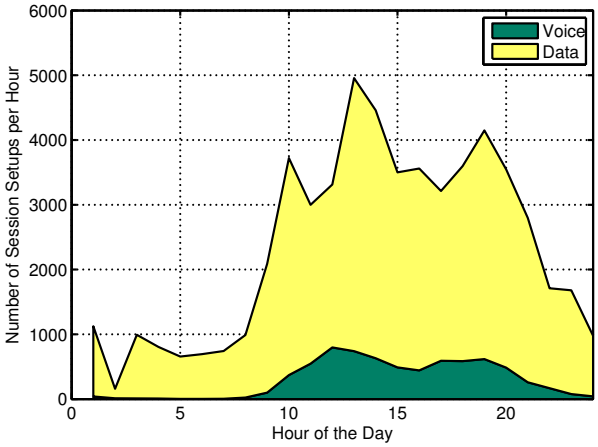

Figure 1.2: Typical weekday traffic pattern within an antenna sector in a city center.

1.3 The SDR Cloud

A MIMO antenna system increases the radio transmission performance by exploiting the spatial diversity where fading affects each one of the multiple communication links differently. Distributed antenna systems, moreover, employ optical fiber or coaxial cable for connecting the antenna system to the remote processing unit. A cooperative distributed antenna system based on a radio over fiber technique was introduced in China's beyond 3G FuTURE project and tested in field experiments. In those experiments, multiple antennas were distributed over the cell and connected to a data centre performing all signal processing tasks [8].

A base station (BS) is the wireless access points of cellular communications systems. It comprises antennas and analog and digital signal processing resources for implementing radio transmitters and receivers. Today's BSs feature a set of heterogeneous processing devices, including application-specific integrated circuits (ASICs), general purpose processors (GPPs), digital signal processors (DSPs), and field programmable gate arrays (FPGAs). Each device executes the tasks that were specified at design time. The network operator deploys base station resources as a function of the expected peak load. The goal is guaranteeing a certain quality figure, for example, the probability of granting a user service request. Providing resources for the worst case scenario however leads to long idle times and resource inefficiencies because of the sporadic use of wireless communications services [9]. Deploying fewer resources would increase the mean resource utilization while increasing the user rejection probability. Base stations may be shared between radio operators, but temporarily unused resources can still hardly be reassigned for other purposes. Radio operators thus purchase, maintain, and update considerably more resources than needed for most of the time (Fig. 1.2).

The software-defined radio (SDR) concept arose for extending the digital-signal processing parts of radio transmitters and receivers to run as software (SDR application) on general-purpose hardware (SDR platform). The ongoing advances in radio engineering and digital signal processing suggest employing processor arrays and automatic resource allocation tools that provide computing resources on demand. The approach of considering a data centre as the computing core of a BS is then a natural evolution of wireless communications.

A SDR cloud comprises a set of distributed antenna sites that connect to one or several data centers through low-latency and high-bandwidth communication links [10]. The antenna sites process the radio frequency (RF) signals and convert signals from analog to digital and vice versa. The digital data is processed entirely in the data centre, employing SDR and cloud

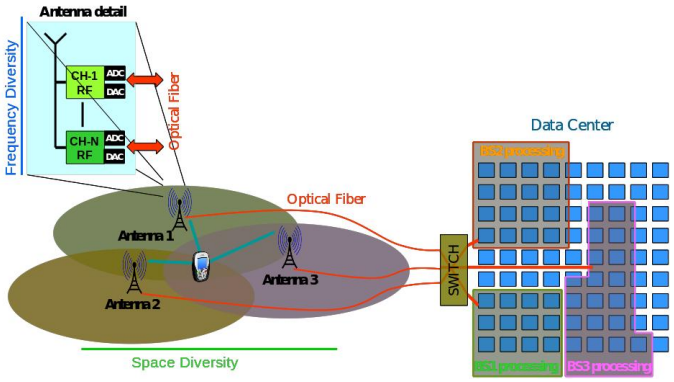

Figure 1.3: The SDR cloud. (CH-1 RF represents the analog signal processing part of radio frequency channel 1.).

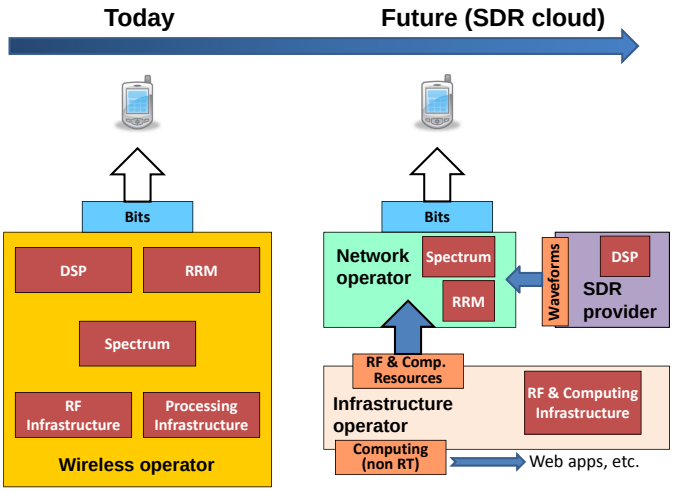

Figure 1.4: Envisaged evolution of the wireless technology market players, enabled by the introduction of SDR clouds.

computing technology. The computing infrastructure is based on commodity servers, which have shown to provide enough computational capacity to support 4G-like standards [11]. The use of SDR technology and commodity infrastructure allows leasing temporary unused resources for other purposes, not necessary related to wireless services (e.g. web applications). Consequently, it is economically sustainable to provide computational resources for the worst case scenario, because unused resources provide income too.

We envisage an infrastructure based on distributed antennas with a limited amount of local processing resources. These antennas connect to a data centre via low latency and high-speed communication links (Fig. 1.3). The SDR cloud naturally enables implementing the MD-MIMO cell-less network architecture.

The SDR cloud will change the role of today's wireless business players. Figure 1.4 illustrates the envisaged evolution of the current vertically integrated market towards an horizontal market with three different entities. We have represented the most important components in the figure:

- *DSP* is the set of signal processing algorithms required for the transmission and reception of the information;
- *Radio resource management (RRM)* is set of policies or algorithms adopted for an efficient management of the radio resources;
- *Spectrum* is the set of frequencies where a license allows a carrier to operate;
- *RF Infrastructure* is the set of physical elements or devices required for the transmission and reception of RF signals, e.g: antennas, power amplifiers, RF filters, construction sites, real estate, air conditioning, etc.
- *Computing Infrastructure* is the set of physical elements or devices required for the processing of RF signals, e.g.: processors, memories, hard disk, data centre security services, software licenses, air conditioning, etc.

On the other hand, we can identify the following players and their relations:

- Wireless Operator: Is the operator or carrier, owner of the entire system: the DSP and RRM algorithms, the spectrum, the RF infrastructure and the processing infrastructure. Here, the term *processing* is used to emphasize the fact that the signal processing is performed on dedicated hardware.
- Infrastructure Operator: Is the owner of the RF and computing infrastructure. Here, the term *computing* instead of processing is used, because signal processing is executed on general purpose computing hardware.
- Network Operator: Owns the rights to use the spectrum and is responsible of the definition of RRM algorithms.
- SDR Provider: Develops SDR applications, that is, waveforms for different radio access technologies (RAT).

A Cloud is a type of parallel and distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned as a function of the service-level agreements between service providers and consumers. Clouds provide services to clients without reference to the infrastructure that hosts the services. As cloud clients, network operators would be able to offer wireless communications services on demand and anywhere in the world without local installation. Network operators may therefore share radio-related infrastructure and focus on developing wireless communications standards and services. An infrastructure operator providing computing resources to network operators facilitates a more efficient and scalable computing resource deployment and management by centralizing storage, memory, processing, and bandwidth resources.

The SDR cloud provides essentially the same benefits as a general purpose cloud. It inherits the resource-as-a-service and pay-per-use business concepts: computing power (Infrastructure as a Service–IaaS), system software (Platform as a Service–PaaS), and applications (Software as a Service–SaaS) will be provided on demand and without knowledge of the physical location and types of CPUs, discs, software repositories, and so forth.

A single data centre is shared between several radio operators and thousands of end users. (Some 100,000 user sessions may be active at the same time in a city of one million or more inhabitants.) Virtualization is employed

- for ensuring secure and fair resource sharing, where one radio operator–the SDR cloud *client*–is not aware of others using the same physical machines;

- for consolidating several virtual resources in a single physical resource, reducing energy consumption;
- for dynamic provisioning of resources without the need of hardware purchase;
- for security, reliability and scalability purposes.

In other words, virtualization transforms physical resources to virtual resources (Fig. 1.5). Different business models or agreements are possible. A minimum set of resources may be guaranteed to each radio operator, for instance. The remaining or unused resources can be shared–fairly or competitively–as a function of the market, environment, or policy, among others. This requires a flexible, though efficient, computing resource management framework as a basis for the SDR cloud business. Such framework, in other words, plays an essential role in the deployment and operation of SDR clouds. It, particularly, needs to ensure real-time resource allocation and execution in dynamic environments with different resource and service constraints.

Computing resources virtualization allows the SDR cloud to isolate the physical infrastructure from its functionality, enabling each market player to concentrate on an isolated part of the business, increasing the specialization and reducing business risks. Nevertheless, it is also possible that a single corporation plays more than one of the roles identified in Fig. 1.4. For instance, the network operator could also be interested in developing the DSP algorithms, tailored to their specific radio resource management (RRM) policies; or the infrastructure operator may provide the DSP algorithms tailored to their specific computing infrastructure (SaaS). Furthermore, it is also possible that a single operator centralizes the baseband processing and applies virtualization techniques without resource sharing functionalities. This is usually called a *private cloud* and can be seen as a subclass of the SDR cloud. Private clouds are already being deployed as part of industry R&D projects or commercial products:

- China Mobile's Cloud RAN (C-RAN) [5] has already been deployed in a few cities of China. Their technology uses IBM commodity servers with Intel Core i7 general purpose processors. The project considers virtualization for consolidation and scalability purposes.
- Alcatel-Lucent's LightRadio [12] is an all-in-one low cost and low energy solution for deploying small cells or remote radio head (RRH) units. The project does not include the baseband processing side on the data centre.
- Nokia-Siemens Networks Liquid Radio [13] centralizes the digital signal processing of several RRH units associated to cells or pico cells. The use of SDR or GPP is not specified hence the flexibility is limited.

1.4 Resource Efficiency in Radio Resource Management

RRM is defined as the set of techniques (strategies or algorithms) that a wireless operator applies to the control plane of the RAN, for controlling system parameters such as transmit power, channel allocation, data rates, handover criteria, modulation scheme, error coding scheme, etc. The objective is typically to utilize the limited radio spectrum resources and radio network infrastructure as efficiently as possible.

Similarly to how virtualization transforms a single physical resource into several virtual resources, there are signal processing techniques that are able to transform physical radio resources (space, time and frequency) into a set of orthogonal virtual resources, for instance orthogonal frequency-division multiplexing access (OFDMA), spatial-division multiplexing

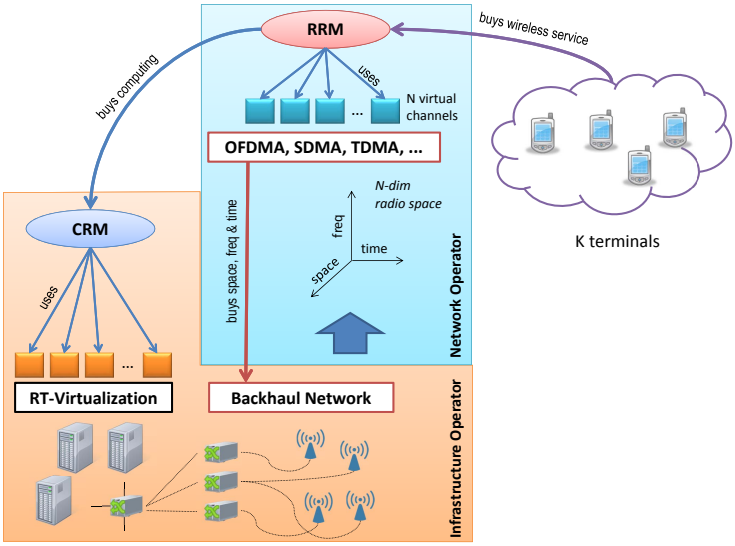

Figure 1.5: Virtualization of radio and infrastructure resources in the SDR cloud.

access (SDMA), time-division multiplexing access (TDMA) and so forth. In an SDR cloud, the network operator buys space, frequency and time utilization to the infrastructure operator. For instance: 10 seconds of signal transmission using 4 antennas and 1 RF channel on each antenna. The network operator moves CAPEX to OPEX, reducing risk and allowing easy scalability and resource provisioning. Physical radio resources are transformed to virtual radio resources (i.e. orthogonal streams) which are then allocated to users (Fig. 1.5).

The fact that RF and computing resources are shared implies that using them has a cost, determined by the competition between different network operators and the market laws. This cost is different than the operational costs of using today's infrastructure, which is acquired once and then amortized during the resource lifetime. Using resources has a cost also if they are not shared but can be consolidated through virtualization (e.g. private clouds), because power consumption is proportional to the utilization.

Today's RRM policies or decisions do not consider resource utilization costs or *pay-per-use* models. Typically, the aim of RRM is to maximize the user experience based on a set of system constraints. System constraints can be, for instance, maximum transmission power, maximum channel bandwidth, number of antennas, maximum processing power and so forth. These constraints may not exist, or be more relaxed in an SDR cloud system, because physical resources can by dynamically provided.

Based on these premises, we argue that today's RRM policies or algorithms, designed for cellular systems, are not valid for tomorrow's SDR cloud virtualized infrastructure. The aim of tomorrow's RRM algorithms should be to maximize how efficiently the virtual resources are employed. Hence, a network operator that applies resource-efficient RRM policies maximizes the service delivered to users while minimizing the costs or the price paid to the infrastructure operator for virtual resources utilization [14, 15].

1.5 Contribution and Outline

This thesis contributes to solve the computing resource management (CRM) and RRM challenges presented in SDR clouds. We study how wireless restraints influence the CRM and

how the SDR cloud architecture alters the RRM problem.

Chapter 2 introduces the technologies that enable the implementation of a cell-less network architecture where all the baseband processing is done in software, in a centralized infrastructure and with resource sharing capabilities. These technologies are SDR, Cloud computing and distributed real-time SDR frameworks. In the topic of real-time SDR frameworks, we will describe the open-source ALOE framework, developed and maintained by this thesis author.

Real-time virtualization is enabled by an entity that manages the computing resources to be assigned to each user. The CRM problem for SDR clouds is introduced in Chapter 3. CRM aims at dynamically allocating a sufficient amount of computing resources required for the processing of each user's signals. The fact that computing resources are used for wireless applications impose two conditions: the processing and allocation time must satisfy a deadline constraint (real-time processing and allocation). Due to the fact that real-time scheduling algorithms have polynomial complexity, as a function of the number of processors they manage, the two conditions impose a limit on the maximum number of processors a single resource allocator can manage. In turn, this imposes a limit on its capacity, in terms of maximum number of users. This chapter identifies this fundamental relation and derives the capacity of a resource allocator, which is then used to introduce a hierarchical CRM architecture for large SDR clouds, where thousands of users must be served in real-time.

The thesis moves from the computing world to the radio world through Chapters 4 and 5. Chapter 4 describes how physical radio resources can be shared between different users, creating virtual radio resources. The general parallel channel model is introduced, to cover many different virtualization techniques (TDMA, OFDMA and SDMA). This unique model allows us to easily characterize the quantity of used radio resources, as defined in Chapter 5. A simple linear resource cost model is proposed to characterize the OPEX or the cost the network operator pays the infrastructure operator for using their resources.

Chapters 4 and 5 are the basis for the formulation of RRM algorithms for SDR clouds. Chapter 6 studies the Efficiency Maximization Power Allocation (EMPA) problem. The problem considers the parallel channel model and resource cost model introduced in Chapters 4 and 5. The optimal power allocation is derived in closed-form for the single-channel case with and without retransmissions due to outage or channel errors. For the multi-channel case, an iterative optimal solution is proposed.

The complexity of this solution scales linearly with the number of channels and hence is not efficient for application in SDR clouds where a large number of channels may be available. To overcome this limitation, a novel sub-optimal solution is introduced in Chapter 7, with complexity independent of the number of channels. This solution uses ordered statistics to compute a set of look-up tables based on the channel statistics to then solve the power allocation problem. Furthermore, the new solution allows deriving important insights on the relation between the system parameters and optimal solution.

Concluding remarks, summary of results and future work are outlined in Chapter 8.

The main contributions of this thesis are:

1. Fundamental analysis and capacity derivation of the traffic a computing resource allocator can manage in an SDR cloud;

2. Derivation of resource-efficient power allocation on single channels with and without channel state information (CSI) and in parallel channels with CSI;

3. Application of ordered statistics to solve power allocation problems based on channel statistics, which allow

- to obtain the power allocation solution with 2-3 orders of magnitude less complexity than state-of-the-art solutions, and
- to study the performance and properties of the solution as a function of the problem parameters.

Chapter 2

Enabling Technologies

2.1 Introduction

In this chapter three technology enablers for the SDR cloud concept are reviewed. Computational aspects, instead of information theoretic aspects, are covered, because cooperative wireless processing techniques (e.g. MCP [6] and IC [7]), while improving the system capacity, are not strictly required for implementing an SDR cloud. On the other hand, virtualization technology and Cloud computing (Section 2.2), Software-Defined Radios (Section 2.3) and a framework for distributed SDR processing (Section 2.4) are the essential components of an SDR cloud.

2.2 Cloud Computing

There are many definitions of *Cloud Computing*. It is often suggested as another way to define what is known as *Software as a Service* (SaaS), very extended thanks to the *Web 2.0*. Others suggest that it is simply a redefinition of old technologies as Virtualization, Grid Computing or Utility Computing. Indeed, Cloud Computing includes these technologies (and many others) in a much wider scope. As a first approximation, we could define Cloud Computing as a solution where resources (hardware, software, storage, network, etc.) are dynamically provided on-demand [16]. Furthermore, the Cloud must guarantee high availability, security and quality of service. To this aim, it is important the ability to scale resources such that the user obtains from the Cloud exactly the amount of required resources, no more no less.

The Cloud aims at solving the following problems, the current IT industry is facing:

- Costs: Hardware and software capitalization and maintenance costs are continuously increasing;
- Scalability: Businesses need to invest more on hardware as the demand increases;
- Flexibility: Hardware is no more needed after a change on the service, or is outdated;
- Availability: Throughout the globe, resources need to be always available;
- Reliability: Redundancy is required for facing potential disasters;
- Updates: New software or hardware updates are difficult and costly to integrate in the business.

These elements, commonly defined for Cloud Computing are directly applicable to the SDR clouds and the wireless industry.

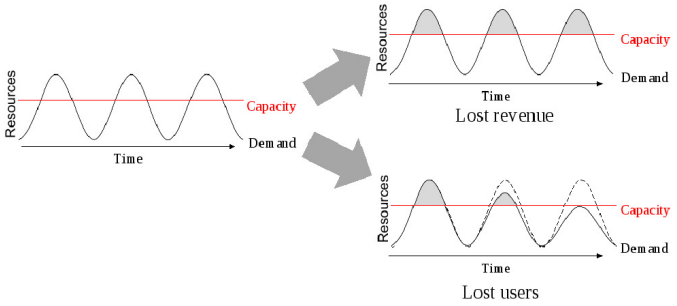

Figure 2.1: Effect of an underestimation of resources given a time-varying demand: loss of revenue due to rejected users and loss of users due to bad service.

2.2.1 Economical Factors of the Cloud Client

The Cloud concept is exploited by the client to transform CAPEX to operational expenditures (OPEX). The client pays for computing time, which can be used at his discretion. For instance, 100 CPU-hours can be used in 1 hour, running 100 processors in parallel, or during 100 days, running 1 hour each day. This is known as *utility computing*. This kind of products are more expensive than buying the material and then amortizing the cost over its lifetime. On the other hand, utility computing offers to important advantages: *Elasticity* and *Risk Transfer*.

Elasticity is the capability of a system to adapt to different environments, for instance, a computing system is elastic if it can adapt to time-varying load. The typical average load of IT servers is 5-20%, because the peak load is 2 to 10 times higher than the average. While the provisioning of physical resources is limited by the time it takes to buy and receive new hardware, the provisioning of virtual resources is done by the Cloud almost instantaneously. Therefore, if resource are allocated for its peak use, resources are wasted and the average cost of utilization time after amortization increases. This fact may already compensate by the higher cost of Cloud resources utilization time.

An underestimation of the required resources can have worse consequences for the service provider. Users that are rejected due to insufficient resources result in loss of revenue and, most important, loss of clients. The average number of users decreases until almost all of them can be served, which then results in the overestimated scenario (Fig. 2.1).

High elasticity implies high *risk*, due to the difficulty of an accurate estimation of the demand. The Cloud allows the client to transfer the risk to the provider. The Cloud provider is subject to a much lower elasticity requirement because it owns a very large infrastructure serving many clients running different applications with different usage profiles. Nevertheless, it is expected that when resources are acquired in advance for a large period of time the cost per time unit is lower.

We can find an insightful parallelism with a much more familiar utility: the electricity. For a large industry, it is often more convenient to buy or build a generator to provide electricity. The factory does not depend on external sources or companies to keep working and the cost per Watt will be probably lower. However, deciding the required generator power is risky. If it is overestimated we get a huge capital which is underused; on the other hand, the factory has to stop if it has not enough power. The solution is to outsource the production of energy and buy it to an electrical provider, even if it is more expensive. On the other hand, often the best solution is a combination of two energy sources, outsourcing and self-generation.

The same hybrid solution is found in the computing market. A private Cloud is a Cloud with a single client. Virtualization technology allows allocating resources dynamically in the

Table 2.1: Economy of scale in the Cloud

Resource	Cost in Medium DC	Cost in Very Large DC	Ratio
Network	$ 95/Mbps/month	$ 13 /Mbps/month	7.1x
Storage	$ 2.2/GB/month	$ 0.4 /GB/month	5.7x
Administration	≈ 150 servers/admin	$ ¿1000 servers/admin	7.1x

private or public Cloud seamlessly. It is feasible to quantify when it is worth to work on the public Cloud. Inequality (2.2.1) compares the net benefit of allocating a task in the Cloud with the net benefit of allocating it on the private Cloud:

$$T_{\text{public}} \left(\text{Income} - \text{Cost}_{\text{public}}\right) \geq T_{\text{private}} \left(\text{Income} - \frac{\text{Cost}_{\text{private}}}{\text{Utilization}_{\text{private}}}\right), \tag{2.2.1}$$

with T denoting utilization time, in time units, Cost is the cost in currency units per time unit and $0 \leq \text{Utilization} \leq 1$. If the inequality holds it is more convenient to outsource the task to the public Cloud. Note that if we assume that $\text{Cost}_{\text{public}} = \text{Cost}_{\text{private}}$, then the inequality hold strictly, except when the utilization is 1, when it holds with equality.

There are two more risks which are worth discussing: software and hardware updates. Paradoxically, if the software is improved and it performs the same task in less time, the utilization time is reduced and the cost per time unit of acquired hardware increases. On the other hand, if the process is executed on the Cloud, it is always beneficial to reduce the task execution time because less time implies less cost per task. Technology advances, on the other hand, imply that the hardware lifetime is more difficult to predict. Then, new hardware technology can not be introduced in the private Cloud as fast as in the public Cloud.

2.2.2 Economical Factors of the Cloud Provider

The Cloud provider enjoys the benefits of the *economy of scale*. Table 2.1 shows an study conducted by Hamilton [17] in 2008. The study compares the cost of acquiring hardware for a medium and a large data centre. The difference is exploited by the Cloud provider to generate benefits.

The Cloud provider also exploits *consolidation* technology for reducing the energy consumption and reduce the costs. Virtualized machines with different loads are consolidated in a single physical machine, so that unused machines can be switched off and the utilization of active machines is maximized [16].

2.3 Software-Defined Radio

The high revenues of the wireless business have increased the variety of wireless communications services and RATs. Coexisting RATs serve a growing number of wireless subscribers with personalized quality of service (QoS) demands. Sophisticated radio standards, such as Worldwide Interoperability for Microwave Access (WiMAX) and LTE, are emerging around two main concepts that arose from the wireless evolution: physical layer management and spectral efficiency. Whereas the former allows for many operational modes providing a higher degree of waveform flexibility, the latter increases the number of bits per Hertz through MIMO or spatial coding techniques, among others. The increasing flexibility at the physical layer requires additional management elements at the radio infrastructure (hardware and software), making the system more complex and less efficient in terms of computing resources per transmitted bit.

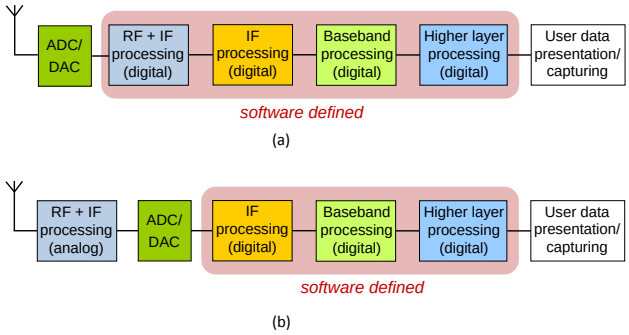

Figure 2.2: Ideal software radio (a) and practical SDR implementation (b).

The term *digital receiver* was coined in 1970 by researchers at a US DoD laboratory. The term defined a transceiver where the baseband processing was performed in software. In the early 90s Mitola envisaged radio transmitters and receivers (transceivers) that implement the entire signal processing chain in software (Fig. 2.2.a). He coined this vision software radio [18, 19]. Software radio describes multistandard, multiservice, and multiband radio systems, which are software-reconfigurable or reprogrammable. It promises to become a pragmatic solution to the variety of available and incompatible radio standards [20]. The technological difficulty for digitalizing radio frequency (RF) signals [21, 22, 23] led to the introduction of SDR [24]. SDR may be considered as a generalization of software radio. It characterizes a transceiver that implements one or more signal processing blocks in software. Since digitalization usually takes place at the intermediate frequency (IF) stage (Fig. 2.2.b), (part of the) digital IF processing, (part of the) baseband processing, and (part of the) higher protocol layer processing can then be implemented in software [25].

An SDR platform stands for software-programmable computing equipment–a handset transceiver or a network element–whereas an SDR application refers to a RAT-specific digital signal processing chain implementing a radio transceiver or part of it. Reconfiguring an SDR platform for executing another SDR application may then change the radio operational mode or establish a completely different wireless communications link. This facilitates deploying the most suitable waveform as a function of environmental stimuli and management policies.

Wireless communications systems were traditionally built around processor arrays: DSPs, GPPs, and ASICs. The multiprocessing concept is currently explored for future SDR platforms, evaluating the applicability of parallel processing devices, such as FPGAs and Multi-Processor System On-Chips (MP-SoCs). Using specialized hardware is more energy and cost efficient than using general-purpose hardware. GPPs, on the other hand, are able to run non-wireless applications more efficiently. This is specially interesting on SDR clouds, because the data centre infrastructure should be used for other tasks during hours of low wireless traffic. Today's many-core GPP have sufficient capacity to run modern signal processing algorithms efficiently [11]. Throughout this thesis, we will assume a data centre consisting on networks of multi-core GPP.

2.4 Distributed Open-Source SDR Frameworks: ALOE

The dynamic deployment of SDR applications on multiprocessor platforms however requires abstracting hardware resources and offering software services for a controlled use of these resources. This control and automatic allocation of resources in a distributed environment is the basis of the virtualization features required for the application of the SDR cloud concept. An execution environment or SDR framework provides these features.

Execution environments for general purpose computing platforms and applications include operating systems (OSs), such as UNIX or POSIX, higher level abstractions or virtual machines, object oriented architectures, such as the Common Object Request Broker Architecture (CORBA), and grid computing environments. While conceptually appropriate for the SDR computing context, these environments do not specifically address digital signal processing applications and introduce timing, memory, and power consumption overheads that make it difficult to meet the real-time computing constraints of SDRs. Although any execution environment or middleware introduces resource overheads, the benefits of isolating SDR platforms from applications overweight the costs:

1. Portability and reuse of hardware and software components,
2. Compatibility and integration of RATs and networks
3. Individual hardware and software development, service diversification, new business models, and so forth.

General-purpose computing knowledge and experience has been adapted to the digital signal processing world and SDR frameworks emerged. The two most popular SDR frameworks are the Software Communications Architecture (SCA) (http://sca.jpeojtrs.mil/) and GNU Radio (http://gnuradio.org/trac). The SCA is developed by the U.S. Department of Defense. It is the most widespread SDR research project that tries to reduce the development and deployment costs of SDRs by employing common middleware techniques (CORBA). Interfaces between waveform components are well specified, enabling compatibility between different SCA implementations. The SCA is deployed in commercial and open-source frameworks. GNU Radio, on the other hand, is an open-source SDR project targeting research and education. It is a low-cost solution for rapidly testing new algorithms and waveforms on GPPs.

Although the SCA and GNU Radio have both been designed for SDRs, they are general enough for being applied in other computing contexts. This increases their scope but decreases the performance of digital signal processing applications and resource-constrained platforms. The abstraction layer and operating environment (Abstraction Layer and Operating Environment (ALOE)) is an alternative SDR framework targeting multiprocessor platforms and embedded systems with tight resource constraints [26, 27] (http://github.com/flexnets/aloe). It is a lightweight, open-source SDR framework with cognitive computing resource management capabilities [28, 29]. The ALOE is not SCA-compliant; it uses a specific message passing scheme instead of CORBA.

ALOE is a continuously evolving SDR framework with its roots at the Platform-Hardware Abstraction Layer (P-HAL) [30]. This section presents the design goals, concepts and functionalities, architecture and services, application modules, and computing resource management capabilities of ALOE.

2.4.1 Design Goals

SDR applications are high performance computing applications with strict timing and power constraints. Power efficiency can be achieved by integrating a large number of heterogeneous processing devices, each specialized for a certain set of operations. This increases the hardware performance, but decreases the system flexibility and may lead to hardware and software compatibility and maintenance issues.

Developing applications for highly constrained processing tasks is time consuming and requires a lot of design effort. Large applications that need to execute on several devices are especially difficult to synchronize and route. Moreover, performance does not scale well with

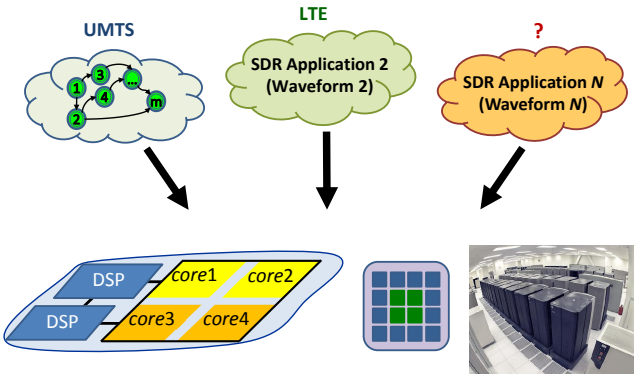

Figure 2.3: ALOE is designed for heterogeneous MP-SoC, multi-core processors or distributed processors in a data centre.

technology, often requiring the redesign of applications. These aspects suggest employing reconfigurable hardware and introducing abstraction or virtualization layers for creating a suitable execution environment or middleware. The application development can then be isolated from the platform design. A careful balance between a flexible solution and a specific design can provide a flexible environment that exploits the SDR application characteristics and meets the required performance. That is, the SDR middleware should provide scalable performance, flexibility, and computing efficiency for digital signal processing applications executed on heterogeneous processing devices with limited computing resources.

2.4.2 Concepts and Functionalities

ALOE supports heterogeneous multiprocessor platforms, where a single silicon device may encapsulate several processing elements (PEs) or distributed networks of processors in a data centre (Fig. 2.3). The aim of ALOE is to seamlessly load different waveforms regardless of the platform architecture. A PE here abstracts any bounded silicon area—static logic (GPP, DSP, ASIC) or dynamically reconfigurable area (DRA). PEs with multi-tasking operating systems access the ALOE services through an application program interface (API). PEs without operating system (e.g. ASICs) access the middleware services through a static logic interface [31, 32].

ALOE abstracts the platform's physical network interfaces. Depending on the network characteristics (mesh, star, shared bus, or any other) different throughputs and latencies are achievable. The task of ALOE is to deal with the network particularities while providing inter-PE communication capabilities as a service. Fig. 2.4 illustrates the ALOE layers. The hardware layer at the bottom shows a cluster of PEs and their physical interconnection. The abstract application layer at the top consists of graphs that model the application tasks and their data flow dependencies. At the real application layer these tasks are treated as individual modules, which use the ALOE services for assembling the waveform. The platform layer provides a pseudo-homogeneous and virtual execution environment, where all tasks see the same abstract platform, the ALOE platform.

The main functionalities of ALOE are:

1. real-time execution,
2. waveform execution control,
3. synchronized distributed computing,

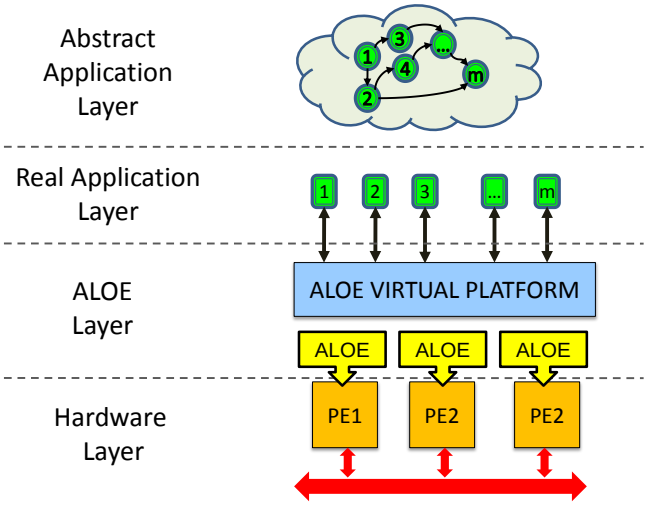

Figure 2.4: The ALOE layers

4. packet-oriented data flows,
5. cognitive computing resource management, and
6. external configuration and management.

2.4.3 Architecture and Services

The ALOE architecture encompasses several components that enable a hardware-independent access to the ALOE services. The framework is divided in two parts:

- **RTDAL** or Real-Time Distributed Abstraction Layer, is a framework for general-purpose real-time execution of tasks in a distributed environment. Provides platform-independence through the abstraction of platform-specific services.
- **OESR** or Operating Environment for Software Radio, provides middleware services designed for SDR applications. It requires RTDAL to execute;

RTDAL

RTDAL facilitates real-time synchronous execution of tasks in a distributed and heterogeneous environment. Tasks are executed periodically on each processor in a pipeline fashion. Each processor creates one thread per core, which runs each task (dynamically loaded as a shared library) one after another. The threads period on each core of each processor are continuously synchronized, offering the user an abstracted virtual platform. It is also possible to synchronize the task execution with a digital converter (AD/DA) for coherent transmission and processing of samples.

Besides, the RTDAL API also provides other functions:

- Task management, the following types of tasks can be created:
  - High-priority synchronous tasks. These tasks are non-preemptable and run synchronously with the platform time slot (synchronized throughout all the distributed environment).
  - Low priority tasks, which are preemptable can be:

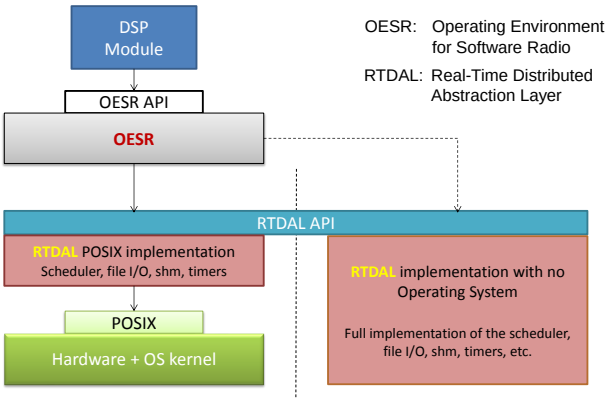

Figure 2.5: ALOE architecture

  - Synchronous, running with a periodicity multiple of the time slot or
  - Asynchronous, running only once.

- Interfaces: Two tasks sharing a common interface can communicate between each other. Interfaces can be:
  - External to communicate tasks running in remote processors, or
  - Internal, to communicate tasks running in the same processor (may be different cores, but sharing a common memory). Internal interfaces, at the same time support two communication mechanisms:
    - Flow interface, where the transmitter task supplies a memory address whose contents are physically transferred to the receiver task, or
    - Zero-copy interfaces, where only a pointer is transfered between the transmitter and receiver modules, minimizing memory bandwidth consumption. The current implementation employs a wait-free SPSC bounded queue.

- Others: AD/DA abstraction, time functions, shared memory, file I/O, etc.

The current implementation of RTDAL is based on POSIX interface and uses GCC atomic extensions. Therefore it targets common Linux distributions. However, the interface is very lightweihgt, thus possible to be ported to other platforms (e.g. Windows, Texas Instruments RTOS, etc.) with minimal implementation effort.

OESR

The OESR, on the other hand, is built on top of the RTDAL. This allows future portability to other platforms. While the distributed communications, synchronization and scheduling is provided by RTDAL, OESR provides functionalities specifically tailored for SDR applications:

- Automatic mapping of waveforms to a set of processing cores (distributed, in a multicore or both).
- Location-transparent inter-module communications.
- Global variables and parameters configuration/visualization
- Logs, counters and others.

OESR is interfaced through two different APIs:

- OESR API is used by the DSP modules to interact with OESR, e.g. create and use interfaces, logs, counters, access parameters and/or variables, etc.
- OESR Manager API is used by the platform manager program (a GUI for instance) to manage the entire platform: load and run waveforms into the distributed platform, view and modify variables, etc.

2.4.4 Application Modules

As for any execution environment, the applications and application modules need to follow a middleware-specific interface and execution pattern. In compliance with the directed data flow of digital signal processing applications, data is processed by a module and propagated to the next. A module retrieves new data from its input interfaces (first-in first-out (FIFO)) whenever needed and writes the processed data to the output interfaces (FIFOs). Notice that data types need to be compatible, an issue which remains at the application design phase. Whereas a module can have several virtual data interfaces, only three control signals– a status switch indicator (input), a current status indicator (output), and the finished flag (output)–are required for controlling its execution.

Fig. 2.6 illustrates the module execution flow: After registering to ALOE, the module falls in an idle state waiting for the initialization order. In the INIT phase, communication interfaces are initialized and configuration parameters, such as filter coefficients, are obtained or computed. The timing constraints are relaxed during the INIT phase. Once the module enters the RUN phase, it performs digital signal processing task in a real-time loop: read new data from the input FIFOs, process data, and write the results to the output FIFOs. ALOE, supported by the underlying OS, if available, schedules this execution flow and ensures a synchronized data processing. Since a module typically needs to meet certain timing constraints, the finished flag is processed by the execution controller for recognizing any timing violation. The STOP phase finally terminates the module execution and deallocates the corresponding computing resources.

ALOE, apart from ensuring real-time execution, coordinates the distributed application processing. The INIT, START and STOP phases are precisely scheduled for not interrupting the continuous data flow despite the interprocessor propagation delays. During a partial or total waveform reconfiguration, ALOE can schedule the modules' load and initialization phases as part of the execution pipeline (described below) for ensuring the data flow continuity.

2.4.5 Computing Resource Management

The multiprocessor mapping and scheduling of real-time constrained applications is a complex management problem that has been thoroughly studied in the heterogeneous computing context. Most approaches target the application speedup. In SDR, however, other objectives are prevailing: (1) meet all real-time computing requirements, (2) support RRM decisions, and (3) achieve higher computing efficiencies. ALOE supports these through computing resource awareness and monitoring (cognition), time management and execution control, coherent computing system modeling and flexible resource management.

Resource Awareness and Monitoring

Computing resource awareness and monitoring enables the dynamic reconfiguration and management of SDR platforms and applications. A cognitive system in this context autonomously

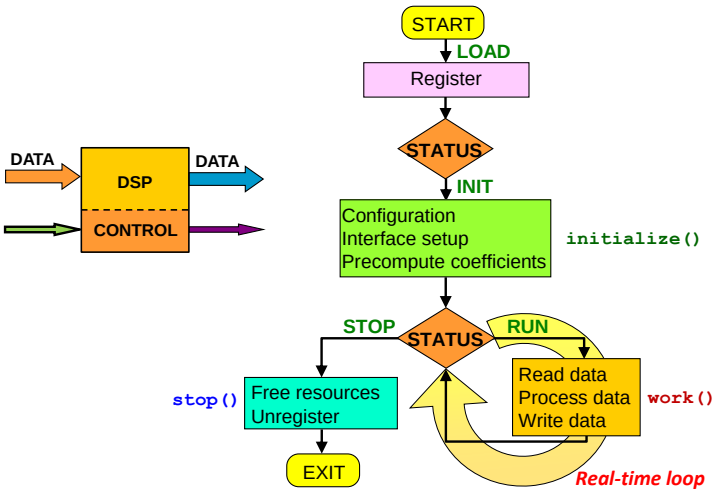

Figure 2.6: Module execution flowchart.

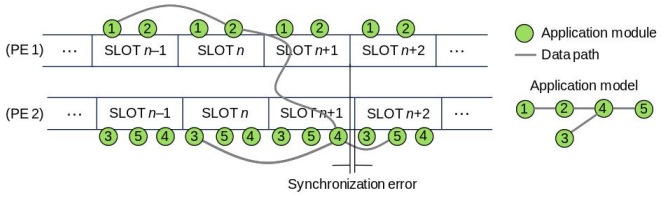

Figure 2.7: Time slots and pipelining.

and continuously tracks the computing resource states and requirements of SDR platforms and applications. The dynamic monitoring of running tasks (application modules) and system processes moreover facilitates providing execution time statistics, such as stack, memory, and silicon area occupation over time [33]. At system boot, each PE runs a common test that measures its computing capacity. The test consists of computing a large sequence of multiply-accumulate (MAC) operations for obtaining the processing capacity in million MAC operations per second. The platform's internal communication bandwidths, memory and FIFO capacities, and so forth are also measured on PE basis. The platform's computing resource information then contains [33]

- the PE classes: processor or reconfigurable area,
- the instruction-set family of each PE: x86, c64x, ARMx, . . . ,
- the number of PEs and their processing capacities,
- the inter-PE communication bandwidths, and
- additional status information, such as energy consumption and PE clock frequencies.

A periodic process continuously reports the computing resource states of PEs, enabling the detection of failures or malfunctioning. If a PE stops reporting its status, it is removed from the platform model. The system, in this case, reallocates the corresponding tasks, if possible, or stops the application, if not.

Time Management and Execution Control:

The ALOE framework splits the computing resource time in discrete computing time slots and executes application modules in a pipelined fashion. This means that a data block produced in time slot n is not consumed before the beginning of time slot $n + 1$ (Fig. 2.7). This facilitates synchronizing the distributed data processing and ensures deterministic computing delays—data processing and propagation.

The proposed execution pattern eliminates the modules' precedence constraints within a time slot. That is, the execution order of modules is irrelevant as long as all data processing and propagation finishes within the time slot boundaries. The task precedence constraints are then maintained throughout the pipeline without the need for a hard real-time OS. Moreover, a proper selection of the time slot duration guarantees meeting the end-to-end latency requirements while preserving the continuous data flow. This is monitored by ALOE at the beginning of each time slot. Since application modules run at highest priority and cannot be interrupted, the start and finish times of modules provide a precise and instantaneous measure of the processor occupation.

The local clocks need to synchronize to a unique and reliable virtual time. SDR applications are constrained by the sampling rate, which is determined by the radio standard. Analog-to-digital and digital-to-analog converters are tuned to this frequency and provide the clock source for the time slot alignment. The entire processing chain needs to synchronize to this clock, ensuring that buffered samples will always be provided or consumed on time. The synchronization process is detailed in [30]. It ensures that the clock synchronization errors will be within a certain fraction of a time slot (Fig. 2.7). Since typical time slots for SDR applications are in the order of milliseconds, errors of tens of microseconds are easily achievable for most PEs, even for MP-SoCs and FPGAs, because of the devices' high operating frequencies. We can then create real-time applications without the knowledge of the platform's real-time capabilities.

Resource Modeling and Management

Because of the differing resource and execution concepts of the different processor types (parallel versus sequential processing - FPGAs versus DSPs), we generally consider the available silicon area and time as the fundamental processing resources. Based on benchmarks or implementations, these resources can be abstracted using appropriate metrics for an efficient resource management. Since dealing with digital signal processing applications, we suggest the MAC as the basic computing operation.

Computing resources are measured on time slot basis. Each PE, for example, is capable of performing a certain number of MACs per time slot. millions of operations per time slot (MOPTS) and megabits per time slot (MBPTS) abstract the processing powers and interprocessor data flow capacities. SDR applications are modeled as directed acyclic graphs (DAG), with nodes representing the application modules and arcs their data flow dependencies. The application's processing and data flow requirements are coherently specified in MOPTS and MBPTS. Considering time as an implicit resource facilitates meeting the waveforms' hard real-time processing requirements: Distributing the computing loads among the available computing resources (mapping) and determining the execution order of software modules (scheduling) on time slot basis ensures that the real-time computing constraints (maximum latency and minimum throughput) will be met with the available computing resources.

Since SDR requires a flexible, though efficient computing resource management, ALOE features a computing resource management API that facilitates implementing any suitable mapping algorithm. We propose the tw-mapping , a general-purpose mapping algorithm

based on the dynamic programming principle. The mapping policy and computing constraints are externally specified by defining a suitable cost function, which guides the tw-mapping process and dynamically updates the resource models. That is, the algorithm does not assume any particular SDR platform, application, or system constraint. It is scalable for different problem sizes, applicable in different scenarios, and adjustable for trading mapping performance against computing complexity.

2.5 Summary

The MD-MIMO cell-less network architecture promises to break the capacity and economical bottlenecks of today's cellular networks. SDR cloud is a technology that will enable to implement this architecture in a cost-efficient and economical sustainable manner. We have argued in this chapter that both network and infrastructure operators can benefit of the SDR cloud ecosystem. The application of the economy of scale and consolidation principles allows the infrastructure operator to amortize the investment. This investment is amortized between several network operators, which will find in the SDR cloud a place to scale smoothly according to their number of subscribers, as well as to adapt to new standards or wireless technologies.

To allow economically-sustainable large cell-less networks based on Cloud technology, radio signals must be processed in general-purpose hardware. This is essential in order to maximize the utilization of the deployed resources (computing) during hours of low wireless traffic. SDR is today a mature technology. Commodity servers can be used for processing modern RAT signals in real-time [11].

An SDR framework is the element that connects SDR with virtualization. We have reviewed the features and architecture of ALOE. The design and development of ALOE provided important insights on the needs and constraints found in distributed environments and signal processing applications as found in the SDR cloud. ALOE provides automatic resource allocation and resource control, essential for supporting processor time virtualization. It is capable to work in a distributed environment, as the one found in the SDR cloud. Resource modeling allows characterizing the available resources in the data centre, enabling the coexistence of real-time and non real-time tasks.

Chapter 3

The SDR Cloud

3.1 Introduction

Whereas current systems provide data rates of a few mega-bits per second (Mbps), 4G systems will offer up to 100 Mbps per user. A few seconds may be necessary today before a connection is established between the user equipment and the network. Long term evolution (LTE) and LTE-Advanced (LTE-A) promise connection establishment times of less than 50 and 10 ms, respectively [34, 35].

The real-time processing and data flow demands of SDR applications are a function of the service, QoS, and channel conditions. Digital signal processing algorithms are continuously evolving. Many processing operations are applied in the same or similar form in various radio standards. This permits reusing software modules for assembling different radio transceivers. An SDR system is, hence, a dynamic system, where radio and computing resources are continuously reassigned.

Managing the SDR cloud computing infrastructure within the tight timing constraints of wireless communications services is a complex task. We therefore suggest dividing the data centre into clusters of a few processors each. A high-level resource manager assigns users to clusters or cluster groups as a function of radio communications and cloud computing conditions. Distributed low-level resource managers then allocate and deallocate cluster computing resources in real-time. That is, whenever a user wishes to initiate or terminate a wireless communications session, the corresponding low-level resource manager allocates or deallocates computing resources. Computing resources should therefore be assigned on a first-come first-serve basis. We can then formulate the resource management requirements for SDR clouds as:

1. Respond, that is, allocate the necessary amount of computing resources, in real-time,
2. Provide real-time control over the processing throughput and latency,
3. Manage different QoS targets,
4. Identify and track the system resource states and handle multiple computing constraints,
5. Provide computing resources as a single abstract resource, shared between different users, and
6. Adapt the resource management strategy to internal (SDR cloud) and external (environmental) influences or conditions.

The rest of the chapter is organized as follows: After providing some background on computing resource management methods and algorithms (Section 3.2), we identify the problem (Section 3.3.1) and elaborate a RA complexity model (Section 3.3.2). In the central part of the chapter we define and solve an optimization problem for assigning computing resources to an RA as a function of the environmental parameters (Section 3.3.3). We finally apply our solution for managing the resources associated to a single radio cell (Section 3.3.4) and multiple cells (Sections 3.4 and 3.5) under different wireless communications traffic characteristic.

3.2 Related Work

Massively parallel computing architectures will dominate the high-performance computing landscape. A platform with a large number of parallel processors is more suitable for executing many applications than a single powerful processor [36]. The high and heterogeneous computing demands of SDR applications, in particular, are executed more efficiently on a multiprocessing execution environment [37, 38]. Empirical studies have shown that scheduling hard real-time tasks on many-core processors is challenging [39, 40]. Sophisticated resource allocation algorithms are consequently necessary for managing the real-time computing demands and the limited computing resources.

Distributed computing has a long research record. The multiprocessor mapping and scheduling problem, in particular, has been vastly investigated in the heterogeneous computing context [41, 42, 43]. Heterogeneous computing refers to a coordinated use of distributed and heterogeneous computing resources [44]. It is similar to grid computing [45] or metacomputing [46].

It is well known that the computing resource allocation problem is NP-complete, in general [47]. Heuristic approaches were therefore proposed, presenting a polynomial relation between the problem size and the computing complexity. Grid or cloud computing RAs dispatch computing jobs or independent task for their distributed execution. Grid computing workloads exhibit little intra-job parallelism, the average job completion time is several hours, and typical job inter-arrival times are in the order of seconds or minutes [48]. Many grid or cloud workloads are data-intensive [49].

Grids and clouds are accessed via the internet, which is relatively slow and has unpredictable delays. They were originally built for providing very high computing power for scientific or popular applications with no stringent real-time constraints. Rather than ensuring real-time allocation and execution, grid or cloud RAs therefore follow other objectives. Doulamis et al. [50], for example, discuss the fair sharing of CPU rates and allocate resources to users as a function of resource availabilities, user demands, and socio-economic values. Lui et al. [51] focus on the joint resource allocation of computing and network resources in federated computing and network systems. They present various resource allocation schemes that can provide performance and reliability guarantees for modern distributed computing applications. Entezari-Maleki and Movaghar [52] develop a probabilistic task scheduling method for minimizing the mean response time of grid jobs.

The SDR cloud concept has been recently introduced [10] and merges three fundamental technologies: centralized baseband processing, automatic computing resource allocation and virtualization. Related work addresses centralized baseband processing [12, 13] and offline, that is, design-time resource allocation [53]. We focus on automatic computing resource allocation, enabling runtime resource management and seamless real-time execution. Each wireless communications service request needs to be served in real time, providing sufficient computing resources for the continuous real-time data processing. Two general approaches exist for scheduling real-time tasks on multiprocessor platforms. Tasks can be statically as-

signed prior to execution or migrate between processors during execution. The former can be achieved through partitioned scheduling, where an application is partitioned among the processing elements (mapping) before being locally scheduled. The latter approach is typically associated with global or dynamic scheduling. The contention for the global scheduling queue and non-negligible migration overheads among processing elements can result in significant scheduling overheads in practice [39]. The migration cost limits the number of cores that a global scheduler can manage [39, 40]. Non-preemptive static partitioned scheduling, on the other hand, is pertinent to high performance many-core and multiprocessor platforms. It facilitates implementation and introduces low run-time resource overheads [54].

A constant execution period and practically deterministic and regular execution patterns characterize SDR applications. The real-time constraints of the DSP processing chains can then be given as minimum throughput and maximum latency constraints and static schedulers can be employed [37]. The mapping and scheduling can thus be calculated only once for each waveform as part of the session establishment process. The SDR cloud resource management performance is then limited by the RA's execution time per invocation (user session request) and the session arrival rate.

3.3 Fundamental Limits of the SDR Cloud

This section elaborates a relation between the wireless communications system requirements or constraints and the SDR cloud computing resource management before deriving optimal solutions for the high-level resource provisioning [55].

Each service request requires loading the corresponding transceiver waveform. Real-time resource provisioning and *hard* real-time execution needs to be ensured for seamless service provisioning. The SDR cloud resource allocator (RA) will therefore determine the mapping of waveforms to the available computing resources on demand and under stringent timing and resource constraints. We show that the maximum traffic load that a single RA can handle is limited. It is a function of the complexity of the resource allocation algorithm, the call setup delay, and the user rejection or blocking probability. The radio access technology specifies the maximum call setup delay, whereas the radio operator determines a blocking probability target. We introduce a general execution time model for characterizing the complexity of different resource allocation algorithms and derive expressions for the average call setup delay and maximum traffic load. The results show that SDR cloud data centers can be efficiently managed in a distributed way. They provide guidelines for designing data centers and distributed resource management methods for SDR clouds.

3.3.1 Problem Formulation

Wireless subscribers access communications services anywhere, anytime, and under different circumstances. Measurements have shown that the average user establishes seven or eight voice sessions per day of 90 seconds in the mean [56]. Data users realize a larger number of shorter sessions. The number of concurrent sessions in a large city may range between 10,000 and 120,000 as a function of place and time.

The SDR cloud RA needs to be able to handle the spatial and temporal variety in the traffic load. A single data centre ideally executes all waveforms and centrally manages all session requests. The corresponding RA then needs to be able to dispatch thousands of requests per second.

Modern wireless communications standards, however, impose restrictions on the maximum session establishment time $t_s^{\max}$. The call setup delay t_s is the transition time from a dormant (camping or idle [34]) state to the transmission or reception state. Each session

establishment here consists of allocating sufficient computing resources to the corresponding transceiver waveform. The shorter the call setup time the better the *always connected* illusion. LTE-A therefore establishes 10 ms as the target call setup delay. Wireless operators moreover define a maximum blocking probability target $p_b^{\max}$, which should be satisfied in the mean. The blocking probability p_b denotes the probability of a user session request being rejected due to insufficient computing resources. Wireless communications systems need to be accordingly dimensioned.

The session establishment time and blocking constraints determine the RA capacity in terms of manageable users. The number of users that can be concurrently served is directly proportional to the available processing resources. More processors ensure a lower p_b, whereas fewer processors a shorter t_s. The objective of this chapter is analyzing the relation between the RA capacity and the call setup time and blocking probability. We identify fundamental SDR cloud management limits and indicate possible SDR cloud data centre design and management solutions.

Table 3.1: Description of parameters

Parameter	**Description**
n	Number of nodes or processors
m	Number of waveform modules or tasks
t_s	Call setup delay
$t_s^{\max}$	Call setup delay constraint
p_b	Blocking probability
$p_b^{\max}$	Blocking probability constraint
t_{RA}	Resource allocator's (RA's) execution time model
F	Scaling factor of RA model
α	Nodes' exponent (n^α) of RA model
β	Modules' exponent (m^β) of RA model
θ	Cost function's weight
ρ	Traffic load in Erlangs
$\rho^{\max}$	Maximum traffic load a single RA can manage
λ	Average session initiation requests per second
$1/\mu$	Average session duration in seconds
$\Phi(n)$	Number of users that can be served with n processors for a given waveform model

3.3.2 Resource Allocator Complexity Model

The algorithmic complexity of any RA is a function of the number of tasks m and the number of processing cores or nodes n. A polynomial expression can be used for modeling the complexity of practical RAs, such as

$$t_{\mathrm{RA}}(n, m) = F n^\alpha m^\beta. \tag{3.3.1}$$

Parameters α and β specify the complexity order of a RA. The same expression also serves as a general execution time model of a RA implementation. Parameters F, α, and β can be found by measuring the RA execution time for different n and m and then performing model fitting. Although other models may be more accurate for certain RA algorithms, (3.3.1) is simple and general.

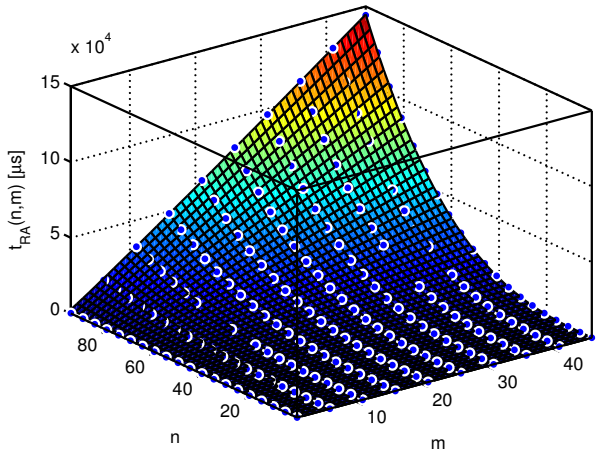

Figure 3.1: Measured and modeled execution times of the g-mapping RA.

Without loss of generality, we suggest a simple and well-known algorithm for providing numerical examples for the analysis performed in this chapter. The g- or greedy-mapping [37] is a baseline mapping algorithm. It maps one process after another, choosing the processor that leads to the minimum mapping cost. Cost metrics are therefore computed based on a suitable cost function. The cost function we suggest manages the limited processing and interprocessor bandwidth resources and, accordingly, distributing the processing load while minimizing the data flows between processors [37].

The algorithm is implemented in C and available as open source code [57]. Measuring the execution time of our implementation as a function of n and m and performing non-linear least-squares model fitting leads to $A = 3.98 \cdot 10^{-9}$, $\beta = 1.04$ and $\alpha = 2.94$. We can thus approximate the execution time model of the g-mapping algorithms as

$$t_{\mathrm{RA}}(n, m) \approx 4mn^3[ns]. \tag{3.3.2}$$

The g-mapping execution time thus increases linearly with the number of waveform modules m and with the number of processors n cubed. Figure 3.1 plots the execution time measurements together with the least-squares model.

3.3.3 Resource Provisioning

Throughout this section we will use the previously derived execution time model (3.3.2). The analysis is also valid for other RA complexity models provided that the complexity increases with the number of processors.

Optimization Problem

We analyze the relation between the call setup delay, the blocking probability, and the RA capacity. To this aim, we derive the optimal number of resources for processing user signals as a function of the environmental conditions and constraints. The wireless communications traffic model is a stochastic birth-death process. The time between consecutive session establishments follows a Poisson distribution with a mean of $1/\lambda$. That is, λ corresponds to the average number of new user session requests per second. The session duration follows an exponential distribution, where $1/\mu$ corresponds to the average session duration in seconds. The traffic load is then $\rho = \lambda/\mu$ Erlangs.

We assume that a single RA needs to handle ρ Erlangs of traffic. The objective is then determining the optimal number of processors n that satisfies the system constraints $t_s^{\max}$ and $p_b^{\max}$. This value is obtained as the solution to an optimization problem maximizing the following objective function:

$$f(n) = (1-\theta)\,U(n) - \theta\frac{n}{\rho},\ \rho > 0, \tag{3.3.3}$$

$U(n)$ is the average number of users that can be served with n processors. Function $f(n)$ weights off the benefit (average number of served users) and the cost (allocated resource per Erlang). Parameter θ weights the importance of one term with respect to the other. Equation (3.3.3) allows minimizing the number of allocated resources n ($\theta = 1$) or maximizing the average number of served users $U(n)$ ($\theta = 0$). Applying Little's law, we can express $U(n)$ as

$$U(n) = (1 - p_b(n))\,\rho. \tag{3.3.4}$$

The optimization problem can then be formulated as follows:

$$\begin{aligned} \max_{n} \quad & f(n) \\ \text{s.t} \quad & p_b(n) \leq p_b^{\max} \\ & t_s(n) \leq t_s^{\max} \\ & n \in \mathbb{N}. \end{aligned} \tag{3.3.5}$$

Before solving this problem, we first need to model the call setup delay $t_s(n)$ and blocking probability $p_b(n)$ constraints.

Constraints

The session establishment process can be modeled as a double-queuing process: New users enter an infinite queue whose service time is the execution time of the RA, that is, $t_{\mathrm{RA}}(n,m)$. They leave this *allocation queue* and enter a second multi-server queue of size c. The service time of the *active sessions queue* is exponentially distributed with an average of $1/\mu$, which corresponds to the average session duration.

This model can be represented by a two-dimensional state transition diagram, where state probability $p_{i,j}$ indicates the probability that there are i users waiting for the allocation queue while j users have active sessions. The model can be simplified if we consider that the mapping time is much shorter than the average session duration, that is, $t_{\mathrm{RA}}(n,m) \ll 1/\mu$. This allows separating the two queues. Following Kendall's representation, we model the allocation queue as an infinite length $M/D/1$ queue and the active sessions queue as a blocking and finite-size $M/M/c/c$ queue with no wait sates. For simplifying the mathematical analysis, here we consider waveforms of $m = 10$ tasks and analyze t_{RA} as a function of n.

User session requests are random and independent from one another. The random session requests lead to random session establishment times. The call setup delay constraint (3.3.5) will thus be satisfied on average despite the deterministic mapping time $t_{\mathrm{RA}}(n)$. Applying the PASTA (Poisson Arrivals See Time Averages [58]) property, we know that $t_s(n)$ follows a Poisson distribution. According to the $M/D/1$ model, the average call setup delay then becomes

$$t_s(n) = \begin{cases} \frac{2-\lambda t_{\mathrm{RA}}(n)}{2(1-\lambda t_{\mathrm{RA}}(n))/t_{\mathrm{RA}}(n)} & \text{if } \lambda t_{\mathrm{RA}}(n) \leq 1 \\ \infty & \text{if } \lambda t_{\mathrm{RA}}(n) > 1. \end{cases} \tag{3.3.6}$$

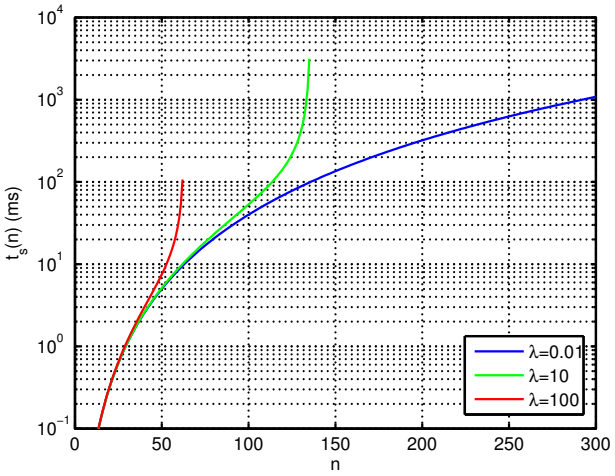

Figure 3.2: Average waiting time $t_s(n)$ as a function of n and λ for $t_{\mathrm{RA}}(n, m = 10)$.

This function is monotonically increasing with n (Fig. 3.2) for $\lambda t_{\mathrm{RA}}(n) \leq 1$. The system becomes unstable and the average waiting time infinite beyond that point. Figure 3.2 shows that the call setup delay limits the maximum number of processors that can be managed. For $t_s(n) = 100$ ms and $\lambda = 10$ user arrivals per second, for example, up to 150 processors can be managed with the g-mapping RA, but less than 100 processors for $\lambda = 100$. Figure 3.2, moreover, shows that a low $t_s(n)$ significantly limits the RA capacity.

The blocking probability of the active sessions queue ($M/M/c/c$ queue) is the probability that c users are occupying all available resources. When this happens, a new user is rejected due to insufficient computing capacity. Parameter c therefore represents the maximum number of waveforms that can be loaded to n processors. This number is difficult to characterize since depending on many factors, including the computig capacity of each processor, the interprocessor communication network, the waveforms' computing characteristics, and the performance of the RA algorithm.

For an analytical treatment the capacity of the queue needs to be abstracted. We propose defining $c = \Phi(n)$, which defines the maximum number of users that n processors can accept. Without loss of generality, we assume the linear model $\Phi(n) = n/k$. Parameter k is a real positive value and indicates the percentage of a single processor that is needed for executing a waveform. For $k > 1$, more than one processor is required for processing a single-user digital transceiver. For $k = 1.8$, for instance, one waveform requires 180% of the processing resources of a single processor for real-time execution. Note that $U(n)$, which provides the average number of users that can be loaded to n processors, depends on the traffic load and blocking probability, whereas $\Phi(n)$ essentially depends on the processor capacity, waveform characteristics, and RA algorithm efficiency. The blocking probability of the $M/M/c/c$ queue with $c = \Phi(n)$ is then

$$p_b(\rho, \Phi(n)) = \mathrm{B}(\rho, \Phi(n)), \tag{3.3.7}$$

where $B(\rho, c)$ is the Erlang-B function [58] for ρ Erlangs and c servers. Figure 3.3 indicates the evolution of the blocking probability as a function of n for different k.

Solution

The objective function (3.3.3) is strictly concave because the blocking probability (3.3.7) is strictly convex [59]. This ensures that the optimization problem has a unique solution. Figure 3.4 plots the objective function $f(n)$ for different weights θ.

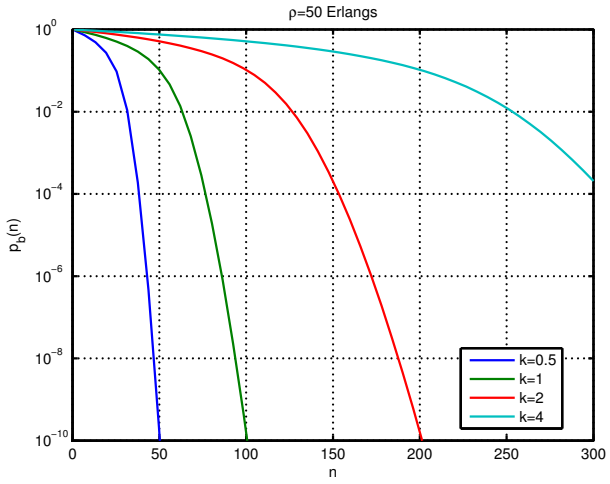

Figure 3.3: Blocking probability $p_b(\rho, n/k)$ as a function of n for different k.

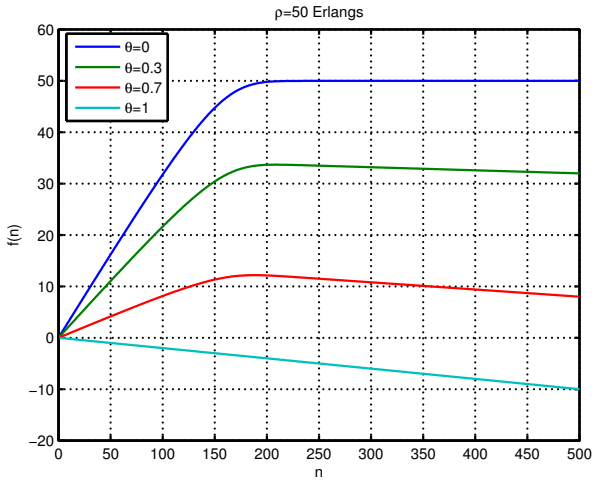

Figure 3.4: Objective function $f(n)$ for $\rho = 50$ for different θ.

The solution is trivial for $\theta = 0$ or $\theta = 1$, because the objective functions and the constraints are monotonic with n over the entire range of processors. More precisely, $p_b(n)$ decreases and $t_s(n)$ increases with n. We therefore define $n_{\min}$ as the minimum number of processors that satisfies the blocking probability constraint $p_b^{\max}$ and $n_{\max}$ as the maximum number of processors that meets the call setup delay limit $t_s^{\max}$. That is, $n_{\min}$ and $n_{\max}$ limit the number of processors to a range that provides the desired quality. They satisfy

$$p_b(n_{\min}) \leq p_b^{\max},\ p_b(n_{\min} - 1) > p_b^{\max} \tag{3.3.8}$$

and

$$t_s(n_{\max}) \leq t_s^{\max},\ t_s(n_{\max} + 1) > t_s^{\max}. \tag{3.3.9}$$

We can say that $n = n_{\min}$ processors minimize the number of resources, whereas $n = n_{\max}$ processors minimize the blocking probability, or maximize the average number of concurrently served users, while still satisfying the call setup delay constraint. If, however, $n_{\min}$ exceeds $n_{\max}$, the two constraints cannot both be satisfied and the problem becomes unsolvable. The solutions that minimize the blocking probability ($\theta = 0$) and the allocated processors ($\theta = 1$)

then become

$$n^*_{\theta=0} = \{n_{\max} \mid n_{\min} \leq n_{\max}\} \tag{3.3.10}$$
$$n^*_{\theta=1} = \{n_{\min} \mid n_{\min} \leq n_{\max}\} \tag{3.3.11}$$

We need to solve the problem numerically for arbitrary θ. The first option is using numerical optimization. Integer optimization problems are very complex to solve, though. We therefore relax the integer nature of the optimization variable n and use a convex solver for finding a non-integer solution. We then evaluate the objective function for the two closest integers, choosing the maximum that satisfies the constraints.

The Erlang's $\mathrm{B}(\rho, c)$ function is defined for natural c. The Erlang's extended B-formula is a continuous representation of the Erlang-B function based on the incomplete Gamma function. Computing this function numerically however requires numerical integration. We rather propose using the recursive method

$$\mathrm{B}(\rho, i) = \frac{\rho \mathrm{B}(\rho, i-1)}{\rho \mathrm{B}(\rho, i-1) - 1}, \tag{3.3.12}$$

where i is a real positive number. If we are able to obtain $\mathrm{B}(\rho, z)$ for a real number $z < 1$, then we can compute $\mathrm{B}(\rho, i)$ for any i. Various approximations for $\mathrm{B}(\rho, z)$ have been published based on parabolic interpolations. We used the expression of [60] for the numerical examples that follow.

Examples

More than one processor is typically needed for executing a modern waveform consisting of 10 or more tasks [37]. The numerical examples therefore consider $\Phi(n) = n/2$ allocatable users, $m = 10$ waveform tasks, and $1/\mu = 40$ s average data session duration. We use the interior-point numerical algorithm for solving problem (3.3.5) and obtaining a non-integer solution (Fig. 3.5).

Assigning n^* processors to the RA maximizes the system efficiency $f(n)$. The optimal number of processors n^* is a function of θ. The curves corresponding to $\theta = 0$ represent the solutions that minimizes the blocking probability ($n^* = n_{\max}$) while meeting the call setup delay constraint. The curves corresponding to $\theta = 1$, at the other extreme, indicate the solutions that minimize the use of processing resources ($n^* = n_{\min}$) while satisfying the blocking probability constraint. The intersection of these two curves provides the maximum system capacity $\rho^{\max}$. Parameter $n_{\min}$ becomes larger than $n_{\max}$ beyond that point and the problem has no solution.

The system capacity is almost 50 Erlangs for a call setup delay constraint of 50 ms, which corresponds to the LTE standard specification (Fig. 3.5.a and 3.5.b). LTE-A indicates call setup times of 10 ms, reducing the RA capacity to some 25 Erlangs in this case (Fig. 3.5.c and 3.5.d). The capacity can be improved by using more powerful processors. Assuming $\Phi(n) = n/1.5$, for example, leads to $\rho^{\max} = 35$ Erlangs for the LTE-A case (Fig. 3.5.e and 3.5.f).

3.3.4 Resource Allocator Capacity

The previous section has indicated that the RA capacity $\rho^{\max}$ is finite. Here we analytically derive this limit. The manageable number of processors is obtained from the tolerable execution time. The blocking probability then determines the RA capacity.

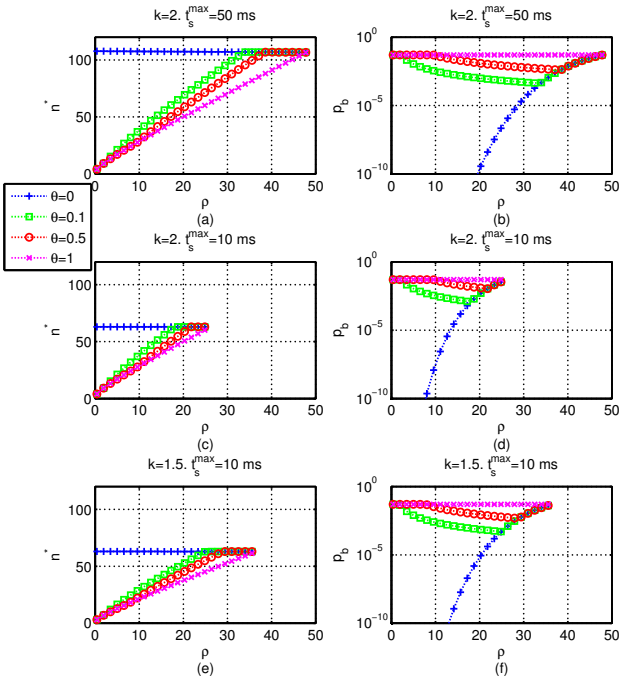

Figure 3.5: Optimal number of processors (a, c, e) and corresponding blocking probability (b, d, f) for different $t_s^{\max}$ and k.

RA execution time limit

The tolerable RA execution time $t_{\mathrm{RA}}^{\max}$ is a function of the call setup delay constraint and the user arrival rate. It is obtained assuming that the average call setup delay (3.3.6) is equal to the call setup delay constraint $t_s^{\max}$:

$$t_{\mathrm{RA}}^{\max} = \lambda^{-1} + t_s^{\max} - \sqrt{\lambda^{-2} + (t_s^{\max})^2} \tag{3.3.13}$$

Equation (3.3.13) can be simplified when $t_s^{\max}$ is either considerably smaller or considerably larger than the user inter-arrival time:

$$t_{\mathrm{RA}}^{\max} \approx \begin{cases} \lambda^{-1} & \text{if } t_s^{\max} \gg \lambda^{-1} \\ t_s^{\max} & \text{if } t_s^{\max} \ll \lambda^{-1}. \end{cases} \tag{3.3.14}$$

When $t_s^{\max} \gg \lambda^{-1}$, the capacity is limited by the stability of the $M/D/1$ mapping queue (see (3.3.6)). The call setup delay is then dominated by the time the user needs to wait before being be served rather than the RA execution time itself. On the other hand, when $t_s^{\max} \ll \lambda^{-1}$ the capacity is limited by the call setup delay constraint. This is the case with modern communications standards, such as LTE and LTE-A, where the call setup delay is dominated by the RA execution time.

The general expression of $t_{\mathrm{RA}}^{\max}$ is a function of λ (3.3.13). Therefore, $n_{\max}$ is also a function of λ.

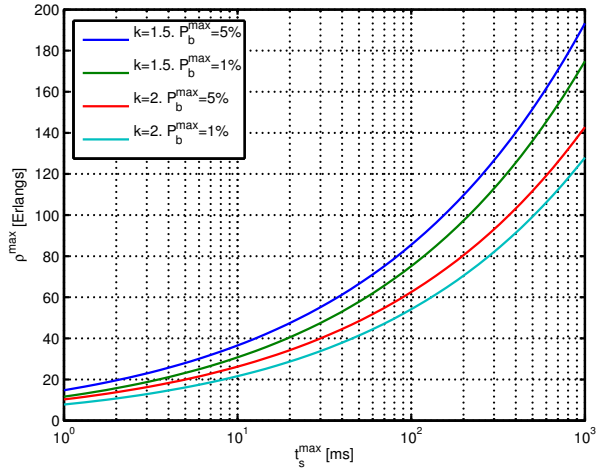

Figure 3.6: Capacity of the g-mapping RA as a function of $t_s^{\max}$ and k.

Processor limit

The maximum number of processors that a RA can manage is a function of $t_{\text{RA}}^{\max}$ and follows from inverting equation (3.3.1):

$$n_{\max} = \left\lfloor \sqrt[\alpha]{\frac{t_{\text{RA}}^{\max}}{Fm^{\beta}}} \right\rfloor . \tag{3.3.15}$$

The expression $\lfloor \cdot \rfloor$ indicates rounding off to the closest lower integer value.

Traffic limit

Considering the blocking probability constraint, the maximum traffic load $\rho_{\max}$ that the RA can manage is then the solution to

$$\begin{aligned} \text{B}\left(\rho_{\max}, c\right) &= p_b^{\max} \\ \Phi\left[n_{\max}(\rho_{\max}\mu)\right] &= c. \end{aligned} \tag{3.3.16}$$

The capacity in Erlangs is thus a function of the average user session duration μ. The expression can be simplified when $t_s^{\max} \ll \lambda^{-1}$. The maximum number of processors $n_{\max}$ then depends only on the maximum call setup delay constraint and the capacity becomes independent of μ. The maximum traffic load $\rho_{\max}$ that the RA can manage is then the solution to

$$\text{B}\left(\rho_{\max}, \Phi\left(n_{\max}\right)\right) = p_b^{\max} \text{ , when } t_s^{\max} \ll \lambda^{-1}. \tag{3.3.17}$$

Figure 3.6 plots the capacity of the g-mapping RA, assuming that $k = 1.5$ and $k = 2$ processors, are needed for digitally processing the signals of a single user. The figure assumes the approximation $t_s^{\max} \ll \lambda^{-1}$ and, thus, is the solution to (3.3.17). It shows that a single RA can handle up to 200 Erlangs, assuming legacy cellular communications standards, which are characterized by loose call setup delay constraints. However, the capacity drops way below 80 Erlangs for the emerging LTE and LTE-A standards, which establish maximum session establishment delays of 50 and 10 ms.

3.4 Distributed Resource Management

SDR clouds will provide wireless communications services to very wide service areas and will, consequently, need to manage huge traffic loads. Incoming user session requests will then need to be assigned to different RAs. Each RA will absorb only a portion of the total traffic demand, managing part of the data centre resources [61, 62, 63]. The assignment of processors to RAs should adapt to the traffic load distribution while satisfying the constraints of (3.3.5).

Here we assume a reduced SDR cloud model, where a data centre of N=100 processors serves two radio cells with a total traffic load of 40 Erlangs. The maximum call setup delay is 10 ms and the target blocking probability 5%. The capacity limit of a single RA is 35 Erlangs for $\Phi(n) = n/1.5$. We thus need at least two RAs, one per radio cell. The problem then consists of splitting the N processors between the two RAs in such a way that all constraints are satisfied. Problem (3.3.5) thus extends to

$$\begin{aligned} \max_{\{n_1,n_2\}} \quad & f(n_1) + f(n_2) \\ \text{s.t} \quad & p_b(n_i) \leq p_b^{\max},\ i = 1,2 \\ & t_s(n_i) \leq t_s^{\max},\ i = 1,2 \\ & n_1 + n_2 \leq N \\ & n_1, n_2 \in \mathbb{N} \end{aligned} \tag{3.4.1}$$

Parameters n_1 and n_2 represent the number of proce ssors allocated to RA1 and RA2. Figure 3.7 shows the optimal solution for RA1. The plot of n_2^* is symmetrical to $\epsilon = 0.5$. The traffic of each cell is $\rho_1 = \epsilon\rho$ and $\rho_2 = (1-\epsilon)\,\rho$, where $\rho = 40$ Erlangs and $0 \leq \epsilon \leq 1$.

For $\theta = 0$ all processors will be employed for maximizing the sum of $U(n)$ (see (3.3.3)). The processors are distributed between the RAs depending on the slope of the Erlang-B function. For $0.1 \leq \epsilon \leq 0.3$ and $0.7 \leq \epsilon \leq 0.9$ more processors are assigned to the cell with higher traffic load. This is different for $0.3 \leq \epsilon \leq 0.7$, because assigning more processors to the cell with lower service demand decreases the overall blocking probability. For $\epsilon \leq 0.1$ or $\epsilon \geq 0.9$, the traffic of one or the other cell exceeds the corresponding RA capacity and the problem becomes unfeasible. The deployment of additional RA are necessary for such traffic distributions.

When $\theta = 1$, the number of processors is directly proportional to the traffic load, because the slope of the objective function is constant with n. For $0 < \theta < 1$, the resources are allocated as a function of the performance increment in relation to the amount of allocated resources. The number of allocated processors linearly increases with ϵ and $(1-\epsilon)$, respectively, until reaching the maximum number of processors $n_{\max}$ that still meets the session establishment delay constraint.

3.5 Simulation Results

We simulate a non-homogeneous traffic demand, where the user session initiation and termination are modeled as a Poisson arrival and departure process. The user arrival rate is 4 times the departure rate, simulating an unstable situation for better analyzing the performance of the different strategies. The *adaptive strategy* solves equation (3.4.1) with 16 RAs instead of 2. The *static strategy* does not track the traffic load distribution, but rather assigns 16 of the 256 processors to each RA. The second variant of the static strategy randomly distributes the 256 processes among the RAs.

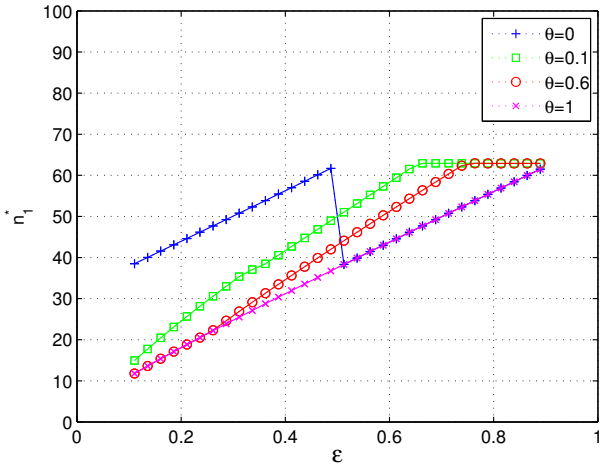

Figure 3.7: Optimal distribution of processors to two g-mapping RAs as a function of ϵ for different θ ($t_s^{\max} = 10$ ms $p_b^{\max} = 5\%$, and $\Phi(n) = n/1.5$).

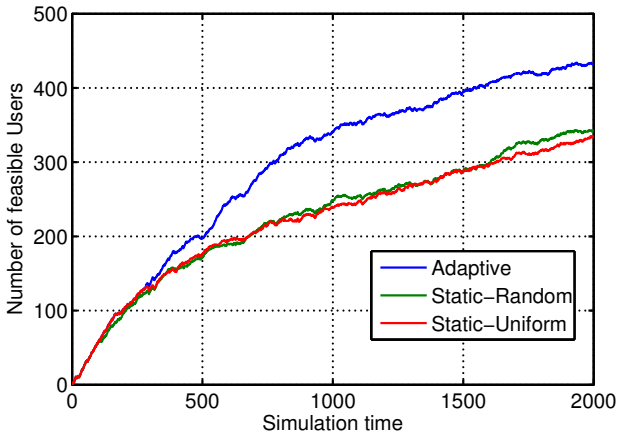

Figure 3.8: Accepted user sessions over simulation time for 256 processors, 16 RAs and a heterogeneous traffic distribution.

Each processor has a capacity of 12 giga-operations per second (GOPS). The waveforms offer 64, 128, 384, and 1024 kbps data rate services, which are solicited with a probability of 0.5, 0.2, 0.2, and 0.1, respectively. The four waveform models are those from [10], requiring between 50 % and 75 % of a processor's capacity ($k < 1$). The users follow a two-dimensional Gaussian distribution, centered and with a variance of 0.25 relative to the service area.

The blocking probability constraint is dropped in order to enable a fair evaluation between the three strategies. The optimal strategy then maximizes the number of served users ($\theta = 0$). The session initialization constraint is set to 50 ms and the average session duration 40 s.

Figure 3.8 shows the accumulated number of accepted user session requests over time. The adaptive strategy assigns processors to RAs according to the traffic demand and optimization parameters, resulting in 2-33 processors assigned to each RA. It accepts considerably more users than both variants of the static strategy. This indicates the performance improvement of adapting the processor assignment to the actual user distribution [61, 62, 63].

3.6 Summary

This chapter has addressed the SDR cloud CRM problem. Defining the concept of a RA that manages a subset of computing resources facilitates separating the signal processing algorithms design from the infrastructure and enables using resources on a pay-per-use basis.

Based on the call setup delay and the blocking probability constraints, we have defined the RA processor allocation problem as a constrained convex optimization problem. The feasibility region provides the maximum traffic capacity that a single RA can manage. The results have shown that modern cellular communications standards, such as LTE and LTE-Advanced, considerably limit the RA capacity. Assuming that two processors or more are required to process a transceiver processing chain in real time, less than 50 Erlangs of traffic can be handled by a single RA employing a greedy mapping algorithm. A distributed resource management is therefore necessary.

The data centre processors need to be distributed among several RAs subject to the call setup delay and blocking probability constraints. The simulation results moreover indicate that the number of accepted users is severely degraded if the processors distribution is not adapted to the traffic distribution. Our solution is optimal, but does not scale well with the problem size. More precisely, the complexity of problem (3.4.1) grows exponentially with the number of RAs.

Chapter 4

Radio Virtualization Model

4.1 Introduction

SDR cloud computing resources can be virtualized, as discussed in Chapter 3. Processing time, for instance, is divided in slices. Time slices are allocated to users by the CRM. Processors are shared between different clients, creating the illusion of exclusive access. We say that users' access to the resources is orthogonal, because there is no interference.

A similar virtualization is applicable to radio resources. Time, frequency and space dimensions can be divided so that different users share a single physical resource, creating the illusion of exclusive access (i.e. orthogonal access). The network operator buys transmission/reception time in a physical location (antennas) and frequency band (RF channels). By employing multiplexation techniques, e.g. TDMA, FDMA, SDMA, physical radio resources become virtual radio resources. Time, frequency and space dimensions are divided in time slots, sub-carriers and spatial streams and assigned to users. The virtualization of radio resources facilitates the radio resource management problem, in particular the channel and power allocation problems.

Radio resources show different (often random) performance depending on the location of the user with respect to the base station (shadowing, propagation loss) and the statistical nature of the wireless channel (multipath). The performance of a radio resource unit (time slot, sub-carrier or spatial stream) is characterized by a real-valued non-negative coefficient named *channel gain*.

The aim of this chapter is identifying, reviewing and providing coherent formulation for the different transmission and processing techniques capable of virtualizing wireless resources. The remaining of the chapter is organized as follows. Section 4.2 review common multiplexation methods for the time, frequency and space dimensions. Section 4.3 introduces the general parallel channel model. Section 4.4 studies the statistical properties of the channel gains.

Throughout this chapter, we will use the following notations. Lowercase letters, boldface lowercase letters and boldface uppercase letters are used for scalars, vectors and matrices, respectively. $()^H$, $()^T$ respresent the conjugate transpose and the transpose operations, respectively. $\Pr(\mathcal{A})$ denotes the probability of occurrence of the event $\mathcal{A}$, $\mathbb{E}\{\}$ denotes the expectation operator, $\mathbf{I}_{N\times M}$ is the identity matrix of size $N \times M$, $\text{diag}(\mathbf{a})$ is an $n \times n$ matrix whose diagonal contains the vector $\mathbf{a}$ and the rest of elements are zero, $\mathcal{CN}(.,.)$ denotes a circular symmetric complex Gaussian vector and $[\mathbf{A}]_{i,j}$ is the (i,j)th element of the matrix $\mathbf{A}$.

Throughout this chapter and for mathematical clarity, it is assumed that the additive white Gaussian noise (AWGN) has unit power.

4.2 Resource Multiplexing

In this section, the *downlink* channel is addressed. One or more BS try to send information to different users. In some cases, however, the same discussion is valid for the *uplink* or peer-to-peer systems as well. The transmitter decomposes the channel in n parallel and independent channels. Each channel is characterized by the channel gain γ_i, a real non-negative coefficient. Based on the channel gains γ_i, $i = 1 \dots n$, the transmitter decides a power and rate allocation. This allocation optimizes a given policy or metric, e.g.: total throughput, power, fairness or efficiency. To be able to carry out this decision, it is essential that the transmitter knows the channel gain γ_i. These coefficients are a function of the channel statistics. In time-division duplexing (TDD) mode, if the channel can be assumed stationary and reciprocal, the channel coefficients can be estimated during a training phase. In frequency-division duplexing (FDD) mode, a feedback channel is needed to transmit a quantified version of the channel measurements.

4.2.1 Time

Time-division multiplexing is a simple virtualization scheme for transmitting independent signals over a common signal path. It consists on dividing the time dimension in slices, of equal length or not. If the slices are assigned to different users, we use the term TDMA. Consider the discrete baseband representation of a frequency-flat channel:

$$y[m] = x[m]h[m] + z[m],\ m \geq 0 \tag{4.2.1}$$

where $x[m]$ and $y[m]$ are the transmitted and received signals, $h[m]$ is the time-varying complex channel and $z[m]$ is AWGN. Assume we divide the time dimension in slots of equal size M symbols. The received signal at the kth slot is

$$y[m] = x[m]h[m] + z[m],\ m = (k-1)M \dots kM - 1, k = 1, 2, \dots \tag{4.2.2}$$

If we consider a finite number of slots K, we can reformulate (4.2.2) in vector notation:

$$\mathbf{y}[m] = \mathbf{H}[m]\mathbf{x}[m] + \mathbf{z}[m],\ m = 0 \dots M \tag{4.2.3}$$

where the kth row of the transmitted and received vectors is the symbol transmitted and received at time m of the slot. That is:

$$\mathbf{x}[m] = [x[m], x[m+M], x[m+2M], \dots, x[m+(K-1)M]]^T. \tag{4.2.4}$$

$\mathbf{y}[m]$ and $\mathbf{z}[m]$ are defined analogously. $\mathbf{H}$ is the *channel matrix* and is a matrix whose elements are zero except in its diagonal, i.e.:

$$\mathbf{H}[m] = \begin{pmatrix} h[m] & 0 & 0 & 0 \\ 0 & h[m+M] & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & h[m+(K-1)M] \end{pmatrix}. \tag{4.2.5}$$

For mathematical clarity, we will drop the time index $[m]$ in the remaining of the chapter. Therefore, the matrix formulation of the time-sliced channel becomes

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z}. \tag{4.2.6}$$

If each time slice is allocated to a different user, the kth row of $\mathbf{y}$ is the signal received by the kth user. The fact that the channel matrix is diagonal, $\mathbf{H} = \text{diag}\,(h_1, h_2, \dots, h_K)$

indicates that each user accesses the channel orthogonally. We can say that the resource *time* is virtualized because time is shared between different users without inter-user interference. This communications scheme is called TDMA.

There are K virtual or parallel channels, with channel-to-noise ratio (CNR)

$$\gamma_k = |h_k|^2. \tag{4.2.7}$$

It is often more convenient to assume that the transmitted signal has unit power, $\mathbb{E}\{|x[m]|^2\} = 1$ and introduce a new parameter, $\sqrt{p_k}$ or the power transmitted in the kth slot. Since the AWGN has unit power, the signal-to-noise ratio (SNR) is equal to

$$\text{SNR} = p_k \gamma_k. \tag{4.2.8}$$

Define the diagonal matrix $\mathbf{P} = \text{diag}\,(p_1, p_2, \ldots, p_K)$, then the input-output model of a TDMA system becomes

$$\mathbf{y} = \mathbf{H}\mathbf{P}^{1/2}\mathbf{x} + \mathbf{z}. \tag{4.2.9}$$

4.2.2 Frequency

Frequency-division multiplexing is a technique by which the total bandwidth is divided in a series of non-overlapping frequency sub-bands, each of which is used to carry a separate signal. We denote frequency-division multiplexing access (FDMA) to a system where multiple users access a BS using different frequency bands.

Consider a wideband frequency selective channel which can be modeled as an L-tap finite impulse response (FIR) filter

$$y[m] = \sum_{l=0}^{L-1} h_l[m]x[m-l] + z[m], \tag{4.2.10}$$

where $h_l[m]$ is the l-th tap gain at time m. Let $H(f)$, $f \in (-B/2, B/2)$ be the channel frequency response. The wideband channel can be decomposed in N_c frequency sub-bands, with constant channel gain.

A practical and more illustrative application of frequency multiplexing is orthogonal frequency-division multiplexing (OFDM). Let a vector of N_c symbols $\mathbf{s} = [s[0], s[1], ..., s[N_c - 1]]^T$ with unit power be scaled by the transmission power matrix $\mathbf{P}$,

$$\mathbf{d} = \mathbf{P}^{1/2}\mathbf{s}, \tag{4.2.11}$$

and prefixed with the last L symbols:

$$\mathbf{x} = [d[N_c - L - 1], \ldots, d[0], \ldots, d[N_c - 1]]^T. \tag{4.2.12}$$

The inter-symbol interference (ISI) extends over the first $L-1$ symbols and the receiver ignores it by considering only the output over the time interval $m \in [L, N_c + L - 1]$. For simplicity, we will assume that for each l, the l-th tap is time-invariant. The channel output can be formulated as

$$y[m] = \sum_{l=0}^{L-1} [h_l d[m - L - l] \ mod \ N_c] + z[m], \tag{4.2.13}$$

where *mod* indicates modulo operation. Denoting

$$\begin{aligned}\mathbf{h} &= [h_0, h_1, \ldots, h_{L-1}, \ldots, 0, \ldots, 0]^T \\ \mathbf{y} &= [y[L], y[L+1], \ldots, y[N_c + L - 1]]^T,\end{aligned} \tag{4.2.14}$$

then (4.2.13) can be written as:

$$\mathbf{y} = \mathbf{h} \otimes \mathbf{d} + \mathbf{z} \tag{4.2.15}$$

where the vector $\mathbf{y}$ has length N_c and $\otimes$ is the circular convolution. Alternatively, we can represent (4.2.15) in matrix notation:

$$\mathbf{y} = \mathbf{H}\mathbf{d} + \mathbf{z} \tag{4.2.16}$$

with

$$\mathbf{H} = \begin{pmatrix} h_0 & 0 & \cdot & 0 & h_{L-1} & h_{L-2} & \cdot & h_1 \\ h_1 & h_0 & 0 & \cdot & 0 & h_{L-1} & \cdot & h_2 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & 0 & h_{L-1} & h_{L-2} & \cdot & h_1 & h_2 \end{pmatrix} \tag{4.2.17}$$

a circulant matrix.

The basic idea of OFDM is to turn the channel matrix into a circulant matrix via the addition of a cyclic prefix to the transmitted sequence. A circulant matrix has the property that its left and right singular vector matrices are respectively discrete Fourier transform (DFT) and inverse discrete Fourier transform (IDFT) matrices. Then, after singular-value decomposition (SVD) of $\mathbf{H}$:

$$\mathbf{H} = \mathbf{U}^{-1}\mathbf{\Lambda}\mathbf{U} \tag{4.2.18}$$

with $\mathbf{\Lambda}$ a diagonal matrix with entries the DFT coefficients of the channel impulse response $\mathbf{h}$, or a discretization of the channel frequency response $H(f)$. The matrix $\mathbf{U}$ is unitary with

$$[\mathbf{U}]_{k,n} = \frac{1}{\sqrt{N_c}} e^{\frac{-j2\pi kn}{N_c}}. \tag{4.2.19}$$

The multiplication of the transmitted symbols by an IDFT matrix at the transmitter and by a DFT matrix at the receiver transforms the channel circulant matrix $\mathbf{H}$ into a diagonal matrix $\mathbf{\Lambda}$, whose elements are the singular values of the circulant matrix, that is:

$$\hat{\mathbf{y}} = \hat{\mathbf{\Lambda}}\hat{\mathbf{x}} + \hat{\mathbf{z}} \tag{4.2.20}$$

where

$$\begin{aligned}\hat{\mathbf{x}} &\triangleq \mathbf{U}^{-1}\mathbf{x} \\ \hat{\mathbf{y}} &\triangleq \mathbf{U}\mathbf{y} \\ \hat{\mathbf{z}} &\triangleq \mathbf{U}\mathbf{z}.\end{aligned} \tag{4.2.21}$$

$\mathbf{U}$ is unitary, thus the distribution and energy of $\hat{\mathbf{z}}$ are preserved.

The original frequency-selective channel becomes a set of parallel flat-fading channels. A diagonal channel matrix $\mathbf{\Lambda} = \text{diag}\,(\lambda_1, \ldots, \lambda_{N_c})$ indicates that transmission can be done independently on each channel. Each channel admits the following base-band frequency-flat representation:

$$y_k = \sqrt{p_k}\lambda_k x_k + z_k,\ k = 1 \ldots N_c \tag{4.2.22}$$

And the equivalent channel gain is

$$\gamma_k = \lambda_k^2 \tag{4.2.23}$$

with λ_k the kth coefficient of the DFT of the channel impulse response, or the kth singular value of the circulant matrix $\mathbf{H}$.

Each channel (sub-carrier) can be assigned to a different user and we can say that the resource *bandwidth* is virtualized. This communications scheme is called OFDMA.

4.2.3 Space

Spatial multiplexing is a technique to transmit independent and separately encoded data signals, called *streams*, from each of the multiple transmit antennas in a MIMO wireless system. We say that a system uses SDMA if streams are used by different users. In order to facilitate the multi-user discussion, we briefly introduce the single-user MIMO operation.

Single-User MIMO

A MIMO channel models the channel a receiver with multiple antennas observes from a transmitter with multiple antennas. Spatial multiplexing requires accurate CSI at the receiver. Transmitter CSI facilitates the receiver implementation, though.

Assume a frequency-flat time-invariant channel with n_t transmit and n_r receive antennas, described by a deterministic channel matrix $\mathbf{H} \in \mathbb{C}^{n_r \times n_t}$, which is known to both the transmitter and receiver. The channel model is

$$\mathbf{y} = \mathbf{H}\mathbf{P}^{1/2}\mathbf{x} + \mathbf{z}, \tag{4.2.24}$$

where $\mathbf{x} \in \mathbb{C}^{n_t}$, $\mathbb{E}\{\mathbf{x}\mathbf{x}^H\} = 1$, $\mathbf{P} = \mathrm{diag}\,(p_1, \ldots, p_{n_t})$, $\mathbf{y} \in \mathbb{C}^{n_r}$ and $\mathbf{z} \sim \mathcal{CN}\,(0, \mathbf{I}_{n_r \times 1})$ denote the transmitted signal, received signal and white Gaussian noise, respectively. The (i,j)th position of the channel matrix $\mathbf{H}$ is the channel gain from transmit antenna i to receive antenna j. In general, the channel matrix $\mathbf{H}$ is not diagonal. Thus, the signal received at the j receive antenna is a linear combination of the signals transmitted by the n_t transmit antennas.

Let

$$\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^H \tag{4.2.25}$$

denote the SVD of the matrix $\mathbf{H}$. $\mathbf{U} \in \mathbb{C}^{n_r \times n_r}$ and $\mathbf{V} \in \mathbb{C}^{n_t \times n_t}$ are unitary matrices. $\mathbf{\Lambda} \in \mathbb{R}^{n_r \times n_t}$ is a diagonal matrix whose elements λ_i are non-negative real numbers. $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n_{\min}}$ are the ordered singular values of $\mathbf{H}$ and $n_{\min} \triangleq \min(n_t, n_r)$ is the rank of the channel matrix. If we define

$$\begin{aligned} \hat{\mathbf{x}} &\triangleq \mathbf{V}\mathbf{x}, \\ \hat{\mathbf{y}} &\triangleq \mathbf{U}^H\mathbf{y}, \\ \hat{\mathbf{z}} &\triangleq \mathbf{U}^H\mathbf{z}, \end{aligned} \tag{4.2.26}$$

we can rewrite (4.2.24) as

$$\hat{\mathbf{y}} = \mathbf{\Lambda}\mathbf{P}^{1/2}\hat{\mathbf{x}} + \hat{\mathbf{z}}, \tag{4.2.27}$$

where it must be noted that the unitary transformations preserve the energy of $\mathbf{x}$ and the distribution of $\mathbf{z}$. Thus, after SVD decomposition, the original MIMO channel can be decomposed in $n_{\min}$ parallel channels, defined by the diagonal matrix $\mathbf{\Lambda}$, which contains the singular values of the original channel matrix $\mathbf{H}$. The channel model is now equivalent to the

models discussed in the preceding sections. The unitary transformation at the transmitter and receiver virtualizes the radio resource *space* into a set of $n_{\min}$ parallel (virtual) channels. The equivalent gain of the kth channel is

$$\gamma_k = \lambda_k^2. \tag{4.2.28}$$

If the power matrix $\mathbf{P}$ is chosen appropriately [64] (i.e. water-filling), the SVD-MIMO scheme achieves capacity, which equals:

$$I = \sum_{k=1}^{n_{\min}} \log\left(1 + \lambda_k^2 p_k\right). \tag{4.2.29}$$

The SVD technique is difficult to implement in practice, because it requires the transmitter and receiver to know the matrix $\mathbf{H}$. The square of the singular values of $\mathbf{H}$ or the eigenvalues of $\mathbf{H}\mathbf{H}^H$ determine the capacity of an SDMA system.

Multi-User MIMO

Let a single base station with $n_t = M$ antennas send information to K users, each equipped with a single antenna. This is also known as the *broadcast MISO* channel. If $K \leq M$, it is possible to exploit the total number of degrees of freedom and transmit independent information to all users [65]. We assume that the base station has perfect CSI of all users, that is, the BS knows the matrix

$$\mathbf{H} = \left[\mathbf{h}_1^T \cdots \mathbf{h}_K^T\right]^T \tag{4.2.30}$$

of size $K \times M$. $\mathbf{h}_k$ is a $1 \times M$ vector of channel gain coefficients between M BS antennas and the user's antenna.

The broadcast channel can be seen as an $n_t = M$, $n_r = K$ MIMO system with the difference that the receivers are not able to cooperate. SVD-MIMO scheme can not be employed, because the kth user knows $\mathbf{h}_k$ but not $\mathbf{H}$.

The signal to each user can be separated at the BS by multiplying the transmitted vector by a *precoding vector*. Precoding is a generalization of beamforming to support multi-stream transmission in multi-antenna wireless communications. Precoding algorithms can be subdivided into linear and nonlinear precoding types. Nonlinear precoding is designed based on the concept of dirty paper coding and include Costa precoding [66], Tomlinson-Harashima precoding [67, 68] and the vector perturbation technique [69].

Linear precoding approaches usually achieve reasonable performance with much lower complexity. Linear precoding strategies include maximum-ratio transmission (MRT) [70], zero-forcing (ZF) precoding [71] and transmit Wiener precoding [71]. The optimal linear precoding does not have any closed-form expression, but it takes the form of a weighted minimum mean square error (MMSE) precoding for single-antenna receivers. MRT only maximizes the signal gain at the intended user and is close to optimal in noise-limited systems, where the inter-user interference is negligible compared to the noise. ZF precoding aims at nulling the inter-user interference, at the expense of losing some signal gain and achieves a performance close to capacity when the number of users is large or the system is interference-limited (i.e. the noise is weak compared to the interference) [72, 73]. A balance between MRT and ZF is obtained by the so-called regularized zero-forcing [74] (also known as transmit Wiener filtering [71].

Let s_k, p_k and $\mathbf{w}_k$ define the data symbol, the transmitted power and a $M \times 1$ precoding vector for user k, respectively. $\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_K]$ is the $M \times K$ precoding matrix, $\mathbf{s} =$

$[s_1 \cdots s_K]^T$ is the $1 \times K$ data vector and $\mathbf{P} = \text{diag}\,(p_1, \ldots, p_K)$ is the $K \times K$ diagonal power matrix. User k receives the following signal

$$y_k = \mathbf{h}_k \mathbf{w}_k \sqrt{p_k} s_k + \sum_{j \neq k} \mathbf{h}_k \mathbf{w}_j \sqrt{p_j} s_j + z_k \tag{4.2.31}$$

and in matrix notation

$$\mathbf{y} = \mathbf{HWP}^{1/2}\mathbf{s} + \mathbf{z}. \tag{4.2.32}$$

Therefore, the signal-to-interference-and-noise ratio (SINR) at user k becomes

$$\text{SINR}_k = \frac{|\mathbf{h}_k \mathbf{w}_k|^2 p_k}{1 + \sum\limits_{j \neq k} |\mathbf{h}_k \mathbf{w}_j|^2 p_j}. \tag{4.2.33}$$

The SINR at the kth user after MRT or regularized ZF precoding is a function of the power p_j allocated to users $j \neq k$. Hence, the resulting parallel channel model is not equivalent to other parallel channels reviewed throughout this chapter. The power allocation solution is iterative in nature, because the power allocated to one user determines the interference caused to the rest of the users. This fact makes the implementation more costly, specially in systems with many users as the SDR cloud.

ZF precoding makes sure that $\mathbf{h}_k \mathbf{w}_j = 0$ for $j \neq k$ and thus the interference can be removed from the SINR expression. The precoding matrix $\mathbf{W}$ is obtained using the Moore-Penrose (right) pseudoinverse [75] of the channel matrix $\mathbf{H}$:

$$\mathbf{H}^+ = \mathbf{H}^H \left(\mathbf{H}\mathbf{H}^H\right)^{-1}. \tag{4.2.34}$$

Then, let

$$\mathbf{W} = \mathbf{H}^+ \mathbf{\Gamma}^{1/2} \tag{4.2.35}$$

where the column-normalizing diagonal matrix $\mathbf{\Gamma}^{1/2}$ contains the reciprocal of the squared norm of the columns of $\mathbf{H}^+$ on the diagonal. $\mathbf{\Gamma} = \text{diag}\,(\gamma_1, \ldots, \gamma_k)$ and

$$\gamma_k = \frac{1}{\left[\left(\mathbf{H}\mathbf{H}^H\right)^{-1}\right]_{k,k}}. \tag{4.2.36}$$

The received signal is $\mathbf{y} = \mathbf{HWP}^{1/2}\mathbf{s} + \mathbf{z}$. Noticing that $\mathbf{HW} = \mathbf{\Gamma}^{1/2}$, we arrive to the parallel channel model

$$\mathbf{y} = \mathbf{\Gamma}^{1/2}\mathbf{P}^{1/2}\mathbf{s} + \mathbf{z}, \tag{4.2.37}$$

where the equivalent channel matrix $\mathbf{\Gamma}^{1/2}\mathbf{P}^{1/2}$ is a $K \times K$ diagonal matrix. Again, we can say that we have converted the resource *space* into a set of K virtual resources, each assigned to a different user.

In a multiuser MIMO system–each user is equipped with n_r antennas–ZF precoding transforms the MIMO broadcast channel into $n_r K$ parallel streams. A generalization of the ZF precoding for users equipped with multiple antennas is the block diagonalization precoding [76, 77]. With this precoding scheme, we can obtain K MIMO channels with n_r receive antennas and M transmit antennas, and obtain multiplexing as well as power gain. The equivalent channel is decomposed in K channels of n_r streams. The gain of the streams follow from the eigenvalues of the equivalent MIMO $n_r \times M$ matrix [78].

MD-MIMO

Consider a MIMO BS with B antennas. Let each antenna be connected to the processing unit by an optical fiber. We can separate the antennas to cover a large area and we obtain the model of the cell-less concept or the SDR cloud (i.e. MD-MIMO). The area covered by these BSs is denoted a *supercell* in [79, 80].

If each BS is equipped with a single antenna and there are K single-antenna users, the model is the same as the multi-user MISO model with $M = B$ transmitter antennas. In this case, however, we have that

$$\mathbf{H} = \mathbf{H}_w \odot \mathbf{\Sigma}, \tag{4.2.38}$$

where $\odot$ is the Hadamard product (i.e. element-wise product). $\mathbf{H}_w \in \mathcal{CN}(0, \mathbf{I}_{K\times B})$ is the multipath fading and $\mathbf{\Sigma}$ accounts for the signal path loss and shadowing. $[\mathbf{\Sigma}]_{ij}$ is the path loss and shadowing between user i and BS j. The matrix $\mathbf{\Sigma}$ is not diagonal because the antenna array can not be assumed to be compact with respect to the distance to the user.

Efficient precoding techniques for MIMO networks is an active area of research. ZF precoding is close to the optimum if $B, K \to \infty$ and $B \gg K$, although it requires the inversion of a very large matrix if B and K are large. Nevertheless, we can study the performance of any SDMA scheme using the capacity equation (4.2.29). The maximum rate the kth channel can support is given by

$$I_k = \log\left(1 + p_k\lambda_k\right) \tag{4.2.39}$$

where λ_k the kth eigenvalue of $\mathbf{H}\mathbf{H}^H$ and p_k is the corresponding allocated power.

4.3 The General Parallel Channel Model

In the last section, we observed that different communication systems can be modeled by the input-output relation

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z}, \tag{4.3.1}$$

where $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are the transmitted, received and AWGN vectors and $\mathbf{H}$ is the matrix of channel gain coefficients. Different transmission techniques can be combined and still be represented by similar models. For instance, a MIMO-OFDM system with N_c sub-carriers can be modeled as

$$\mathbf{y}_j = \mathbf{H}_j\mathbf{x}_j + \mathbf{z}_j,\ j = 1 \ldots N_c \tag{4.3.2}$$

where $\mathbf{H}_j$ is the channel gain in the jth sub-carrier.

We address the time, frequency and spatial multiplexing transmission schemes, which allows virtualizing the radio resources and facilitates the RRM problem. Under some assumptions, (4.3.1) is equivalent to a bank of n parallel and independent channels which are represented in matrix form as:

$$\mathbf{y} = \mathbf{D}\mathbf{P}^{1/2}\mathbf{s} + \mathbf{z}, \tag{4.3.3}$$

with $\mathbf{D}$ a diagonal matrix of complex elements and $\mathbf{P} = \mathrm{diag}\left(p_1, \ldots, p_n\right)$ a diagonal matrix of real elements. The kth user sees then the scalar channel

$$y_k = \sqrt{p_k}d_kx_k + z_k,\ k = 1 \ldots n. \tag{4.3.4}$$

The coefficient d_k is the complex coefficient of the equivalent kth channel. The equivalent channel gain is

$$\gamma_k = |d_k|^2. \tag{4.3.5}$$

Equation (4.3.4) is also valid for combinations of different communications systems, with the difference that γ_k has a different physical interpretation. For instance, a MIMO-OFDM system can be represented as $n = N_c \min(n_t, n_r)$ parallel channels after SVD decomposition on each sub-carrier.

4.4 Statistics of γ_k

The ith channel coefficient has a different physical interpretation depending on the communications system it represents:

1. In a time-division multiplexing (TDM) system, γ_k is a combination of the signal path loss, shadowing and multipath fading;
2. In an OFDM system, γ_k is the kth coefficient of the N_c-point DFT of the channel impulse response;
3. In a MIMO system...
   (a) γ_k is the kth eigenvalue of $\mathbf{H}\mathbf{H}^H$ in several models, e.g.: in a SVD-MIMO, in multiuser MIMO systems with block diagonalization precoding and, in general, determines the maximum rate of an SDMA technique;
   (b) $\gamma_k = \left[\left[(\mathbf{H}\mathbf{H}^H)^{-1}\right]_{k,k}\right]^{-1}$ if ZF precoding is applied at the transmitter;

4.4.1 TDMA

The statistics of γ_k in TDMA depend on the channel propagation characteristics. The equivalent channel gain (Section 4.2.1) is

$$\gamma_k = |h_k|^2 \tag{4.4.1}$$

with h_k the channel coefficient between the kth user and the BS, which is assumed to be constant in a given time slot. The statistical behaviour of γ_k can be divided in three terms

$$\gamma_k = L(d) \cdot S \cdot M \tag{4.4.2}$$

1. the path loss $L(d)$, due to the power loss the signal experiences while travelling a distance d,
2. the *shadowing* S or slow fading, caused by obstacles affecting the signal propagation and
3. the *multipath fading* M or fast fading, due to the different paths of the signal being combined at the receiver.

There are many models to characterize the path loss $L(d)$, specialized for different environments (e.g. urban, rural, indoor, etc.). In Chapter 7 we will assume the log-distance path loss model, where:

$$L(d) = c_0 \left(\frac{d_0}{d}\right)^{\alpha}, \tag{4.4.3}$$

and α is the path loss exponent, typically between 2 and 4. c_0 is the gain at the reference distance d_0 (typically 1 m) and depends on the carrier frequency.

The shadowing is modelled by a log-normal distribution with probability density function (p.d.f.)

$$f_S(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\log x-\mu)^2}{2\sigma^2}},\ x \geq 0. \tag{4.4.4}$$

The multipath fading M depends on whether there is line of sight (LOS) between the transmitter and receiver or not. If there is no LOS, the real and imaginary parts of the complex channel are modelled as zero-mean Gaussian random variables. The fading amplitude or envelope of the received signal follows a Rayleigh distribution and the received signal power is exponentially distributed. Thus, the term M in (4.4.2) has p.d.f.:

$$f_{M,\text{ray}}(x,\mu) = \frac{1}{\mu} e^{-\frac{x}{\mu}},\ x \geq 0. \tag{4.4.5}$$

If one of the signal paths is much stronger than the others the channel is characterized by a Rician distribution. It is described by two parameters: K and μ. K is the ratio between the power in the direct path and the power in the scattered paths. μ is the total average power from all paths. The received signal amplitude (not the received power) is Rice distributed with pdf:

$$f_{\sqrt{M},\text{ric}}(x;K,\mu) = \frac{K+1}{\mu} \exp\left(-K - \frac{(K+1)x}{\mu}\right) I_0\left(2\sqrt{\frac{K(K+1)}{\mu}x}\right),\ x \geq 0, \tag{4.4.6}$$

where $I_0(\cdot)$ is the 0th order modified Bessel function of the first kind.

The Rice distribution is mathematically hard to manipulate, due to the modified Bessel function. A simpler model of a LOS channel is the Nakagami-m distribution, which is based on the Gamma distribution. The p.d.f. is:

$$f_{\sqrt{M},\text{nak}}(x;m,\mu) = \frac{1}{\Gamma(m)} \left(\frac{m}{\mu}\right)^m x^{m-1} \exp\left(\frac{-m}{\mu}x\right),\ x \geq 0 \tag{4.4.7}$$

where m is a parameter indicating the ratio of the LOS component to the total power and μ is the total average power.

The combination of Rayleigh multipath with log-normal shadowing is modeled by the p.d.f. [81]:

$$f_{S\cdot M}(x;\sigma) = \int_0^\infty f_{M,\text{ray}}(x|\mu = y) f_S(v;\sigma) dv,\ u \geq 0. \tag{4.4.8}$$

It is also possible to combine a Rice or Nakagami-m distribution with log-normal shadowing. The exact form of the p.d.f. can not be obtained for these combinations of log-normal and Rayleigh or Rice models, though. A more convenient distribution is the K distribution, extensively used for modelling diverse scattering phenomena such as troopospheric propagation of radio waves or radar clutter, among others [82]. It assumes a Gamma distribution with mean σ and shape parameter L, with σ being a random variable having another Gamma distribution, this time with mean μ and shape parameter v. The result is that the product of the shadowing and multipath fading $S\dot{M}$ has the following density function:

$$f_{S\cdot M}(x;v,L)=\frac{2}{x}\left(\frac{Lvx}{\mu}\right)^{\frac{L+1}{2}}\frac{1}{\Gamma(L)\Gamma(v)}K_{v-L}\left(2\sqrt{\frac{Lvx}{\mu}}\right),\ x\geq 0, \tag{4.4.9}$$

where K is the a modified Bessel function of the second kind. The average signal amplitude is $\mathbb{E}\{X\}=\mu$.

In summary, the exponential, log-normal and Gamma distributions characterize the random variations of the channel gain a user experiences in TDMA. The exponential distribution models NLOS fading, the log-normal distribution models shadowing and the Gamma distribution (or equivalent) models Rician fading and the combination of fading and shadowing.

4.4.2 OFDMA

The sub-carrier gain is the DFT response of the frequency-selective channel $H(f)$ which in general can have an arbitrary shape. Shadowing fading, however is time-selective rather than frequency-selective. Thus, the set of channels belonging to the same user have equal mean but are affected differently by multipath fading, which is frequency-selective.

If each channel belongs to a different user, each users sees an independent flat-fading channel to the BS and the statistics of γ_k are exactly the same as in TDMA case: each user is affected by independent path loss, shadowing and multipath fading.

Suppose all sub-carriers belong to the same user and let the channel impulse response of the multipath fading channel be modelled as a FIR filter with L taps $g[l]$, $l=0..L-1$. Assume that the envelope of the signal of the lth tap can be modelled as a Rayleigh random variable with power $\sigma_n^2[l]=\mathbb{E}\{|g[l]|^2\}$. Then, the channel frequency response is the DFT of the channel impulse response. $\sqrt{\hat{\gamma}_i}$ becomes:

$$\sqrt{\hat{\gamma}_i}=\sum_{l=0}^{L-1}g[l]e^{-j2\pi l(i-1)/N_c},\ i=1\ldots N_c. \tag{4.4.10}$$

Assuming that the cyclic prefix is larger than L and perfect timing and frequency synchronization are achieved at the receiver, it is shown in [83] that the distribution of envelope of the channel frequency response $\hat{\gamma}_i$ is

$$f(|x|)=\frac{|x|}{\sigma_G^2}e^{-\frac{|x|^2}{2\sigma_G^2}}, \tag{4.4.11}$$

where the variance is $\sigma_G^2\triangleq\sum_{i=0}^{L-1}\sigma_n^2[l]$. Equation (4.4.11) is the p.d.f. of a Rayleigh distribution. Hence, $\hat{\gamma}_i$, $i=1\ldots N_c$ is finally exponentially distributed with mean $\mu=\sqrt{\sum_{i=0}^{L-1}\sigma_n^2[l]}$.

It must be noted that not all the channel frequency coefficients are independent. This observation is of special interest for the derivations that follow in Chapter 7. The correlation depends on the coherence bandwidth of the channel. The coherence bandwidth is inversely proportional to the multipath spread [65][Section 2.3]:

$$B_{\text{coh}}=\frac{1}{2T_d}=\frac{B}{2L} \tag{4.4.12}$$

with B the signal bandwidth, L the channel impulse response length and T_d the channel delay spread. Since each sub-carrier is B/N_c wide, there are approximately

$$\frac{N_cB_{\text{coh}}}{B}=\frac{N_c}{2L} \tag{4.4.13}$$

adjacent sub-carriers whose channel coefficients are heavily correlated.

In OFDM, thus, we can expect the channel gain of the sub-carriers to be exponentially distributed in the NLOS fading case. In the general case with LOS fading and shadowing the distribution is unknown.

4.4.3 SDMA

In SDMA, the channel gains γ_k depend on the channel matrix $\mathbf{H}$ and processing technique. In this section we study the statistics of the eigenvalues of the matrix $\mathbf{H}\mathbf{H}^H$ and the equivalent channel gains after ZF precoding.

The matrix $\mathbf{H} \in \mathbb{C}^{n_r \times n_t}$ models a downlink system with n_r receive antennas and n_t transmitter antennas. For simplicity, we will assume that there is no LOS between the kth user and the corresponding BS, so that multipath fading is modeled by a zero-mean complex Gaussian process. Furthermore, we will neglect any transmit or receive antenna correlation effects.

Define the matrix $\mathbf{H}_w \sim \mathcal{CN}(0, \mathbf{I}_{n_r \times n_t})$, the diagonal matrix $\mathbf{T} \in \mathbb{R}^{n_r \times n_r}$ and the matrix $\mathbf{\Sigma} \in \mathbb{R}^{n_r \times n_r}$. The channel matrix $\mathbf{H}$ can be represented by one of the following models (among others):

- $\mathbf{H} = d \cdot \mathbf{H}_w$, with d a scalar models a single-user MIMO system where d accounts for path loss and shadowing;
- $\mathbf{H} = \mathbf{H}_w\mathbf{T}$ models a multi-user MISO downlink. If each user is equipped with a single antenna, the kth row of $\mathbf{H}$ is the channel vector it sees from the n_t BS transmit antennas. The kth element of the diagonal of $\mathbf{T}$ is the path loss and shadowing coefficient of the kth user, which affects the entire kth row of $\mathbf{H}_w$.
- $\mathbf{H} = \mathbf{H}_w \odot \mathbf{\Sigma}$ models a multi-user multi-cell distributed MIMO network (e.g. MD-MIMO or SDR cloud). Assuming each user and each BS has a single antenna, the ij element of the matrix $\mathbf{\Sigma}$ denotes the path loss and shadowing coefficient from the jth user to the ith BS in the downlink. The product $\odot$ is the element-wise matrix product, then, each coefficient of the matrix is affected differently.

Eigenvalues of $\mathbf{H}\mathbf{H}^H$

The eigenvalues of the complex hermitian matrix $\mathbf{H}\mathbf{H}^H$ are the most important characterization of the channel gains, because they determine the MIMO capacity and hence the maximum performance of a RRM algorithm.

They have been extensively studied in the literature for simple channel models [64, 84, 85, 86, 87, 88, 89, 90]. The joint p.d.f. of the ordered eigenvalues is derived in [64]. From the joint p.d.f. it is possible to extract the marginal cumulative density function (c.d.f.) and p.d.f. of the ordered or unordered eigenvalues, although the resulting expressions are mathematically hard to manipulate.

When the dimensions of the channel matrix are very large, Random Matrix Theory tools can be applied [91]. If the entries of $\mathbf{H}$, of size $n_r \times n_t$, are zero-mean independent and identically-distributed (i.i.d.) complex random variables with variance $1/n_t$, as $n_r, n_t \to \infty$ and $n_r/n_t = \beta$, the p.d.f. of the eigenvalues of $\mathbf{H}\mathbf{H}^H$ converges almost surely to a distribution described by the Marčenko and Pastur law [92]:

$$f(x) = \left(1 - \frac{1}{\beta}\right)^+ \delta(x) + \frac{\sqrt{(x-a)^+(b-x)^+}}{2\pi\beta x},\ a \le x \le b \tag{4.4.14}$$

where

$$a = (1 - \sqrt{\beta})^2,\ b = (1 + \sqrt{\beta})^2. \tag{4.4.15}$$

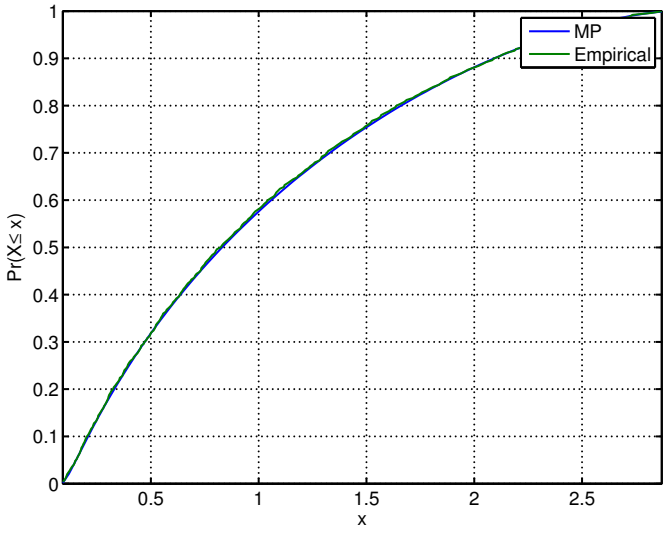

Figure 4.1: Marčenko and Pastur (MP) and empirical distributions of one realization of the eigenvalues of $\mathbf{H}\mathbf{H}^H$ when the entries of $\mathbf{H}$ are $\mathcal{CN}(0, 1/n_t)$.

The c.d.f. of (4.4.14) is derived in [93]:

$$F(x) = \frac{1}{2} + \frac{1}{2\pi\beta}\left[\sqrt{R(x)} - (1-\beta)\,S(x) - (1+\beta)\,T(x)\right] \tag{4.4.16}$$

where

$$\begin{aligned} R(x) &= -x^2 + 2\,(1+\beta)\,x - (1-\beta)^2 \\ S(x) &= \arcsin\left(\frac{(1+\beta)x - (1-\beta)^2}{2x\sqrt{\beta}}\right) \\ T(x) &= \arcsin\left(\frac{1+\beta-x}{2\sqrt{(\beta)}}\right). \end{aligned} \tag{4.4.17}$$

The p.d.f. of γ_k is (4.4.14) if the variance of the channel matrix entries is $1/n_t$. If $\mathbb{E}\{|h_{i,j}|^2\} = \nu$ for all i and j, then

$$\mathbb{E}\{\gamma_k\} = n_t \cdot \nu. \tag{4.4.18}$$

Figure 4.1 plots the empirical distribution of the eigenvalues of $\mathbf{H}\mathbf{H}^H$ when the entries of the random 250×500 matrix $\mathbf{H}$ are zero-mean complex Gaussian random variables. The figure shows that the Marčenko and Pastur distribution is a very good approximation of the true distribution.

The case $\mathbf{H} = \mathbf{H}_w\mathbf{T}$ with $\mathbf{T}$ a random or deterministic matrix is equivalent to the right-sided correlation model or the variance profile model [91]. The entries of the matrix $\mathbf{H}$ are independent but non-identically distributed random variables (e.g. they have different mean or variance). The distribution of the eigenvalues is unknown, albeit the mutual information was obtained in [94].

In the case $\mathbf{H} = \mathbf{H}_w \odot \mathbf{\Sigma}$ (MD-MIMO), the entries are also non-identically distributed and the distribution of the eigenvalues is unknown either. The mutual information is obtained in [95] and if the BSs follow a linear structure, the distribution of eigenvalues is given in [96].

Figure 4.2 plots the c.d.f. of the eigenvalues of $\mathbf{H}\mathbf{H}^H$ for an MD-MIMO network of $n_t = 500$ BS and $n_r = 250$ users uniformly distributed in the interval $(50, 1000)$ meters.

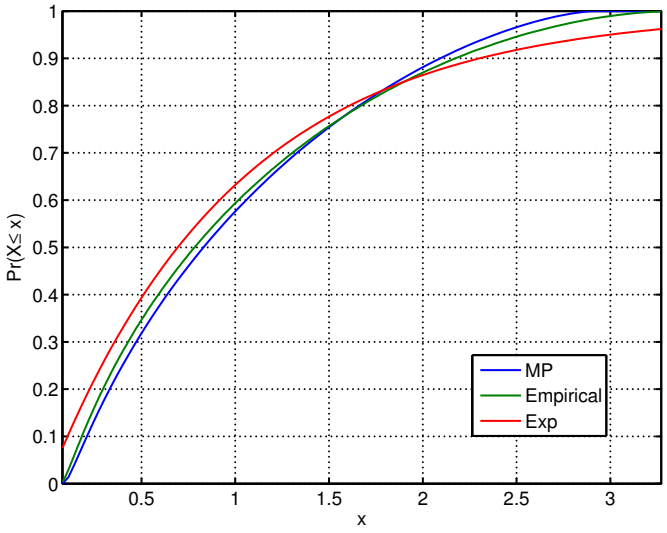

Figure 4.2: Exponential, Marčenko and Pastur (MP) and empirical distributions of one realization of the eigenvalues of $\mathbf{H}\mathbf{H}^H$ for $\mathbf{H} = \mathbf{H}_w \odot \mathbf{\Sigma}$ the channel of a MD-MIMO system. Users are uniformly distributed in the interval $(50, 1000)$ meters and the log-normal shadowing deviation is $\sigma = 6$ dB.

The figure shows that the Marčenko and Pastur distribution does not approximate the true distribution, because the entries of $\mathbf{H}$ are not i.i.d..

We rather propose to use the exponential distribution with average $\mu = n_t \cdot \nu$ to model the empirical distribution of eigenvalues. This distribution has the advantage of being mathematically simpler, which is important for the derivations in Chapter 7. Figure 4.2 shows that the exponential distribution provides a reasonably good approximation when $\mathbf{H} = \mathbf{H}_w \odot \mathbf{\Sigma}$ and $\beta = n_r/n_t < 0.9$.

Equivalent Channel Gain with ZF Precoding

When the precoding matrix is designed to cancel the interference (i.e. ZF), the equivalent kth channel gain becomes

$$\gamma_k = \frac{1}{\left[(\mathbf{H}\mathbf{H}^H)^{-1}\right]_{k,k}}. \tag{4.4.19}$$

If $\mathbf{H} \sim \mathcal{CN}(0, \mathbf{I}_{n_r \times n_t})$ and $n_t \geq n_r$, the equivalent gain γ_k is Chi-squared with $2(n_t - n_r + 1)$ degrees of freedom, with M and K the number of transmitter antennas and users, respectively [97]:

$$f(x) = \frac{1}{(n_t - n_r)!} x^{n_t - n_r} e^{-x},\ x \geq 0. \tag{4.4.20}$$

Figure 4.4 plots the c.d.f. of a Chi-squared distribution with $2(n_t - n_r + 1)$ degrees of freedom and the empirical distribution of γ_k according to (4.4.19) for a realization of a Rayleigh channel. The theoretical distribution well-represents the statistics of the channel gain. It is worth noting that the difference between the strongest and weakest channel is less than 2 dB. This difference increases as n_t, n_r, increase, though.

The p.d.f. of γ_k when ZF precoding is employed and $\mathbf{H}$ has arbitrary distributed entries is a topic of current research. If n_t is sufficiently large, some approximations can be obtained, though. Equation (4.4.19) can also be expressed as

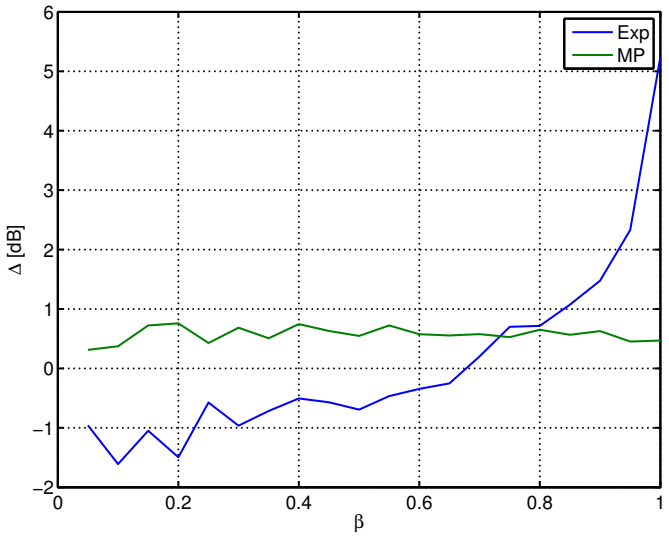

Figure 4.3: Plot of the error $\Delta = \frac{1}{n}\sum_{k=1}^{n}\Delta_k$ (in dB), where Δ_k is the ratio of the theoretical kth eigenvalue to the true kth eigenvalue of $\mathbf{H}$. The theoretical eigenvalues follow from the exponential and the Marčenko and Pastur distributions. The MD-MIMO scenario is the same as in Fig. 4.2.

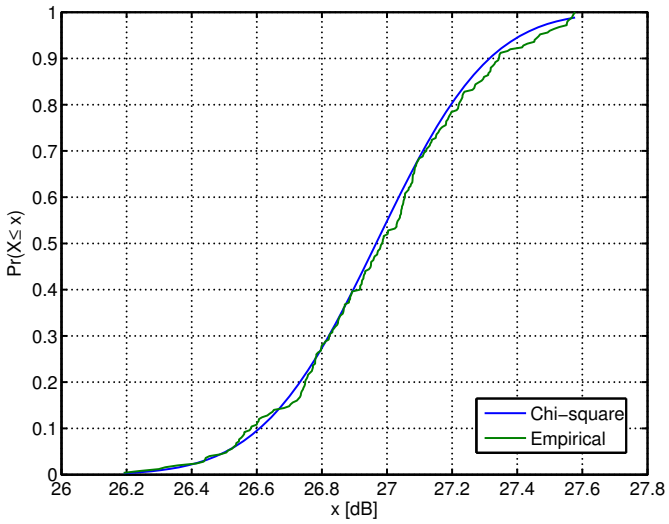

Figure 4.4: Chi-square and empirical distributions of one realization of the equivalent channel gain of the kth user, γ_k, after ZF precoding. The entries of the channel matrix $\mathbf{H}$ are $\mathcal{CN}(0,1)$, $K = 250$ and $B = 500$.

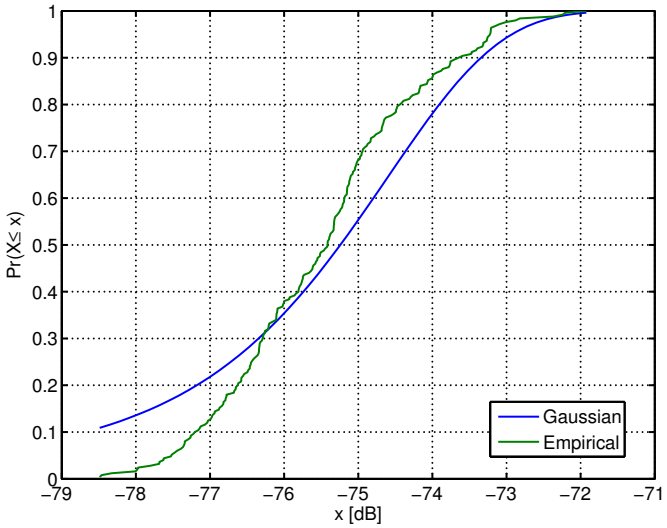

Figure 4.5: Gaussian and empirical distributions of one realization of the equivalent channel gain of the kth user, γ_k, after ZF precoding. The channel matrix is $\mathbf{H} = \mathbf{H}_w \odot \mathbf{\Sigma}$ with $\mathbf{H}_w$ the Rayleigh fading matrix and $\mathbf{\Sigma}$ the path loss and shadowing coefficient matrix. There are $n_r = 250$ users and $n_t = 500$ BS. The distance between a user and all BS is a uniform random variable in the interval $(50, 1000)$ m.

$$\gamma_k = \mathbf{s}_k^H \mathbf{s}_k = \sum_{i=1}^{n_t - n_r + 1} |s_{k,i}|^2 \tag{4.4.21}$$

where $\mathbf{s}_k$ is the projection of $\mathbf{h}_k$ in the orthogonal complement of the left-singular vectors of $\mathbf{H}$ [98]. If n_t is sufficiently large, (4.4.21) can be approximated as the sum of $n_t - n_r + 1$ i.i.d. random variables, which, by the Central Limit Theorem, converges almost surely to a normal distribution if the random variables have a well-defined mean and variance. The condition may not hold when the coefficients of the channel matrix are due to signal path loss, which follows a heavy-tailed distribution. This is observed in Fig. 4.5 were non-negligible differences between the Gaussian c.d.f. and the empirical distribution can be appreciated.

If $\mathbb{E}\{|h_{i,j}|^2\} = \nu$, for all i and j, the average channel gain after ZF precoding is

$$\mathbb{E}\{\gamma_k\} = (n_t - n_r + 1) \cdot \nu. \tag{4.4.22}$$

Our simulations show that an exponential distribution with average $(n_t - n_r + 1) \cdot \nu$ more accurately represents the distribution of γ_k if the entries of the channel matrix are heavy-tailed distributed. Figure 4.6 shows that the average difference between the true channel gains and those predicted by the theoretical distribution is lower for the exponential distribution than for the Gaussian distribution, specially for $K/B \to 1$.

Statistics of γ_k in a MD-MIMO system

Assume a MD-MIMO system with B base stations and K users, each equipped with one antenna. Let $\mathbf{H} = \mathbf{H}_w \odot \mathbf{\Sigma}$ model the channel matrix of an MD-MIMO system. If the distance between the mth BS and the kth user is an i.i.d. random variable and all users are affected by i.i.d. shadowing and multipath fading, we can assume that the entries of $\mathbf{\Sigma}$ are i.i.d., hence the entries of the matrix $\mathbf{H}$ are i.i.d. too.

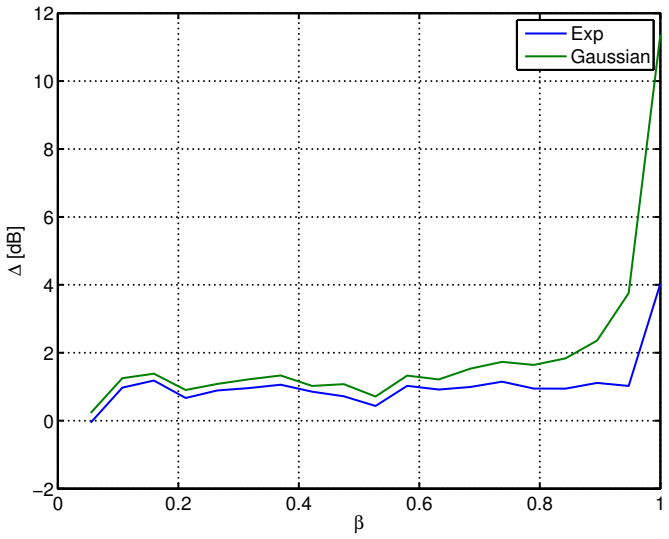

Figure 4.6: Plot of the error $\Delta = \frac{1}{n}\sum_{k=1}^{n}\Delta_k$ (in dB), where Δ_k is the ratio of the theoretical γ_k to the true γ_k after ZF precoding. The theoretical channel gains follow from the exponential and the Gaussian distributions. The MD-MIMO scenario is the same as in Fig. 4.2.

Let $\mathbb{E}\{[\mathbf{H}\mathbf{H}^H]_{i,j}\} = \nu$ for any i, j, the average of the equivalent channel gains after SVD decomposition of $\mathbf{H}$ is:

$$\mathbb{E}\{\gamma_k^{\mathrm{SVD}}\} = B\nu. \tag{4.4.23}$$

If the transmitter performs ZF precoding, the channel gains have expectation:

$$\mathbb{E}\{\gamma_k^{\mathrm{ZF}}\} = (B - K + 1)\,\nu. \tag{4.4.24}$$

The assumption $\mathbb{E}\{[\mathbf{H}\mathbf{H}^H]_{i,j}\} = \nu$ is valid as long as the multipath, shadowing and path loss random processes are i.i.d. for all user-BS pairs. If $h_w \sim \mathcal{CN}(0,1)$ and $\sqrt{g} \in \mathbb{R}$ describe the amplitude gain due to the Rayleigh multipath and shadowing/path loss effects, respectively:

$$\nu = \mathbb{E}\{|h_w|^2\} \cdot \mathbb{E}\{g\} = \mathbb{E}\{g\}. \tag{4.4.25}$$

The random variable g depends on the random variable d, which is the distance between a random user and a random BS. According to (4.4.3) and (4.4.4), it can be described as

$$g(d) = c_0 d_0^{-\alpha} 10^{\frac{Y}{10}}, \tag{4.4.26}$$

where α is the path loss exponent coefficient, c_0 is the reference gain at a distance of 1 m and Y is a zero-mean Gaussian random variable with variance σ^2 (log-normal shadowing). The path loss and shadowing effects are independent, thus

$$\mathbb{E}\{g(d)\} = \mathbb{E}\{L(d)\} \cdot \mathbb{E}\{S\} \tag{4.4.27}$$

According to (4.4.26):

$$\mathbb{E}\{S\} = \mathbb{E}\{10^{Y/10}\} = 10^{\log(10)\cdot(\sigma/10)^2/2}. \tag{4.4.28}$$

$\mathbb{E}\{L(d)\}$ can be found integrating the path loss over the p.d.f. of the random distance d. Considering d uniformly distributed in (d_0, d_1),

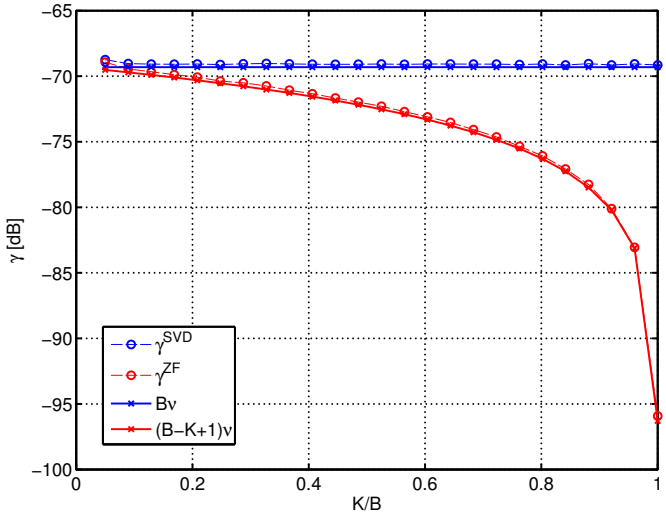

Figure 4.7: Plot of $\gamma = \mathbb{E}\{\gamma_k\}$ for ZF and SVD transmission schemes averaged over 20 realizations of $\mathbf{H}$ and from equations (4.4.24) and (4.4.23). Users are uniformly distributed in the interval $(50, 1000)$ meters.

$$\begin{aligned}\mathbb{E}\{L(d)\} &= c_0 \int_{d_0}^{d_1} \frac{1}{d_1 - d_0} x^{-\alpha} dx \\ &= c_0 \frac{d_0^{1-\alpha} - d_1^{1-\alpha}}{(\alpha - 1)(d_1 - d_0)} \\ &\approx c_0 \frac{d_0^{1-\alpha}}{\alpha - 1} \frac{1}{d_1},\end{aligned} \tag{4.4.29}$$

is the expected value of the path loss. The approximation assumes $d_1 \gg d_0$.

It is reasonable to consider d_0, α and σ as environment parameters and d_1 (the diameter of the super cell) a design parameter. Then, ν is a decreasing function of d_1:

$$\nu(d_1) = \Upsilon(d_0, \alpha, \sigma) \frac{1}{d_1}, \tag{4.4.30}$$

where

$$\Upsilon(d_0, \alpha, \sigma) = c_0 \cdot 10^{\log(10)\cdot(\sigma/10)^2/2} \frac{d_0^{1-\alpha}}{\alpha - 1} \tag{4.4.31}$$

We have simulated a MD-MIMO system with $B = 500$ antennas. Users are uniformly distributed. The path loss exponent is $\alpha = 3$. and the log-normal shadowing variance is $\sigma^2 = 6$ dB. Figure 4.7 plots the average channel gain as a function of the system load in terms of number of users K. The figure shows that while SVD scheme achieves full multiplexing and power gain, the ZF precoding loses the power gain as $K/B \to 1$. The results prove that the approximation assumed in (4.4.29) is accurate.

4.5 Summary

Throughout this chapter, we have reviewed different radio resource virtualization techniques. Time, frequency and space dimensions are divided in so-called time slots, sub-carriers and spatial streams that can be used by different users. The access is orthogonal, which implies

that no interference is caused between users. This facilitates the radio resource management problem, in particular the channel and power allocation problems.

Each independent channel is characterized by a non-negative real scalar value or channel gain. The channel gain is modeled statistically in terms of its p.d.f.. The p.d.f. of the channel gain depends on the multiplexing technique and channel model. Among others, some of the typical distributions are:

- The K-distribution or a combination of log-normal with exponential, Rice or Nakagami-m distributions models TDMA and OFDMA systems.
- The exponential distribution models Rayleigh fading, the OFDM sub-carrier gains when channel taps are Rayleigh distributed, and is a good approximation for the eigenvalues of $\mathbf{H}\mathbf{H}^H$ or the equivalent channel gains after ZF precoding if $\mathbf{H} = \mathbf{H}_w \odot \mathbf{\Sigma}$ is the channel matrix of a MD-MIMO system.
- The Marčenko and Pastur distribution characterizes the distribution of the eigenvalues of a large random matrix $\mathbf{H}\mathbf{H}^H$ under some conditions. The eigenvalues are used to determine the MIMO capacity.
- The Chi-squared or Gaussian distribution model the equivalent channel gains when ZF precoding is employed in a MISO network.

The parallel channel model characterizes a multi-channel system after frequency, time or space virtualization. We have studied the different statistics of the equivalent channel gains and concluded that the exponential distribution would be the most relevant distribution to model the channel gains in an SDR cloud.

Chapter 5

Resource Cost Model

5.1 Introduction

Centralizing the baseband processing increases the spectral, economical and energy efficiency of communication systems. The SDR Cloud goes one step further and adopts the concept of resource sharing and consolidation. A *virtually unlimited* amount of computational resources are offered to the client (network operator). During non-peak hours, processing workload is consolidated saving energy and costs. Furthermore, unused resources can be leased for other purposes (e.g. web services). Both techniques, resource sharing and consolidation, maximize the resource utilization and hence the benefits of the SDR cloud operator. Radio resources, on the other hand, are a *scarce* resource. Radio resources, by definition are meaningful in wireless communications only and can not be leased for other purposes.

The natural consequence of this market of virtualized resources is that there is a cost associated to their operation. The computing resources are unlimited, but the utilization cost is influenced by several factors, for instance:

- amortization costs,
- maintenance costs,
- auxiliary infrastructure costs: cooling, security, software licensing, etc. and
- market demand of 1 hour of computing time (e.g. for non-wireless purposes).

The spectrum usage rights are licensed to the network operator, who pays the infrastructure operator for using everything required to operate the spectrum. For example, let the cost of 1 second/antenna/channel be 1 currency unit (cu). The network operator would pay 100 cu/s for the operation of 10 RF channels at 10 different antennas or 100 cu/s for the operation of 1 RF channel in 100 antennas.

From the point of view of a network operator, it is clear that these costs must be considered during the resource and power allocation process. As will be shown in Chapter 6, maximizing the number of served users, or the total capacity is not the best strategy.

In this chapter, we introduce the concept of resource efficiency (Section 5.2) as the ratio of the delivered service to the consumed resources. In Section 5.3 we introduce a simple linear model for characterizing the operational and power costs involved in the SDR cloud. Section 5.4 describes the operational costs of the cost model.

5.2 Resource Efficiency

The efficiency of a system can be generally defined as the ratio between what the system provides and what it consumes. The business of a wireless carrier is delivering information to users. Therefore, a wireless system provides b *bits* and *consumes* u resource units. The system efficiency is then

$$\eta = \frac{b}{u} \tag{5.2.1}$$

in units bits per resource unit.

By changing the physical magnitude that u represents, different metrics can be optimized with a single model:

- Computing Efficiency: if $u = T_{\text{exec}}$ is the execution time of a baseband algorithm processing, the efficiency is given in transmitted bits per processor second time;
- Energy Efficiency: if $u = t \cdot (p_c + p_t)$ the processing and transmitted power consumption multiplied by the bit transmission time t, then η is the energy efficiency in bits/Joule,

among others.

Efficiency, in general, describes the extend to which the consumed resources u are well used for the intended task of delivering a service to the user. In other words, it measures the capability of the system to produce the desired outcome with the minimum amount of waste or unnecessary effort. The objective of the wireless carrier is to maximize this efficiency. That is, to maximize the delivered service while minimizing the consumed resources.

Consumed resources need to be described by a unique physical unit. This is hard to quantify in practice. Typically consumed resource units describe different things: number of man-months, kilo-watts-hour, amortization time, etc. On the other hand, all classes of consumed resources can be measured in the equivalent economical cost they represent, according to the market price. This cost is measured in *currency units* (cu). In the SDR cloud scenario and from the perspective of the network operator, this conversion is provided by the SDR cloud.

If the communications system is modeled as a bank of parallel channels, the global or total efficiency is the ratio of the total transmitted bits to the total consumed resources:

$$\eta_{\text{total}} = \frac{\sum_{i=1}^{n} b_i}{\sum_{i=1}^{n} u_i} \tag{5.2.2}$$

where b_i is the number of bits delivered to the ith channel and u_i are the resources consumed by the ith channel.

In Cloud computing, consumed resources are paid for their utilization time, in a *pay-per-use* fashion. Let

$$c = \frac{u}{t} \tag{5.2.3}$$

be the cost in cu/s of using u resource units for 1 second. The bit rate of a communications system is the number of transmitted information bits per second, or $r = b/t$. Hence, the resource efficiency (5.2.1) may also be expressed as

$$\eta = \frac{r}{c} \tag{5.2.4}$$

or

$$\eta = \frac{\sum_{i=1}^{n} r_i}{c} \tag{5.2.5}$$

if the communications system is modeled as a bank of parallel channels and r_i is the rate of the ith channel.

The objective of the network operator becomes determining the optimal bit rate (for each channel) that achieves the maximum efficiency or, in other words, delivers the maximum amount of bits consuming the minimum amount of resource units.

5.3 Costs Model

We propose dividing the costs in operational and power costs. Operational costs follow from the utilization of the SDR cloud infrastructure resources. The power cost is the cost of transmitting signals over the space. Then,

$$c = c_o^0 + c_o^r \sum_{i=1}^{n} r_i + c_o^n n + \sum_{i=1}^{n} c_p p_t(r_i) \text{ [cu/s]} \tag{5.3.1}$$

is the cost model, where:

- c_o^0 is the *constant cost*: it measures all costs of operation which are independent of the transmitted rate, power or number of allocated channels. It has units currency units per second of operation, cu/s.
- c_o^r is the *rate-dependent cost*: measures the cost of transmitting more or less bits per second. It has units cu/bit.
- c_o^n is the *n-dependent cost*: measures the cost of allocating degrees of freedom, or independent channels, e.g. bandwidth, time, space. The multi-channel baseband representation allows modelling the degrees of freedom as the number of independent channels n, regardless of the physical nature of the degree of freedom. The units of c_o^n are cu/channel.
- c_p is the *cost of transmission power*. It is measured in currency units per Joule, or cu/J. The product $c_p p_t$ has then units cu/s. It is assumed that c_p is equal for all channels. It includes any kind of power amplifier (PA) inefficiencies, which for simplicity are assumed to be independent of the transmit power. The function $p_t(r_i)$ is the power required to transmit the information to the receiver with arbitrary small error probability. This function is, in general, non-linear and, by definition, a non-decreasing function of the rate r_i.

More accurate models could also be defined. In [99, 100, 101], a model of a receiver's iterative decoder computational complexity is introduced. The model estimates the average number of iterations required to achieve certain bit error probability. The problem becomes mathematically intractable and algorithmic solutions become too complex to be of practical use. On the other hand, simple models offer interesting insights on how the optimal working points behave as a function of the problem parameters.

5.4 Operational Costs

In this section, we describe the operational costs c_o^0, c_o^r and c_o^n in equation 5.3.1.

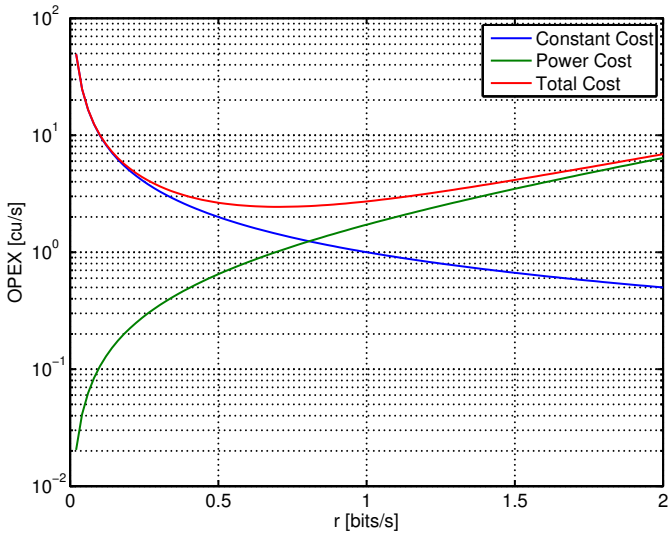

Figure 5.1: Operational costs (OPEX) per utilization time as a function of the transmission rate, for $c_o^0 = 1$, $c_o^r = 0$ $c_o^n = 0$ and $c_p = 1$.

5.4.1 Constant Cost c_o^0

Despite a small part of the processing costs are rate-dependent, the largest part are independent (e.g. data conversion, up and downsampling, etc.). This hypothesis is widely accepted in the literature [102, 103, 104] and has been proven experimentally for 802.11 baseband chips [105]. The results show that, in terms of power consumption, increasing the link speed does not imply extra circuit power consumption. For instance, receiving the fastest rate (405 Mbps) uses 1.6 W, while the slowest rate (40.5 Mbps) consumes 1.48 W.

Reference [106] reports a processing power consumption of 36-54 W for Global System for Mobile Communications (GSM) and 73-127 W for Universal Mobile Telecommunications System (UMTS) signals. This is a small fraction of the total base station power consumption, around 1500-4000 W. Most of the power consumption reported by this and other studies is due to cooling or power supply inefficiencies, among others [107]. Many other operational costs are independent of the rate, for instance: server cooling, security, software licensing, real estate, etc.

Figure 5.1 plots the operational costs per utilization time. For increasing transmission rate, the power costs is an increasing function. Since the costs of operation are independent of r, the cost per second is a decreasing function. The total costs have a minimum, where the maximum number of bits per currency unit is achieved (maximum efficiency).

5.4.2 Rate-dependent Cost c_o^r

The rate-dependent cost models those parts of the infrastructure which are dynamic enough to adapt to the bit rate so that the cost is proportional to the rate. For instance, GPP or DSP share a single processing unit by scheduling users sequentially, as discussed in Chapter 3. The cost is then proportional to the execution time of the processing task. FPGAs also have the capability to dynamically load multiple functionalities, each using a part of the total area.

If the resource can not be shared the operational costs may also be a function of their use. For instance, it is well known that the power dissipation in a chip can be modelled as the sum of a static and a dynamic terms [108],

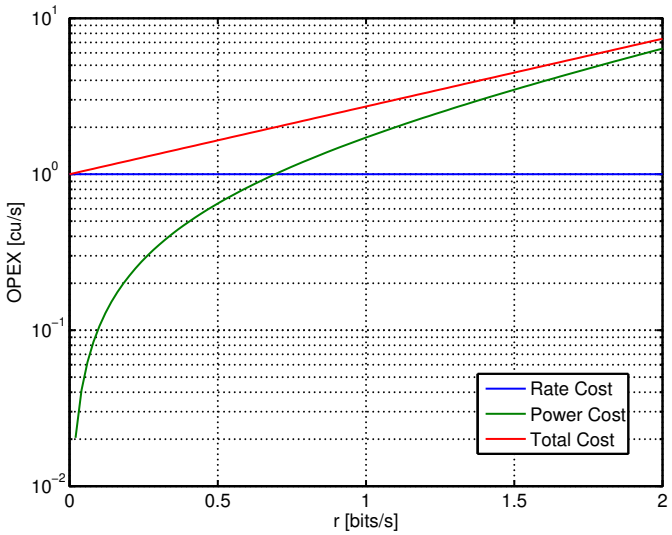

Figure 5.2: OPEX per utilization time as a function of the transmission rate, for $c_o^0 = 0$, $c_o^r = 1$, $c_o^n = 0$ and $c_p = 1$.

$$P_C = V_{dd} \cdot I_{\text{leak}} + a \cdot f \cdot C \cdot V_{dd}^2, \tag{5.4.1}$$

where a is the effective fraction of gates switching, f is the clock frequency and C is the circuit capacitance. Assuming that the frequency is dynamically scaled with the transmission rate, the power consumption can be modelled as a linear function of the transmission rate.

The rate costs increase with the transmitter bit rate. If we neglect the power costs, the operational costs in currency units per second is constant. On the other hand, if both rate-dependent and power costs are considered, the optimal strategy is always transmitting at the lowest rate possible as shown in Fig. 5.2.

5.4.3 n-dependent Cost c_o^n

The term c_o^n in the cost model describes the cost of using degrees of freedom or number of channels n. According to the channel model introduced in Chapter 4, the network operator divides the space, time and frequency resources in n independent channels, which are assigned to users. This model allows quantifying the cost of operating more or less radio resources while facilitates the resource allocation problem.

Consider, for instance, a network operator that rents 1 RF channel in 1 antenna. Operating these resources has a cost of 1 cu/s. The operator divides the spectrum in $n = 1000$ channels (e.g. using OFDM) and allocates channels to users. The cost per channel is $c_o^n = 10^{-3}$. According to (5.3.1), if we allocate to one or more users $n = 20$ channels, the total cost becomes $c_o^n n = 0.02$ cu/s.

Given a fixed total cost and number of partitions, this model enables assessing whether assigning more channels to a user (or a set of users) is beneficial or not. If n is increased, there are more frequency, time or space degrees of freedom. More degrees of freedom offer power gain and diversity, but also higher costs because the utilization of more RF channels, antennas or transmission time has to be paid to the SDR cloud.

If the network operator applies SDMA for serving K users using $B \geq K$ base stations, the number of independent channels is $n = K$. However, there are B signal flows being transmitted by B antennas, the cost per channel should be proportional to B instead of n. For this special case, we can redefine $c_o^n = \frac{B}{n} c_o^b$ where c_o^b is the cost of using one antenna,

in cu/s. Then, it is possible to study how increasing the number of transmitting antennas affects the operational costs.

5.5 Summary

In this chapter we introduced the concept of resource efficiency as the ratio of the delivered bits to the consumed resource units, expressed in equivalent currency units as a unified measure of resource operation costs. The consumed currency units are a function of the transmitted rate, power and number of channels, plus a constant term.

The scenario in which radio resource allocation is performed by the network operator was considered. Infrastructure resources (e.g. computing, antenna sites, RF frontends, etc.) are managed by the SDR cloud infrastructure operator, while the network operator chooses the best allocation of radio resource to users, according to a preferred policy. Different policies or strategies from different network operators create a competitive market in radio resource allocation algorithms and the relations between the infrastructure and radio resources in the SDR cloud.

Another conceivable scenario is a single entity managing the radio and infrastructure resources. Despite the infrastructure is not rented, there is always a cost associated to it, due to amortization and maintenance costs. Consequently, the cost model is valid as well.

The cost model (5.3.1) and the parallel channel model deals with virtual channels rather than physical ones, this allows us to seamlessly address frequency, space and time degrees of freedom. The benefits and costs of using these degrees of freedom can be easily measured and introduced in the radio resource management policies.

Chapter 6

Efficiency Maximization Power Allocation

6.1 Introduction

Throughout the preceding chapters, we discussed the benefits and costs of SDR clouds. The network operator acquires space, time and frequency resources for serving wireless users. Signals are processed in the SDR cloud computing infrastructure. The network operator pays the SDR cloud operator for the use of radio and computing resources, as discussed in Chapter 5. The infrastructure is not acquired once and then amortized. On the contrary, all CAPEX are converted to OPEX from the point of view of the network operator. The network operator objective becomes minimizing the OPEX per delivered bit.

The fact that radio and computing resources have an associated cost has important consequences in the adopted RRM algorithms and policies. Current RRM algorithms either maximize the user experience subject to radio operational constraints (e.g. power, spectrum, etc.) or minimize the radio operational costs subject to a minimum user experience constraints. New algorithms have to be designed that optimize the delivered bits (income) per infrastructure cost unit. This chapter introduces the *efficiency maximization power allocation* (EMPA) problem.

We assume the general parallel channel model with null inter-channel interference and n channels. The problem formulation is similar to the energy efficiency maximization problem, which has recently gained attention both in academia and industry. Section 6.2 reviews the literature in this topic. The problem is formulated as a constrained quasi-convex optimization problem in Section 6.3. In Section 6.4 a closed form solution for the case $n = 1$ is derived. A novel sub-optimal power adaptation for $n = 1$ with receiver CSI is derived in Section 6.5. The optimal solution for the general case $n \geq 1$ is derived in Section 6.6.

6.2 Related Work

The EMPA problem for application in SDR clouds, has never been tackled in the literature. The problem formulation is, however, equivalent to the energy efficiency maximization (EEM) problem. If the costs constants are set to:

$$\begin{aligned} c_o^0 &= P_c \\ c_o^r &= 0 \\ c_o^n &= 0 \\ c_p &= \epsilon, \end{aligned} \tag{6.2.1}$$

and P_c and ϵ are the rate-independent circuit power consumption and the power amplifier efficiency, the EMPA and EEM problem are equivalent. Therefore, the analysis carried out throughout this chapter can be applied to EEM problems as well.

As observed in [109], with the explosive growth of high-data-rate applications in wireless networks, energy efficiency in wireless communications has recently drawn increasing attention from the research community. Several international research projects dedicated to energy-efficient wireless communications are being carried out: Green Radio [110], EARTH [111], OPERA-Net [112] and eWIN [113], among others. Reference [114] provides a good survey on the last advances of these and other projects related to the EEM problem.

Early work on energy-efficient communications focused on optimizing the transmitted energy per bit [115, 116] while neglecting the processing power consumption. When transceivers became more complex, the significance of the processing power consumption increased. This was firstly addressed in [117] for uncoded and coded transmission over a single flat-fading channel. The authors consider a single channel and CSI available at the transmitter. The power and modulation level was then adapted according to the channel conditions for maximizing the energy efficiency. The problem solution is based on an numerical algorithm for obtaining the root of an equation, which yields to the final solution.

The authors of [118] introduced a closed form solution based on the Lambert-W function for the single channel case. This solution is adapted in Section 6.4 to the SDR cloud case. In [119], the optimal amplifier to processing power ratio for maximum energy efficiency was derived. The case with receiver CSI and no channel retransmissions was introduced in [120] for the Rayleigh channel. In [121] we extended the results to the Nakagami-m distribution with and without channel retransmissions.

The parallel channel model was considered in [102, 103, 104] for frequency-selective channels. Reference [102] proves the convexity of the optimization problem, providing the necessary and sufficient conditions for a unique and globally optimal solution. It proposes an iterative ascent algorithm for obtaining the optimal rate for each channel. The complexity of this solution prevents its applicability in adaptive systems with a large number of parallel channels. Reference [103] shows the feasibility of EEM link adaptation under rate and power constraints. The solution however lacks computing-efficient algorithms for practical implementation. Isheden et. al [104] propose an iterative algorithm based on the bisection method for finding the optimal waterlevel before solving the water-filling problem. The paper provides a thorough analysis of the optimality conditions. The parallel channel model and its iterative solution is discussed in Section 6.6.

6.3 Problem Formulation

The problem addressed in this chapter is determining the optimal power allocation on a bank of n parallel channels maximizing the resource efficiency defined as follows:

$$\eta = \frac{\sum_{i=1}^{n} r_i}{c}, \tag{6.3.1}$$

with r_i the transmitted rate in bits or nats per second of the ith channel and c are the total consumed resources in currency units per second.

Without loss of generality, the downlink is considered: the network operator acquires the rights of operation on a set of RF channels in a set of geographically distributed antennas during certain time. The operation applies one (or more) of the multiplexing techniques described in Chapter 4, dividing the radio resources time, space and frequency in a bank of n parallel channels.

The n channels are allocated to users following an arbitrary policy, the details of which are out of the scope of this thesis. It is worth mentioning that, assuming the cell-less architecture enabled by the SDR cloud, all users can employ the same frequency at the same time, provided that the number of users is lower than the number of antennas. Hence, the channel allocation problem is less important, in the cell-less architecture, than the power allocation problem.

The parallel channel model is described by the input-output matrix equation

$$\mathbf{y} = \mathbf{D}\mathbf{P}^{1/2}\mathbf{x} + \mathbf{z}, \tag{6.3.2}$$

where $\mathbf{x}$ and $\mathbf{y}$ are $n \times 1$ vectors containing the transmitted and received vectors, respectively, $\mathbf{D}$ and $\mathbf{P}$ are $n \times n$ diagonal matrices containing the complex channel coefficients and the transmitted powers, respectively. The $n \times 1$ vector $\mathbf{z}$ is AWGN complex noise.

The maximum achievable rate of the ith channel, according to model (6.3.2) is

$$r_i = \log\left(1 + p_i\gamma_i\right), \text{ [nats/s/Hz]} \tag{6.3.3}$$

The logarithm is in natural base and p_i is the transmitted power and the ith element of the diagonal of $\mathbf{P}$. The equivalent CNR is

$$\gamma_i = \frac{|d_i|^2}{\Gamma(P_b)\sigma^2} \tag{6.3.4}$$

where d_i is the ith element of the diagonal of $\mathbf{D}$ and σ^2 is the noise power. The function $\Gamma(P_b)$ is often denoted *gap to capacity* and is a measure of loss of a modulation/coding technique with respect to the theoretically optimum performance (i.e. channel capacity), for a given bit error rate P_b.

A multi-user multi-cellular power allocation problem should consider the following constraints:

- Per user minimum rate constraints, which are applied to the sum of the rates supported by the channels belonging to a user;
- per BS maximum power constraint, applied to the sum of the power allocated to each BS antenna; and
- per antenna maximum power constraint, applied to the power allocated to each BS antenna.

These constraints are introduced in the problem for two reasons. The first reason is a practical one: a power amplifier can not transmit infinite power, certain user applications have latency constraints, which imply a minimum rate constraint, and so forth. The second reason is for comparison purposes: traditional sum-rate maximization and sum-power minimization algorithms *distribute* a finite amount of power among a bank of parallel channels. In order to compare different algorithms, we need to distribute the same amount of power.

The aim of the EMPA problem is not distributing a finite power, but rather determining the optimal power allocation maximizing the efficiency. Thus, the constraints are not necessary for comparison purposes. They are interesting from a practical point of view, though. Mathematically speaking, the cost function of the EM problem is not monotonic, while the rate maximization (RM) and margin maximization (MM) problems cost functions are monotonic and thus they require the constraints. Nevertheless, throughout this chapter we will consider some basic constraints for the sake of illustration:

- Minimum total rate: $\sum_{i=1}^{n} r_i \geq R_{\min}$;

- Maximum total transmission power: $\sum_{i=1}^{n} p_i \leq P_{\max}$ and
- Maximum power per channel: $p_i \leq P_{\text{peak}}$, $i = 1 \ldots n$.

The total costs c are modeled as a sum of an offset cost c_o^0, a rate-dependent cost c_o^r, a cost dependent on the number of channels c_o^n and a power-dependent cost c_p (see Chapter 5):

$$c = c_o^0 + c_o^r \sum_{i=1}^{n} r_i + c_o^n n + c_p \sum_{i=1}^{n} p_i \text{ [cu/s]}. \tag{6.3.5}$$

The EMPA problem, thus, maximizes (6.3.1), given (6.3.5), the rate per channel (6.3.3) and the system constraints. The solution follows from the optimization problem:

$$\begin{aligned} \max_{p_i} \quad & \frac{\sum_{i=1}^{n} \log\left(1 + p_i \gamma_i\right)}{c_o^0 + c_o^r \sum_{i=1}^{n} \log\left(1 + p_i \gamma_i\right) + c_o^n n + c_p \sum_{i=1}^{n} p_i} \\ \text{s.t.} \quad & \sum_{i=1}^{n} \log\left(1 + p_i \gamma_i\right) \geq R_{\min} \\ & \sum_{i=1}^{n} p_i \leq P_{\max} \\ & 0 \leq p_i \leq P_{\text{peak}},\ i = 1..n. \end{aligned} \tag{6.3.6}$$

A ratio of functions $h_1(x)/h_2(x)$ is semistrictly quasi-concave if $h_1(x)$ is concave and non-negative and $h_2(x)$ is convex and positive. If $h_1(x)$ and $h_2(x)$ are differentiable, then $h_1(x)/h_2(x)$ is pseudoconcave. Quasi-convexity is a generalization of convexity. A quasi-convex function might have local minima, which are not global, or critical points, which are not local minimizers (e.g. points of inflection). In a semistrictly quasi-convex function, though, any local minima is global.

Problem (6.3.6) has $2n+2$ inequality constraints, which are concave. Hence, the problem can be reformulated as a quasi-convex optimization problem. In its standard form, the problem becomes:

$$\begin{aligned} \min_{\mathbf{x}} \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_k(\mathbf{x}) \leq 0,\ k = 1 \ldots 2n+2 \end{aligned} \tag{6.3.7}$$

with $\mathbf{x}$ the vector of transmission powers p_i, $i = 1..n$ and

$$\begin{aligned} f_0(\mathbf{x}) &= -\frac{\sum_{i=1}^{n} \log\left(1 + x_i \gamma_i\right)}{c_o^0 + c_o^r \sum_{i=1}^{n} \log\left(1 + x_i \gamma_i\right) + c_o^n n + c_p \sum_{i=1}^{n} x_i} \\ f_1(\mathbf{x}) &= R_{\min} - \sum_{i=1}^{n} \log\left(1 + x_i \gamma_i\right) \\ f_2(\mathbf{x}) &= \sum_{i=1}^{n} x_i - P_{\max} \\ f_j(\mathbf{x}) &= x_{j-2} - P_{\text{peak}},\ j = 3 \ldots n+2 \\ f_j(\mathbf{x}) &= -x_{j-2-n},\ j = n+3 \ldots 2n+2 \end{aligned} \tag{6.3.8}$$

6.4 Case $n = 1$

Let us begin with the simple case of a single channel, or $n = 1$. For mathematical clarity, we now assume that $c_o = c_o^0 + c_o^n$. The problem has one variable only: x denotes the transmission power p_i, x^* the optimal transmission power and γ is the CNR. It is clear that the constraint $x \geq 0$ can be removed because $R_{\min} \geq 0$ by definition, thus the minimum rate already imposes a minimum power constraint. Also, we will assume for simplicity that $P_{\text{peak}} \geq P_{\max}$ and will remove the peak power constraint.

The problem becomes a convex problem with one variable and 2 constraints:

$$\begin{aligned} f_0(x) &= -\frac{\log(1+x\gamma)}{c_o + c_o^r \log(1+x\gamma_i) + c_p x} \\ f_1(x) &= R_{\min} - \log(1+x\gamma) \\ f_2(x) &= x - P_{\max} \end{aligned} \tag{6.4.1}$$

Slater's condition [122] holds for any $x > 0$, the objective and constraint functions are differentiable and the problem is quasi-convex. Then, strong duality holds and Karush-Kuhn-Tucker (KKT) conditions provide necessary and sufficient conditions for optimality. Let λ_k, $k = 1, 2$ be the Lagrange multipliers, then if x^*, λ_k^*, $k = 1, 2$ exists and satisfy:

$$\begin{aligned} f_k(x^*) &\leq 0, k = 1, 2 \\ \lambda_k^* &\geq 0, k = 1, 2 \\ \lambda_k^* f_k(x^*) &= 0, k = 1, 2 \\ \nabla f_0(x^*) + \lambda_1^* \nabla f_1(x^*) + \lambda_2^* \nabla f_2(x^*) &= 0, \end{aligned} \tag{6.4.2}$$

then point x^* is a global optimum.

The last KKT condition is hard to solve analytically for $\lambda_1^*, \lambda_2^* > 0$. Let us assume that

$$\begin{aligned} f_1(x^*) &< 0 \\ f_2(x^*) &< 0. \end{aligned} \tag{6.4.3}$$

Then, according to strong duality, $\lambda_1^* = 0$ and $\lambda_2^* = 0$. This allows simplifying the stationary KKT condition to

$$\nabla f_0(x') = 0. \tag{6.4.4}$$

If the assumption (6.4.3) holds, $x^* = x'$, otherwise $\lambda_1^*, \lambda_2^* > 0$. From inspection of the problem, we see that the minimum rate constraint can be formulated as a minimum power constraint. Then, the constraints are box constraints and the feasible set is

$$x_{\min} \leq x \leq x_{\max}, \tag{6.4.5}$$

where $x_{\min} = (e^{R_{\min}} - 1)/\gamma$ and $x_{\max} = P_{\max}$. Since $f_0(x)$ is quasi-convex and x' is a global minimizer, then $x_1 \leq x' \Leftrightarrow f_0(x_1) \geq f_0(x')$ and $x_2 \geq x' \Leftrightarrow f_0(x_2) \geq f_0(x')$. Therefore,

$$\begin{aligned} &\text{if } x' \leq x_{\min} \Rightarrow x^* = x_{\min} \\ &\text{if } x' \geq x_{\max} \Rightarrow x^* = x_{\max}. \end{aligned} \tag{6.4.6}$$

In summary, the procedure solving the problem is first assuming the constraints are inactive and then finding the unconstrained solution x'. The constrained solution follows from:

$$x^* = \left[x'\right]_{x_{\min}}^{x_{\max}}, \tag{6.4.7}$$

where the operator $[x]_a^b \triangleq \min(\max(x, a), b)$.

The unconstrained solution x' satisfies (6.4.4), i.e.:

$$\begin{aligned} \frac{\partial}{\partial x'}\left(-\frac{\log\left(1+x'\gamma\right)}{c_o + c_o^r \log\left(1+x'\gamma\right) + c_p x'}\right) &= 0 \\ -\gamma\left(c_o + c_p x'\right) + c_p \log\left(1+x'\gamma\right)\left(1+x'\gamma\right) &= 0. \end{aligned} \tag{6.4.8}$$

Setting $r' = \log\left(1 + x'\gamma\right)$ and rearranging terms yields:

$$\begin{aligned} e^{r'}\left(r'-1\right) &= \gamma\frac{c_o}{c_p} - 1 \\ e^{r'-1}\left(r'-1\right) &= \left(\gamma\frac{c_o}{c_p} - 1\right)e^{-1}. \end{aligned} \tag{6.4.9}$$

This equation can be solved analytically using the principal branch of the Lambert-W function, which satisfies $W(x)e^{W(x)} = x$. Applying $W(\cdot)$ at both sides we obtain:

$$r' = 1 + W\left(\left(\gamma\frac{c_o}{c_p} - 1\right)e^{-1}\right), \tag{6.4.10}$$

which is the optimal transmission rate.

The asymptotic behaviour of the resource-efficient rate is obtained using the first term of the series expansion of the Lambert-W function (Appendix A)

$$W(x) \approx \log(x),\ x \to \infty, \tag{6.4.11}$$

then, for $\gamma \to \infty$, the optimal rate behaves like:

$$r' \sim \log\left(\gamma\frac{c_o}{c_p}e^{-1}\right). \tag{6.4.12}$$

Moreover, in Appendix A it is shown that $W(x) < \log x$ for $x > e$. Then, in fact $r' < \log(\gamma\frac{c_o}{c_p}e^{-1})$. For large CNR the capacity behaves like $C \sim \log(\gamma p)$ with p the transmission power. Hence, we conclude that the optimal rate is lower than the capacity of a channel where the transmission power is equivalent to $p = \frac{c_o}{c_p}e^{-1}$, albeit the asymptotic behaviour with respect to the CNR is equivalent.

$W(z)$ is strictly increasing and defined for $z \in (-e^{-1}, \infty)$, and takes values $W(z) \in (-1, \infty)$. The solution $r' = 0$ only occurs if $\gamma\frac{c_o}{c_p} = 0$. Paradoxically, if the operational costs $c_o = 0$, the most efficient strategy is switching off the transmitter (or $r' = R_{\min}$ after applying the constraints).

Equation (6.4.10) also reveals that the optimal rate

- increases with the channel gain γ,
- increases with the constant cost c_o and
- decreases with the transmission power cost c_p.

Figure 6.1 plots the optimal rate as a function of the ratio $\gamma\frac{c_o}{c_p}$. When the receiver is close to the transmitter, γ is high and the required transmission power is small. The best efficiency is achieved by transmitting with a high rate. When the distance increases, γ decreases and the power required to overcome the channel attenuation becomes too high and therefore it is more efficient to reduce the transmitted rate and power.

The costs are measured in currency units per second. If the time required for processing 1 bit of information is reduced (increasing the rate), the total currency units per transmitted

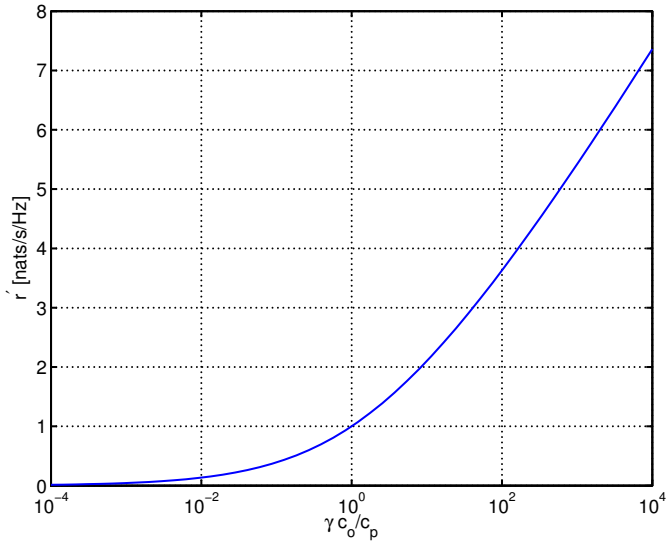

Figure 6.1: Optimal spectral efficiency as a function of the ratio $\gamma \frac{c_o}{c_p}$.

bit is lower. If the operation cost c_o is high, we increase the rate so that the infrastructure is operated for less time. If the power is expensive (c_p is high), we decrease the rate, reducing the transmission power. Note that low channel gain are counterbalanced by low power costs, that is, if the channel attenuation is very high but the power is cheap, the optimal solution is transmit at high rate.

The solution to the original problem, or optimal transmission power is:

$$x^* = \begin{cases} \left[\dfrac{\frac{c_o}{c_p} - \frac{1}{\gamma}}{W\left(\left(\gamma \frac{c_o}{c_p} - 1\right) e^{-1}\right)} - \dfrac{1}{\gamma} \right]_{x_{\min}}^{x_{\max}} & \text{if } \gamma \dfrac{c_o}{c_p} \neq 1 \\ \left[\dfrac{e-1}{\gamma} \right]_{x_{\min}}^{x_{\max}} & \text{if } \gamma \dfrac{c_o}{c_p} = 1 \end{cases} \tag{6.4.13}$$

with $x_{\min} = (e^{R_{\min}} - 1)/\gamma$ and $x_{\max} = P_{\max}$. The singularity at $\gamma \frac{c_o}{c_p} = 1$ is solved using the fact that $\lim_{x \to 0} W(x) = x$.

According to (6.4.13), the optimal power allocation is independent of the cost associated to the transmission rate, c_o^r. The c_o^r cost decreases the optimal efficiency but the slope (or partial derivative) of the resource efficiency as a function of x is independent of c_o^r.

The optimal resource efficiency η^*, follows from inserting (6.4.13) in the η function:

$$\eta^* = \left[c_o^r + c_p \frac{\frac{c_o}{c_p} - \frac{1}{\gamma}}{W\left(\left(\gamma \frac{c_o}{c_p} - 1\right) e^{-1}\right)} \right]^{-1}. \tag{6.4.14}$$

c_o^r is an offset that increases or decreases the efficiency, without affecting the optimal power. Using the first derivative of the $W(\cdot)$ function (Appendix A), it is straightforward to see that η^* is strictly decreasing with c_o, c_o^r and c_p and increasing with γ. Fig. 6.2 plots the optimal resource efficiency as a function of the distance between the transmitter and receiver. γ is obtained assuming the free-space path loss model (4.4.3) with a path loss exponent $\alpha = 3$ and a carrier frequency of 2 GHz and a receiver with a noise factor of 7 dB and 10 MHz channel bandwidth.

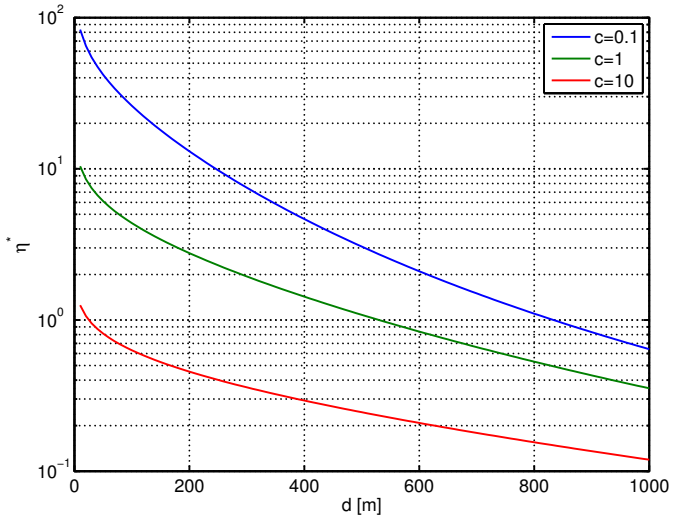

Figure 6.2: Optimal resource efficiency η^* as a function of the distance, for different ratios of offset to power costs $c = \frac{c_o}{c_p}$.

As $\gamma \to \infty$, $\eta^* \to \infty$ and behaves asymptotically as $\eta^* \sim W(\gamma) \sim \log(\gamma)$. Thus, the rate as well as the efficiency increase with the CNR as the capacity of an AWGN channel.

Figure 6.3 plots the gains in resource efficiency when power is adapted for maximizing the resource efficiency compared to the channel inversion power adaptation –power is adapted to maintain a constant rate. The optimal solution is up to 6 times more efficient, depending on the distance between the transmitter and receiver and the cost ratio $c = \frac{c_o}{c_p}$. We have assumed the same propagation parameters than in Fig. 6.2. When the optimal rate coincides with the fixed rate, at 200 m, 500 m and 1000 m for $c = 0.1$, $c = 1$ and $c = 10$ respectively, the constant rate solution is optimal.

6.5 Case $n = 1$ with Outage and Retransmissions

If the transmitter has no CSI, the problem is still mathematically tractable if $n = 1$ [121]. Let the CNR γ be a random variable with p.d.f. $f(x)$. Given a transmitted power x, the maximum rate (in nats/s/Hz) for reliable communications is $C = \log(1 + x\gamma)$, which is a random variable too. The instantaneous γ is unknown to the transmitter. The encoding rate is then fixed and independent of the channel characteristics: $r = \log(1 + x\gamma_0)$. The information is recovered by the receiver if $r \leq C$. This occurs when $\gamma_0 \leq \gamma$. If $\gamma_0 > \gamma$, then $r > C$ and the system operates above capacity. Decoding is not possible and the system is said to be in outage. The outage probability is $\mathcal{P}_{out} = \mathcal{P}(\gamma < \gamma_0) = F_\gamma(\gamma_0)$, where $F_\gamma(x)$ is the c.d.f. of γ.

The transmitter sends the information in packets of L bits. If an outage occurs during the transmission of a packet, the receiver requests a packet retransmission. Retransmissions are repeated until the packet is correctly decoded. We can neglect the resources consumed for requesting retransmissions if the information packet is larger than the retransmission request packet. We also assume that the probability of outage is independent for each bit, corresponding to a perfectly interleaved and memoryless channel. The average number of retransmissions N_r was derived in [123]:

$$\mathbb{E}\{N_r(x)\} = (1 - \Phi(x))^{-L}\,, \tag{6.5.1}$$

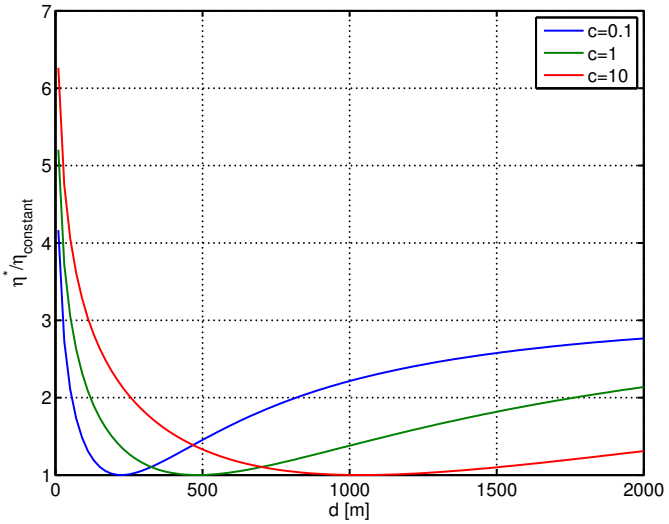

Figure 6.3: Resource efficiency difference between constant rate link adaptation (η_{constant}) and optimal power allocation (η^*) strategies. The rate is fixed to 2 nats/s/Hz.

where $\Phi(x)$ is the bit error probability. In our analysis, $\Phi(x) = F_\gamma(\gamma_0)$ and the bit error probability corresponds to the outage probability.

The resource efficiency is divided by the average number of retransmissions to become the *average resource efficiency*. Then, the optimization problem cost function depends on two variables:

$$\begin{aligned} \min_{x,\gamma_0} \quad & -\frac{\log\left(1+x\gamma_0\right)}{c_o + c_p x}\left(1 - F_\gamma(\gamma_0)\right)^L \\ \text{s.t.} \quad & \log\left(1+x\gamma_0\right) \geq R_{\min} \\ & 0 \leq x \leq P_{\max} \\ & \gamma_0 \geq 0 \end{aligned} \tag{6.5.2}$$

where we have removed the rate-dependent cost since we have shown that is irrelevant to the optimal power allocation.

The special case $L = 1$ models the *effective resource efficiency* when retransmissions are disabled. Note that $\log\left(1+x\gamma_0\right)\left(1 - F_\gamma(\gamma_0)\right)$ is the effective bit rate of a fading channel with receiver CSI.

6.5.1 Convexity Analysis

In Section 6.4 we showed that, since the objective function was quasi-convex, the existence of a unique solution was guaranteed. Unfortunately, the function (6.5.2) is not jointly quasi-convex with respect to the two variables x and γ_0. The Hessian matrix of (6.5.2) is not positive-semidefinite for $x, \gamma_0 \geq 0$ and any $F_\gamma(x)$. On the other hand, a ratio of functions $h_1(z)/h_2(z)$ is semistrictly quasi-concave if $h_1(z)$ is concave and non-negative and $h_2(z)$ is convex and positive. With the negative sign of $f_0(x)$ the quasi-concave fractional function becomes a quasi-convex one. Let define

$$h_1(x) = \log\left(1+x\gamma_0\right) \tag{6.5.3}$$

$$h_2(x) = \left(c_0 + c_p x\right)\left(1 - F_\gamma(\gamma_0)\right)^{-L}. \tag{6.5.4}$$

$h_1(x)$ is bi-concave, i.e. concave with respect to each variable and non-negative. Thus, it is only possible to prove quasi-convexity of $f_0(x, \gamma_0)$ with respect to each variable. As opposed to convex optimization problems, biconvex optimization problems are global optimization problems, which may have a large number of local minima. The biconvex nature of the problem can be exploited for finding a solution more efficiently [124], though.

A product of two positive convex functions is convex [122]. To show that $h_2(x)$ is convex, we need to show that the average number of retransmissions is convex with respect to γ_0. In continuation we show that (6.5.1) is strictly convex if bit error function $\Phi(y)$ is due to channel noise or outage. In particular, (6.5.1) is convex if the BER function is convex and retransmissions are due to channel noise. When retransmissions are due to channel fading, (6.5.1) is convex if the channel model p.d.f. is log-concave. As a consequence, our results are applicable to system models where retransmissions are due to channel noise or channel outage errors.

Lemma 6.1. *Let Y be a random variable with distribution function $F(y)$ and density function $f(y)$. If $f(y)$ is log-concave, then $\bar{F}(y) = 1 - F(y)$ is log-concave, too.*

Proof. See [125, Theorem 2]. □

Theorem 6.1. *Let $g(y)$ be a log-concave function, then $g(y)^{-L}$ is strictly convex for $L \geq 1$.*

Proof. $g(y)^{-L}$ is strictly convex if its second derivative is positive. The first derivative is

$$\frac{d}{dy} g(y)^{-L} = -Lg(y)^{-(L+1)} g'(y) \tag{6.5.5}$$

and the second derivative is

$$\frac{d^2}{dy^2} g(y)^{-L} = g'(y)^2 (L+1) - g''(y)g(y). \tag{6.5.6}$$

$g(y)$ is log-concave if $\log(g(y))'' \leq 0$, or

$$g''(y)g(y) \leq g'(y)^2. \tag{6.5.7}$$

Then, $\frac{d^2}{dy^2} g(y)^{-L} > 0$ if

$$g''(y)g(y) < g'(y)^2 (L+1) \tag{6.5.8}$$

which holds if $g(y)$ is log-concave and $L \geq 1$. □

If $\Phi(y)$ is a convex function, $1 - \Phi(y)$ is concave and the conditions of Theorem 6.1 are satisfied (concavity implies log-concavity). The convexity of bit and symbol error probabilities have been recently proven in the high SNR regime for arbitrary constellations and bit mappings with coding and maximum-likelihood decoding [126]. Convexity in the low SNR regime exists for many common modulations.

If errors are due to channel outages, $\Phi(y)$ is the channel c.d.f. and y is the cutoff rate γ_0. According to Lemma 6.1, the channel model p.d.f. needs to be log-concave for Theorem 6.1 to be applicable. Most probability density functions are log-concave [125]. The log-normal shadowing model, for instance, has a log-normal p.d.f.. Hence, the average number of retransmissions is convex [127]. The CNR under the Rayleigh fading channel follows an exponential p.d.f., which is log-concave [120]. The Nakagami distribution is a popular channel model where the direct signal path is received with higher power. This model can also account for Rician fading and maximum ratio combining with N-branches of i.i.d. Rayleigh fading

signals. The p.d.f. of the received power under the Nakagami fading model follows the Gamma distribution

$$f(y; m, \sigma) = \frac{\left(\frac{m}{\sigma}\right)^m}{\Gamma(m)} y^{m-1} e^{-\frac{m}{\sigma}y}, \tag{6.5.9}$$

where σ is the average reception power, Γ the gamma function and m the shaping parameter. Taking the second derivative of the logarithm of (6.5.9) leads to

$$(\log f(x; m, \sigma))'' = \frac{1-m}{x^2}. \tag{6.5.10}$$

Hence, the gamma distribution is log-concave for $m \geq 1$ and Theorem 1 can be applied. The p.d.f. is log-convex for $m < 1$ and the c.d.f. $F_\gamma(y)$ is log-concave. The complementary c.d.f. $\bar{F}_\gamma(y) = 1 - F_\gamma(y)$ is log-convex and Theorem 1 cannot be applied. The average number of retransmissions $(1 - F_\gamma(y))^{-L}$ is concave for small y and convex thereafter. Non-convex solvers should be employed for this channel model with $m < 1$.

Log-concavity is preserved after multiplication [122, pp. 105] and, in several cases, after integration [122, pp. 106]. Thus, combining shadowing and fading may result in log-concave density functions. This conjecture, though, needs to be formally proved for all shadowing-fading combinations.

6.5.2 Optimal and Sub-Optimal Solutions

In the last section we have shown that the optimization problem is semistrictly quasi-convex with respect to each variable for $x, \gamma_0 > 0$. A biconvex problem is solved dividing it into a set of convex problems, which are solved iteratively for each variable while all other variables are kept constant. A semistrictly quasi-biconvex problem can be divided into a set of semistrictly quasi-convex problems. This means that we first need to verify that each stationary point is a local minimizer (with respect to each variable).

We will proceed equivalently as in Section 6.4: we first solve the unconstrained optimization problem and then apply the rate and power constraints as lower and upper bounds to the final solution. The stationary points satisfy $\nabla f_0(x^*, b) = 0$ and $\nabla f_0(a, \gamma_0^*) = 0$, where a and b are arbitrary positive real numbers.

The curve of stationary points with respect to x is analogous to (6.4.13):

$$x^*(b) = \begin{cases} \left[\dfrac{\frac{c_o}{c_p} - \frac{1}{b}}{W\left(\left(b\frac{c_o}{c_p} - 1\right)e^{-1}\right)} - \dfrac{1}{b} \right]_{x_{\min}}^{x_{\max}} & \text{if } b\dfrac{c_o}{c_p} \neq 1 \\ \left[\dfrac{e-1}{b} \right]_{x_{\min}}^{x_{\max}} & \text{if } b\dfrac{c_o}{c_p} = 1, \end{cases} \tag{6.5.11}$$

which is non-negative for $b > 0$. The upper and lower bounds $x_{\max}$ and $x_{\min}$ are defined in Section 6.4. The optimal power allocation is zero when $b\frac{c_o}{c_p} = 0$. It can be shown that $\frac{d^2}{dx^2} f_0(x, b)|_{x=x^*(b)} > 0$ for any $b\frac{c_o}{c_p} > 0$. The solution given by (6.5.11) is, therefore, a local minimizer with respect to x.

The curve of stationary points w.r.t γ_0 satisfies

$$\frac{F'(\gamma_0^*)}{1 - F_\gamma(\gamma_0^*)} \log(1 + \gamma_0^* a)(1 + \gamma_0^* a) = \frac{a}{L}. \tag{6.5.12}$$

It is difficult to verify that the solution to (6.5.12) corresponds to a local minimum for any channel distribution. For the Rayleigh channel model with average CNR $\bar{\gamma}$, the function c.d.f. becomes

$$F_{\text{ray}}(\gamma_0) = 1 - e^{-\frac{\gamma_0}{\bar{\gamma}}} \tag{6.5.13}$$

Inserting (6.5.13) in (6.5.12), we get the optimal cutoff value for the Rayleigh fading case

$$\gamma^*_{0,\text{ray}}(a) = \frac{\bar{\gamma}}{L \cdot W(a\bar{\gamma}/L)} - \frac{1}{a}. \tag{6.5.14}$$

It can be shown that $\frac{d^2}{dy^2} f_o(a,y)|_{y=\gamma^*_{0,\text{ray}}(a)} > 0$ for $a > 0$, which indicates that (6.5.14) corresponds to a local minimizer. Given the curves of stationary points an iterative solver (e.g. Alternate Convex Search [124]) can be used to find the optimal point (x^*, γ^*_0). Iterative methods can be avoided if we assume that the system works in the low SNR regime. The first order Taylor approximation of $\log(1+x)(1+x)$ around $x = 0$ simplifies (6.5.12) to

$$\frac{F'_\gamma(\gamma^*_0)}{1 - F_\gamma(\gamma^*_0)}\gamma^*_0 = \frac{1}{L}, \tag{6.5.15}$$

which is independent of a. The iterative procedure is then unnecessary because the optimal power is $x^*(\gamma^*_0)$ with γ^*_0 given by (6.5.15). Equation (6.5.15) becomes

$$\frac{\gamma^*_{0,\text{ray}}}{\bar{\gamma}} = \frac{1}{L} \tag{6.5.16}$$

for the Rayleigh channel model.

For more complex channel models, the c.d.f. is usually hard to manipulate and further simplifications are necessary. If we assume that the outage probability approaches zero, $F_\gamma(x) \approx 0$ and the c.d.f. is eliminated from the expression. The optimal cutoff rate for the Nakagami-m channel model with average CNR $\bar{\gamma}$ then becomes

$$\frac{\gamma^*_{0,\text{nak}}}{\bar{\gamma}} = -W\left(-\frac{1}{m}\left(\frac{\Gamma(m)}{L}\right)^{1/m}\right). \tag{6.5.17}$$

Note that $\gamma^*_{0,\text{nak}}$ is never zero. The domain of the principal branch of the Lambert-W function is $(-e^{-1}, +\infty)$. Our approximation is thus valid for $L \geq \Gamma(m)/(me^{-1})^m$. If $m < 6.5$, the packet length has to be $L > 1$ which means that for $m < 6.5$ our approximation is valid only when retransmissions are enabled. This restriction is explained because the outage probability approaches zero if either m or L are very large.

6.5.3 Numerical Results

The closed form expressions (6.5.16) and (6.5.17) allow deriving some interesting conclusions. First, the outage probability is determined only by the ratio $\gamma^*_0/\bar{\gamma}$ and parameter m for the Nakagami channel. In other words, the optimal cut-off rate γ^*_0 is independent of the operational or power costs. This ratio decreases very fast with increasing packet length L. The probability that a packet experiences outage grows exponentially with L. Therefore, a EM strategy reduces the outage probability, preventing retransmissions.

Comparing the power allocation with and without CSI (in section 6.4 the case with CSI was derived) we note that when CSI is not available to the transmitter, the information is encoded assuming a lower CNR. For the Rayleigh case, the optimal encoding rate assumes a L times lower channel gain (equation (6.5.16). Therefore, channel outages are combated by increasing the encoding robustness.

Figures 6.4 and 6.5 plot the outage probability and resource efficiency of the optimal and approximated solutions. $\bar{\gamma}$ is obtained assuming the free-space path loss model (4.4.3) with a path loss exponent $\alpha = 3$ and a carrier frequency of 2 GHz and a receiver with

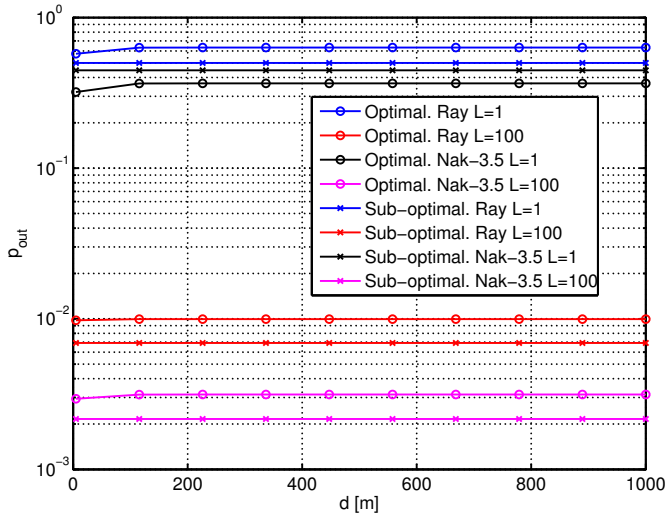

Figure 6.4: Outage probability for the optimal and approximated solutions with ($L = 100$) and without ($L = 1$) retransmissions. Ray and Nak indicate Rayleigh and Nakagami fading, respectively.

a noise factor of 7 dB and 10 MHz channel bandwidth. Our solutions are compared with the resource efficiency when the instantaneous channel gain is perfectly known (AWGN in the figure). The loss of resource efficiency due to the lack of CSI at the transmitter is considerable, specially when retransmissions are required. This indicates that despite CSI feedback have a considerable overhead costs, the SDR cloud resources are used much more efficiently. Whether this increased efficiency pays back the extra costs due to CSI feedback is a topic of future research.

6.6 Case $n \geq 1$

In the general case with CSI available at the transmitter, the problem functions $f_k(\mathbf{x})$, $k = 0 \ldots 2n+2$ take the form:

$$
\begin{aligned}
f_0(\mathbf{x}) &= -\frac{\sum_{i=1}^{n} \log\left(1 + x_i \gamma_i\right)}{c_o^0 + c_o^r \sum_{i=1}^{n} \log\left(1 + x_i \gamma_i\right) + c_o^n n + c_p \sum_{i=1}^{n} x_i} \\
f_1(\mathbf{x}) &= R_{\min} - \sum_{i=1}^{n} \log\left(1 + x_i \gamma_i\right) \\
f_2(\mathbf{x}) &= \sum_{i=1}^{n} x_i - P_{\max} \\
f_j(\mathbf{x}) &= x_{j-2}, \; -P_{\text{peak}} = 3 \ldots n+2 \\
f_j(\mathbf{x}) &= -x_{j-2-n}, \; j = n+3 \ldots 2n+2
\end{aligned}
\tag{6.6.1}
$$

The constraint functions are convex and the objective function is quasi-convex. Problem (6.6.1) satisfies the Slater conditions so that the KKT conditions provide necessary and sufficient conditions for optimality:

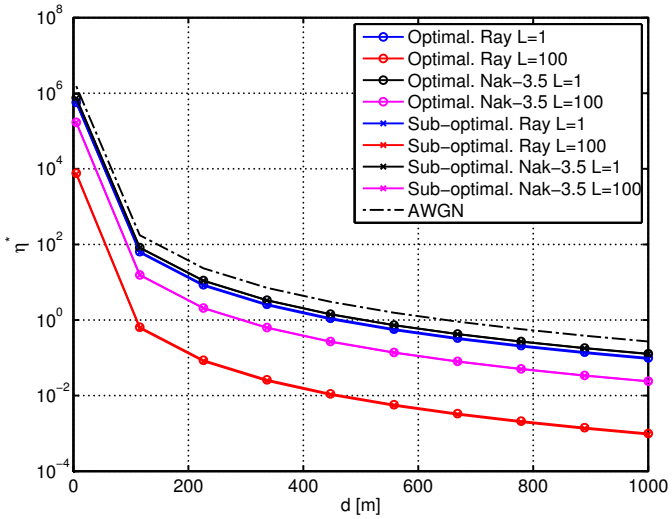

Figure 6.5: Average resource efficiency per bit for the optimal and approximated solutions with ($L = 100$) and without ($L = 1$) retransmissions. Ray and Nak indicate Rayleigh and Nakagami fading, respectively.

$$\begin{aligned} f_k(\mathbf{x}^*) &\leq 0, k = 1..n+2 \\ \lambda_k^* &\geq 0, k = 1..n+2 \\ \lambda_k^* f_k(\mathbf{x}^*) &= 0, k = 1..n+2 \\ \nabla f_0(\mathbf{x}^*) + \sum_{j=1}^{2n+2} \lambda_j^* \nabla f_j(\mathbf{x}^*) &= 0. \end{aligned} \tag{6.6.2}$$

6.6.1 Equivalent Parametric Problem

In this case, it is not possible to solve the problem as in the $n = 1$ case, because constraint functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ depend on all x_i variables. Let us assume that the optimal powers $x_i^* \geq 0$, so that $\lambda_j = 0$, $j = n+3 \ldots 2n+2$. The non-negativity constraint will be imposed to the final solution.

Solving the last equation of the KKT conditions is still cumbersome. Inspired by [103, 104] we convert the fractional problem into a parametric one (see Appendix B). The idea is to define a convex optimization problem with a parameter q using the objective function:

$$F(x,q) = q\left[c_o^0 + c_o^r \sum_{i=1}^{n} \log\left(1 + x_i\gamma_i\right) + c_o^n n + c_p \sum_{i=1}^{n} x_i\right] - \sum_{i=1}^{n} \log\left(1 + x_i\gamma_i\right). \tag{6.6.3}$$

Since $F(x,q)$ is a sum of convex functions of x, it is possible to solve the parametric problem for a given q (q is a parameter, not a variable). Dinkelbach proofs [128] that the solution to the original problem (x^*, q^*) satisfies

$$F(x^*, q^*) = \min\left\{F(x, q^*) | x \in X\right\} = 0. \tag{6.6.4}$$

For clarity, we will change the notation of the Lagrange multipliers $\lambda_{j+2} = \nu_j$ for $j = 2 \ldots n+2$. Thus, the last equation of the KKT conditions becomes:

$$\nabla F(x^*, q) + \lambda_1^* \nabla f_1(x^*) + \lambda_2^* \nabla f_2(x^*) + \sum_{j=1}^{n} \nu_j^* \nabla f_{j+2}(\mathbf{x}^*) = 0, \tag{6.6.5}$$

which is possible to solve analytically. The advantage of this solution methodology is that we divide a hard problem in two steps: a simpler optimization problem plus a root finding.

Equation (6.6.5) becomes

$$\begin{aligned} \nabla \left(q \left[c_o^0 + c_o^r \sum_{i=1}^{n} \log\left(1 + x_i^* \gamma_i\right) + c_o^n n + c_p \sum_{i=1}^{n} x_i^* \right] - \sum_{i=1}^{n} \log\left(1 + x_i^* \gamma_i\right) \right) \\ -\lambda_1^* \nabla \left(\sum_{i=1}^{n} \log\left(1 + x_i^* \gamma_i\right) + R_{\min} \right) \\ +\lambda_2^* \nabla \left(\sum_{i=1}^{n} x_i^* - P_{\max} \right) \\ + \sum_{j=1}^{n} \nu_j^* \nabla \left(x_j^* - P_{\text{peak}} \right) = 0, \end{aligned} \tag{6.6.6}$$

and rearranging terms yields:

$$\left(q c_o^r - 1 - \lambda_1^*\right) \nabla \sum_{i=1}^{n} \log\left(1 + x_i^* \gamma_i\right) + \left(q c_p + \lambda_2^* + \nu_i^*\right) \nabla \sum_{i=1}^{n} x_i^* = 0. \tag{6.6.7}$$

Solving for x_i^* and applying the non-negative and peak power constraints yields the final solution:

$$x_i^* = \left[\frac{1 + \lambda_1^* - q c_o^r}{q c_p + \lambda_2^* + \nu_i^*} - \frac{1}{\gamma_i} \right]_0^{P_{\text{peak}}}. \tag{6.6.8}$$

Equation (6.6.8) reminds the well-known water-filling power allocation with individual waterlevels

$$x_i^* = \left[\frac{1}{\mu_i} - \frac{1}{\gamma_i} \right]_0^{P_{\text{peak}}}, \tag{6.6.9}$$

with

$$\mu_i = \frac{q c_p + \lambda_2^* + \nu_i^*}{1 + \lambda_1^* - q c_o^r}. \tag{6.6.10}$$

Water-filling power allocations are present in many power allocation problems. The water-filling principle states that more power is allocated to the *best* channels and a channel is allocated zero power if $\gamma_i \leq \mu_i$. When a channel is allocated zero power it is said to be *inactive*.

6.6.2 Constrains

The waterlevel equation (6.6.10) still depends on the optimal dual variables $\lambda_1^*, \lambda_2^*, \nu_i^*$ corresponding to the sum-rate, sum-power and peak power constraints, respectively. The peak power constraint adds a significant degree of complexity to the problem. It can be seen, however, that a positive ν_i^* corresponds to a rise of the waterlevel. In section 7.5 we propose

a sub-optimal solution when $\nu_i^* > 0$ for some $i = 1 \dots n$. In the remaining of this section, we assume $x_i^* < P_{\text{peak}}$ and $\nu_i^* = 0$ and redefine:

$$\mu \triangleq \mu_i = \frac{qc_p + \lambda_2^*}{1 + \lambda_1^* - qc_o^r}, \text{ for } i = 1 \dots n. \tag{6.6.11}$$

The sum-power and sum-rate constraints are mathematically simpler. The following solution methodology will be applied [103]:

1. Assume $\lambda_1^* = \lambda_2^* = 0$ and find the unconstrained solution $(\mathbf{x}^*, q^*)$.

2. *Apply minimum rate constraint:* If $f_1(\mathbf{x}^*) \leq 0$, then by complementary slackness, $\lambda_1^* = 0$. Otherwise, it is observed in (6.6.8) that $\lambda_1^* > 0$ corresponds to a rise of the waterlevel, until the rate constraint is fulfilled. The solution is then equivalent to minimize the transmission power subject to a minimum rate constraint.

3. *Apply maximum power constraint:* If $f_2(\mathbf{x}^*) \leq 0$, then by complementary slackness, $\lambda_2^* = 0$. Otherwise $\lambda_2^* > 0$ is equivalent to lower the waterlevel until the maximum power constraint is satisfied; in other words, the problem becomes to maximize the sum-rate subject to a maximum power constraint.

The mathematical proof of this solution method is sketched for the maximum power constraint case:

Proof. The objective function $f_0(\mathbf{x})$ is quasi-convex and unimodal. Consequently, any local minima is global. Let $\mathbf{x}^*$ be a local minimizer of $f_0(\mathbf{x})$ and $\bar{\mathbf{x}} = \sum x_i^*$. Then, $f_0(\mathbf{x})$ is decreasing until $\bar{\mathbf{x}} = \mathbf{x}^*$ and then increases. If $P_{\max} < \mathbf{x}^*$, the minimum is at $\bar{\mathbf{x}} = P_{\max}$, because the function is decreasing. The optimal value of the denominator chooses $\sum x_i^* = P_{\max}$ to be constant and the function is minimized by minimizing the numerator, that is, by maximizing the rate subject to a maximum power constraint. □

Note that both constraints can not be active, otherwise the problem is infeasible.

Let us now focus on the unconstrained case, $\lambda_1^* = \lambda_2^* = 0$. To find the optimal parameter q^*, we need to solve $F(q^*, \mathbf{x}^*) = 0$, using (6.6.8). Let us define the change of variable

$$q' = \frac{1}{c_p}\left(\frac{1}{q^*} - c_o^r\right), \tag{6.6.12}$$

then (6.6.7) becomes:

$$q' \sum_{i=1}^{n} \log(q'\gamma_i)^+ - \sum_{i=1}^{n} \left(q' - \frac{1}{\gamma_i}\right)^+ = c, \tag{6.6.13}$$

with

$$c = \frac{c_o^0 + c_o^n n}{c_p}. \tag{6.6.14}$$

Equation (6.6.13) is called the *waterlevel equation*. Its solution finally solves the power allocation problem. The left-hand side is an piecewise-linear increasing function of q' with breakpoints at γ_i^{-1}, so equation (6.6.13) has a unique solution. It is straightforward to obtain both iterative and exact algorithms to solve this equation. Iterative algorithms are obtained by fixing the waterlevel to some value and then adjusting the waterlevel iteratively until the constraint is satisfied. This can be done by modifying the waterlevel with small decreasing steps or by bisection.

6.6.3 Iterative Solution

Exact algorithms are based on hypothesis testing. The idea is to form a hypothesis of the active ($x_i > 0$) and inactive ($x_i = 0$) channels and check whether a consistent solution is found conditioned to the hypothesis. With an exact method, it is possible to obtain worst-case complexity of n iterations. Assuming $x_i > 0$ for $i = 1..n$, (6.6.13) yields

$$q' \left(n \log(q') + \sum_{i=1}^{n} \log(\gamma_i) \right) - nq' + \sum_{i=1}^{n} \frac{1}{\gamma_i} = c. \tag{6.6.15}$$

Define:

$$\hat{h} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\gamma_i} \tag{6.6.16}$$

$$\hat{g} = \frac{1}{n} \sum_{i=1}^{n} \log \gamma_i \tag{6.6.17}$$

$$\bar{c} = \frac{c}{n} \tag{6.6.18}$$

where $\hat{h}$ is the inverse of the harmonic mean and $\log(\hat{g})$ is the geometric mean. $\bar{c}$ is the *normalized cost*. Rearranging terms

$$q' \left(\log q' + g - 1 \right) = \bar{c} - h, \tag{6.6.19}$$

and applying the change of variable $s = \log q'$ and multiplying both sides by $e^{\hat{g}-1}$ yields:

$$e^{s+\hat{g}-1} \left(s + \hat{g} - 1 \right) = \left(\bar{c} - \hat{h} \right) e^{\hat{g}-1}. \tag{6.6.20}$$

Equation (6.6.20) is a transcendental equation whose solution can be given by the Lambert-W function (Appendix A). Applying $W(x)$ to both sides of the equality and using the identity $\exp(W(x)) = x/W(x)$ yields to the closed form solution:

$$q' = \frac{\bar{c} - \hat{h}}{W\left[\left(\bar{c} - \hat{h} \right) e^{\hat{g}-1} \right]}. \tag{6.6.21}$$

$W()$ is the principal branch of the Lambert-W function, which is defined for $x \geq -1/e$ and takes values in the range $(-1, +\infty)$. A careful look at (6.6.21) reveals that q' never takes negative values. $W(x) \in [-1, 0)$ if $-1/e \leq x < 0$, which happens only when $\bar{c} < \hat{h}$ in which case the numerator is also negative.

If $F(q^*, \mathbf{x}^*) = 0$, we see from (6.6.13) that the optimal resource efficiency η^* is

$$\eta^* = q^*. \tag{6.6.22}$$

Therefore, if the hypothesis that all channels are active is true and q' is the optimal waterlevel, the optimal resource efficiency is:

$$\eta^* = \left(c_o^r + c_p \frac{\bar{c} - \hat{h}}{W\left[\left(\bar{c} - \hat{h} \right) e^{\hat{g}-1} \right]} \right)^{-1}. \tag{6.6.23}$$

Assuming $c_o^r = 0$, η^* becomes

$$\eta^* = \frac{1}{c_p} \frac{W\left[\left(\bar{c} - \hat{h} \right) e^{\hat{g}-1} \right]}{\bar{c} - \hat{h}}, \tag{6.6.24}$$

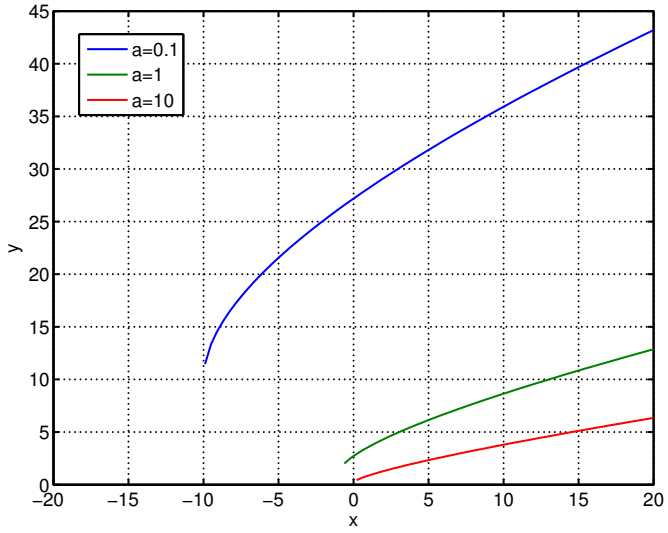

Figure 6.6: Plot of the function $y(x) = \frac{x}{W(xae^{-1})}$. It represents equation (6.6.21) with $a = e^{\hat{g}}$ and $x = \bar{c} - \hat{h}$.

The Lambert-W function is defined in the real domain if and only if

$$\bar{c} \geq \hat{h} - e^{-\hat{g}} \triangleq \kappa. \tag{6.6.25}$$

Otherwise, (6.6.19) has no solution for real q'. Fig. 6.6 plots (6.6.21), showing the points where the original transcendental equation has a real solution. The main branch of the function $W(xae^{-1})$ is undefined for real values if $x < -10$, $x < -1$ and $x < 0$ when $a = 0.1$, $a = 1$ and $a = 10$, respectively.

On the other hand, $\hat{h}^{-1}$ and $\log(\hat{g})$ are the harmonic and geometric mean of the channel gains, respectively. According to the inequality of arithmetic and geometric means, we know that κ is a non-negative number for any sequence of channel gains. In fact, $\kappa = 0$ if and only if all channel gains are equal. Therefore, depending on the distribution of channels, equation (6.6.19) has no solution. This occurs because the assumption that all channels are active is not valid. We can intuit that the fraction of active channels in the final solution depends on the relation between the geometric and harmonic means.

For the general case, the power allocation policy is

$$x_i^* = \left(q' - \frac{1}{\gamma_i}\right)^+ \tag{6.6.26}$$

with q' the solution to (6.6.13). It is clear that the optimal EMPA is independent of the sum-rate cost c_o^r for $n \geq 1$. q' is a function of $\bar{c}$ and the channel gains γ_i. As it was shown in the case $n = 1$, the sum-rate cost c_o^r is an offset to the cost function and modifies the final cost, but the optimal power allocation is independent of c_o^r. Therefore, it is safe to ignore the cost c_o^r from the optimization problem.

6.7 Summary

In this chapter, we have introduced the EMPA problem for $n \geq 1$ parallel channels. The network operator pays for using these channels as modeled by the costs model defined in Chapter 5 (*pay-per-use*). The EMPA problem aims at delivering the maximum amount of bits per consumed resources, in equivalent currency units.

Closed form solutions of this problem were obtained for the case $n = 1$. The results show that the optimal transmission rate is an increasing function of $\gamma \frac{c_o}{c_p}$. Thus, high rates should be employed if either the channel is very good or the operational costs are very high, so that the infrastructure is used less time per transmitted bit. If $c_o \ll c_p$ or even $c_o = 0$, the most efficient strategy is transmitting the minimum rate required by the service, i.e. $r' = R_{\min}$. On the other hand, if $c_o \gg c_p$ or $c_p = 0$ the optimal strategy is transmitting the maximum power possible.

If $n = 1$ and the transmitter has no CSI, channel outages occur and the optimization problem becomes more complex. We showed that the problem is bi-concave for certain channel statistics and introduced closed-form and approximated solutions for the Rayleigh and Nakagami-m fading cases, respectively. The results show that information should be encoded more robustly in order to prevent retransmissions.

The general case $n \geq 1$ with complete CSI is a constrained convex optimization problem. The problem is solved by converting the fractional program into an equivalent parametric one. The solution has the form of water-filling power allocation. It is shown that if the unconstrained solution violates the power constraint, the resource efficient solution is the rate maximizing power allocation. On the other hand, if the solution violates the rate constraint, the resource efficient solution is the margin maximizing power allocation.

The constant $\bar{c}$ has been defined as $\bar{c} = c/n$, where $c = \frac{c_o^0 + c_o^n n}{c_p}$ is the ratio of operational to power costs. An iterative solution has been introduced, and is a function of $\bar{c}$ and the harmonic and geometric mean of the channel gains, suggesting that the distribution of channel gains influences the optimal solution. The complexity of the iterative solution scales linearly with the number of channels. The optimal waterlevel is obtained solving an equation at each iteration, thus preventing its practical implementation in time-varying environments with a large number of channels, as is the case of the SDR cloud.

Chapter 7

Ordered Statistics Based EMPA

7.1 Introduction

In Chapter 6 a solution for the EMPA problem was derived. The solution is iterative, though, hiding important insights about how the optimal resource efficiency behaves as a function of the problem parameters. For instance, determining the average number of active channels or optimizing certain parameter requires time-consuming simulations with different channel realizations.

The main contribution of this chapter is a novel methodology for solving multi-channel power allocation problems based on ordered statistics. The methodology covers the efficiency maximization (EM) (and energy-efficient) cases but also the RM and MM problems. The methodology presented in this chapter has two main advantages:

- the computational complexity is 1-2 orders of magnitude faster than the related work, and
- it improves the understanding of the relationship between the optimal solution and the problem parameters.

Order statistics is a mathematical tool that studies the properties of ordered sequences of random variables. It is recently gaining interest in the wireless communications domain [129]. Despite being deployed in many different areas of wireless communications, this tool has never been applied for solving the water-filling problem.

The theory is briefly introduced in Appendix C. The general idea is that if a sequence of random variables $\alpha_1, \alpha_2, ..., \alpha_n$ is ordered according to their magnitude:

$$\alpha_{1:n} \leq \alpha_{2:n} \leq \alpha_{n:n}, \tag{7.1.1}$$

then $\alpha_{i:n}$ is called the ith *order statistic* [130]. Regardless of the characteristics of α_i, the ordered variables $\alpha_{i:n}$ are dependent, because of the inequality relations among them. The p.d.f. of the ith order statistic can be derived based on the distribution of the unordered sequence. If the number of random variables n is sufficiently large and they are i.i.d. with f and F the p.d.f. and c.d.f., it is possible to show [130, Ch. 9] that almost all values $\alpha_{i:n}$ have normal distribution $\mathcal{N}(m_i, \sigma_i^2)$ with

$$m_i = F^{-1}(p) \tag{7.1.2}$$

$$\sigma_i^2 = \frac{p(1-p)}{n\left[f(F^{-1}(p))\right]^2}, \tag{7.1.3}$$

where $p = i/n$.

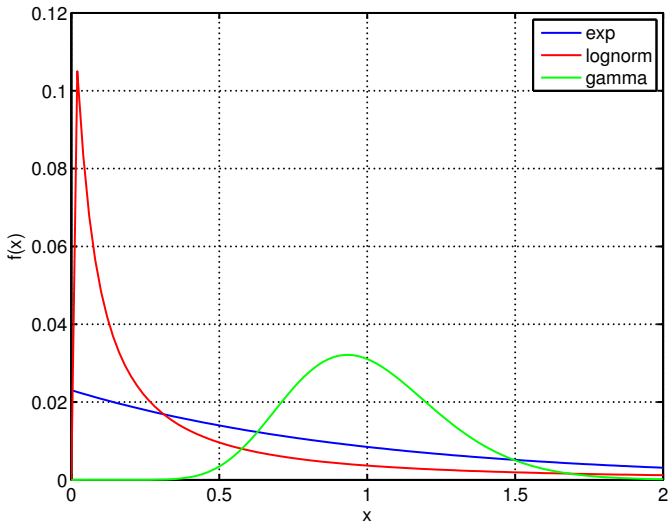

Figure 7.1: Plot of the p.d.f. of the three distributions that will be used throughout this section.

The random variables in the EMPA problem are the channel gains γ_i, for $i = 1 \dots n$. Thus, choosing the appropriate p.d.f. and c.d.f. of the channel gains is crucial in order to obtain an accurate solution. Throughout this chapter, we use the following distributions:

- **exp** is the exponential distribution with mean 1.
- **lognorm** is the log-normal distribution with mean 1 and variance 8.
- **gamma** is the Gamma distribution with mean 1 and variance 1/15.

The p.d.f. of these distributions is shown in Fig. 7.1. We have chosen these distributions and their parameters based on the following reasons:

1. *Shape*: the three distributions have different types of p.d.f. shapes. The exponential has decreasing p.d.f., the log-normal has a large peak close to the origin and heavy tails and the Gamma has a smooth bell-shape curve.
2. *Utility*: the three distributions are found in different channel models. For instance, the exponential distribution approaches the channel gains of a MD-MIMO system and the Gamma distribution models a family of distributions related to the Chi-squared, the Nakagami, Rician or Rayleigh-Rician distributions.
3. *Validation*: Since they have different characteristics but equal mean, they allow validating the theory developed throughout this section.

7.2 Related Work

It is well-known that optimal water-filling solutions are complex. This explains why they are rarely used in practical systems. Many computationally efficient algorithms have been introduced in the literature. They tackle the RM or MM problems, rather than the EM problem. [131] introduces an optimal integer-bit power allocation algorithm for discrete multi-tone modulation. The authors use a bisection method for obtaining the waterlevel. The

main contribution is using a look-up table (LUT) for obtaining the optimal rate based on the channel gain. Sub-optimal algorithms achieve greater computational complexity reduction [132, 133]. [132], for instance, achieves a complexity of $O(kn)$, with k the number of iterations (around 10) and n the number of channels. Other low-complexity algorithms are presented in [134, 135, 136].

Ordered statistics is a mathematical theory that studies the properties of sequences of ordered random variables [130]. The main advantage is that a problem of n random variables can be formulated as a function of their statistical properties, instead of their realization. [129] applies the theory to resource management in MIMO and OFDM systems. [137] uses ordered statistics to derive bounds on the capacity of delay-limited OFDM fading channels. In [138], the authors derive the marginal p.d.f. of the ordered eigenvalues of a MIMO system to derive the capacity a nx2 MIMO. The solution for larger systems, found numerically, is left as a function of the waterlevel. Ordered statistics is also used to group channels of similar gains before applying WF, reducing the problem space [139]. Other applications are found in OFDMA, simplifying the problem by assuming certain order between the channel gains different users experience [140, 141]. Dardari [142] analyses the bit error probability of using k out of the n available channels, for discrete uncoded modulation. The authors propose an empirical relation between k and the channel average SNR, which is then used to transmit a constant power on the best k channels.

All state-of-the-art algorithms have higher complexity than the solution introduced in this chapter. Our solution, based on ordered statistics, only requires a LUT access for obtaining the waterlevel. The optimal waterlevel is computed offline based on the channel statistics, rather than in their realization.

7.3 Assumptions

The parallel channel model

$$y_i = \sqrt{\gamma_i} x_i + z_i,\ i = 1 \dots n \tag{7.3.1}$$

is assumed, with x_i and y_i the transmitted and received symbols and z_i the complex noise. γ_i is the channel-to-noise ratio (CNR) of the ith channel. It is assumed that γ_i is known to the transmitter and that the channel is sufficiently slow faded so that the transmitter is able to adapt to the channel variations.

Assumption 7.1. *Assume all channels γ_i are i.i.d. and $\mathbb{E}\{\gamma_i\} = \bar{\gamma}$ for $i = 1 \dots n$. Let $F(x) \in (0,1)$, $x \in (0, +\infty)$ be the c.d.f. of the normalized random variable $\frac{\gamma_i}{\bar{\gamma}}$ so that $F(x)$ satisfies*

$$\int_0^\infty x dF = 1. \tag{7.3.2}$$

That is, the random variables described by the c.d.f. $F(x)$ have unit mean and $\bar{\gamma}$ is the mean of γ_i. The ordered sequence of channel gains

$$\gamma_{1:n} \geq \gamma_{2:n} \geq \cdots \geq \gamma_{n:n} \tag{7.3.3}$$

becomes approximately deterministic and equals

$$\gamma_{i:n} \approx \bar{\gamma} F^{-1}\left(1 - \frac{i}{n}\right) \tag{7.3.4}$$

if $n \gg 1$.

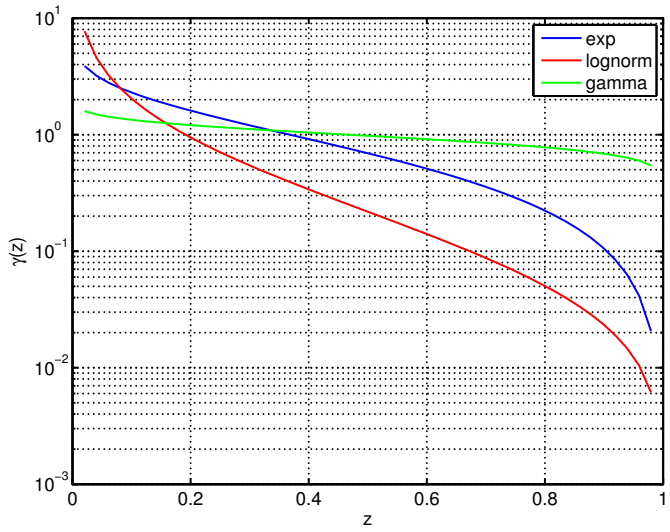

Figure 7.2: Plot of the $\gamma(z)$ function for the three distributions for $n = 100$ random variables.

Definition 7.1. *$z \in (0, 1)$ is the fraction of active channels so that $k = \lfloor zn \rfloor$ is the number of active channels.*

Then the function

$$\gamma(z) \triangleq F^{-1}(1 - z) \tag{7.3.5}$$

is the value of the $\lfloor zn \rfloor$-th strongest channel, i.e. $\gamma(i/n) = \bar{\gamma}\gamma_{i:n}$.

Assumption 7.2. *The function $\gamma(z) = F^{-1}(1 - z)$ is continuous in the interval $z \in (0, 1)$.*

Assumption 7.3. *$\gamma(z) > 0$ for $0 < z < 1$.*

Figure 7.2 plots the $\gamma(z)$ function for the three distributions for $n = 100$. The three distributions have unit mean but different variances. The slope of the $\gamma(z)$ function indicates the difference between the largest and smallest random variables. We observe that the log-normal distribution has the highest slope while the Gamma distribution is the smoothest of the three. The exponential is somewhere in between. If n random variables are drawn from these distributions, the log-normal will show the greatest variation between the highest and lowest value. As shown by the p.d.f. (Fig. 7.1) it is very probable that most of the values are small, but there is a non-zero probability that some of them are very large (the distribution is said to be heavy tailed). The Gamma distribution, on the other hand, has a bell-shape p.d.f., which implies that, with high probability, all values fall around the mean.

Assumption 7.4. *The waterlevel q' satisfies $\gamma_{n:n} \leq \frac{1}{q'} \leq \gamma_{1:n}$.*

Then, q' can be expressed as a function of $\gamma(z)$:

$$q' = \frac{1}{\bar{\gamma}\gamma(z)}. \tag{7.3.6}$$

7.4 Unconstrained Solution

The waterlevel equation (6.6.13) becomes

$$\frac{1}{\gamma(z)}\left[\sum_{i=1}^{\lfloor zn\rfloor}\log\gamma\left(\frac{i}{n}\right)-n\log\gamma(z)\right]-\frac{n}{\gamma(z)}+\sum_{i=1}^{\lfloor zn\rfloor}\frac{1}{\gamma\left(\frac{i}{n}\right)}=\bar{\gamma}c. \tag{7.4.1}$$

Note that the operator $(\cdot)^+$ has been removed, because all channels are active by definition. So far, the only approximation we have considered is the deterministic assumption of the ordered channel gains. The second approximation we take is to express the summation of the channel inverses of the $\lfloor zn\rfloor$ strongest channels as the integral:

$$h(z)\triangleq\frac{1}{n}\sum_{i=1}^{\lfloor zn\rfloor}\frac{1}{\gamma\left(\frac{i}{n}\right)}\approx\int_0^z\frac{dy}{\gamma(y)}, \tag{7.4.2}$$

which is independent of n due to the change of variable $z=i/n$. Similarly, we define the summation of the logarithms of the $\lfloor zn\rfloor$th strongest channels as the integral:

$$g(z)\triangleq\frac{1}{n}\sum_{i=1}^{\lfloor zn\rfloor}\log\gamma\left(\frac{i}{n}\right)\approx\int_0^z\log\gamma(y)dy. \tag{7.4.3}$$

Both integrals (7.4.2)-(7.4.3) exists for $z<1$ because $\gamma(z)$ is continuous and $\gamma(z)>0$. Using $h(z)$ and $g(z)$, equation (7.4.1) becomes

$$\frac{1}{\gamma(z)}\left(g(z)-z\log\gamma(z)\right)-\frac{z}{\gamma(z)}+h(z)=\bar{\gamma}\bar{c}. \tag{7.4.4}$$

We note that the left hand side of (7.4.4) is a function of z and $\gamma(z)$ only. Since we have not imposed any restriction on the p.d.f. of the channel model, we conclude that, for all channel distributions, the term $\bar{\gamma}\bar{c}$ can be separated from the functions involving the variable z in equation (7.4.4).

The solution to (7.4.4) solves the water-filling problem provided that the number of active channels k satisfies $1<k<n$.

The variable z indicates the fraction of active channels in a water-filling power allocation and the waterlevel constant through the transformation (7.3.6). It also allows compact and simple representations of different magnitudes. In the remaining of this chapter, we will use the following functions, which we denote the *z-functions*:

- $\pi(z)$ is the average transmitted power per channel normalized to $\bar{\gamma}=1$, so that $\frac{n}{\bar{\gamma}}\pi(z)$ is the total transmission power;
- $\rho(z)$ is the average rate per channel and $n\rho(z)$ is the total rate;
- $\omega(z)$ is the left hand side of the waterlevel equation (7.4.4) and
- $\eta(z)$ is the resource efficiency.

Using the results of Chapter 6, the total transmitted power is:

$$\begin{aligned}P_t&=\sum_{i=1}^{\lfloor zn\rfloor}x_i=\sum_{i=1}^{\lfloor zn\rfloor}\left(q'-\frac{1}{\gamma_{i:n}}\right)=\frac{n}{\bar{\gamma}}\left[\frac{z}{\gamma(z)}-\frac{1}{n}\sum_{i=1}^{\lfloor zn\rfloor}\frac{1}{\gamma\left(\frac{i}{n}\right)}\right]\\&=\frac{n}{\bar{\gamma}}\left[\frac{z}{\gamma(z)}-h(z)\right].\end{aligned} \tag{7.4.5}$$

If we define $\pi(z)\triangleq\frac{\bar{\gamma}}{n}P_t$, then

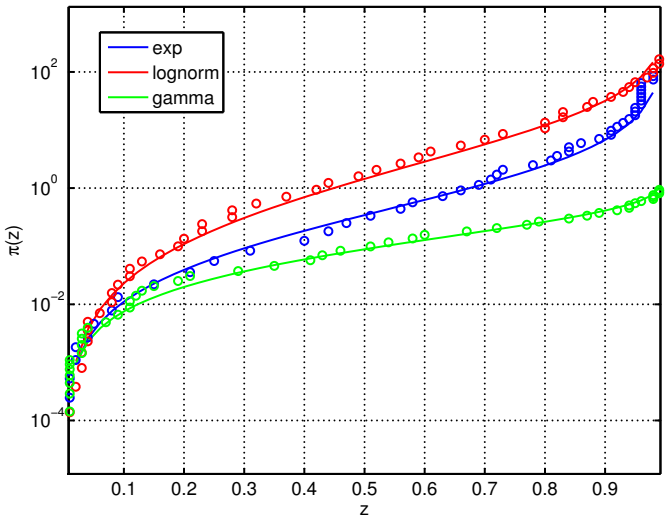

Figure 7.3: $\pi(z)$ function (line) and the average power per channel (circles) as a function of the active channels before any approximation with $n = 100$.

$$\pi(z) = \frac{z}{\gamma(z)} - h(z). \tag{7.4.6}$$

Figure 7.3 plots the function $\pi(z)$ for the three distributions. The figure compares the function $\pi(z)$ with the sum-power resulting of a water-filling power allocation as a function of the waterlevel. $\pi(z)$ only assumes a water-filling power allocation with waterlevel q' satisfying Assumption 7.4. Therefore, the predicted average sum-power also can be applied to other kind of power allocations with different objectives. The accuracy shown in the figure is reasonable and suggest that the function $\pi(z)$ can be used for fast LUT-based evaluations of water-filling power allocations.

The slope of the $\pi(z)$ function explains how the sum-power is increased by using more ordered channels. For instance, the slope of the Gamma distribution is very smooth, which indicates that using more or less channels has little effect in the total transmitted power. In the log-normal distribution, however, using more channels implies more power, because there is a high probability that the lowest channels have very low gain (see Fig. 7.1).

The total rate is

$$\begin{aligned} R_t &= \sum_{i=1}^{\lfloor zn \rfloor} \log(1 + \gamma_{i:n} x_i) = n \left[\log q' + \frac{1}{n} \sum_{i=1}^{\lfloor zn \rfloor} \log \gamma_{i:n} \right] \\ &= n \left[g(z) - z \log \gamma(z) \right]. \end{aligned} \tag{7.4.7}$$

If we define $\rho(z) \triangleq \frac{1}{n} R_t$, then

$$\rho(z) = g(z) - z \log \gamma(z). \tag{7.4.8}$$

Remark 7.1. *Let n channels be i.i.d. with c.d.f. satisfying Assumptions 7.2 and 7.3, then the sum-rate $R_t = \sum_{i=1}^{n} r_i$ is independent of the average channel gain for all waterlevel satisfying Assumption 7.4.*

This observation contrasts with the transmitted power, which scales linearly with the inverse of the average channel gain. Figure 7.4 plots the $\rho(z)$ function or average rate per

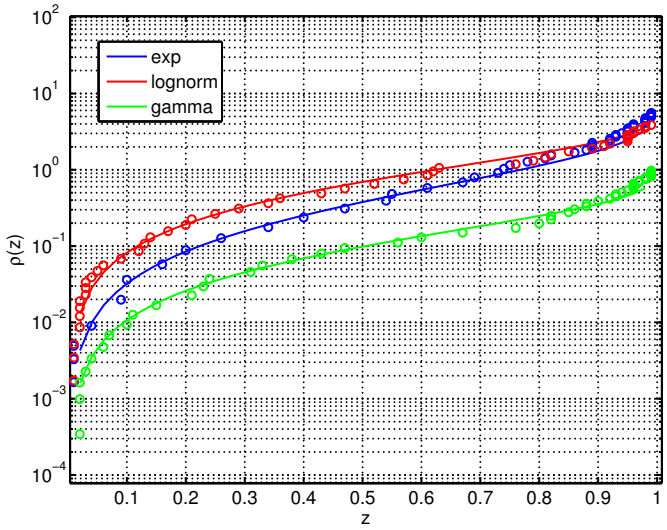

Figure 7.4: $\rho(z)$ function (line) and the average rate per channel as a function of the active channels before any approximation with $n = 100$.

channel of a water-filling power allocation before any approximation. Again, the accuracy of our approximation, and the fact that it only relies on the channels statistics, makes it useful for fast LUT-based sum-rate computation of water-filling power allocations.

The slope of $\rho(z)$ now indicates how the sum-rate increases when the waterlevel increases and more ordered channels are used. In the log-normal distribution, we observe that when z approaches 1, the rate shows very small increments, while the power ($\pi(z)$ function) in that region showed a large increment. Since $\gamma_{i:n}$ is very small if i is close to n, a lot of power is required to obtain a small rate increment.

Equation (7.4.4) can now be expressed in terms of the $\pi(z)$ and $\rho(z)$:

$$\frac{\rho(z)}{\gamma(z)} - \pi(z) = \bar{\gamma}\bar{c}. \tag{7.4.9}$$

Define

$$\omega(z) \triangleq \frac{\rho(z)}{\gamma(z)} - \pi(z), \tag{7.4.10}$$

so that the solution to the waterlevel equation, that is, the optimal fraction of active channels that maximizes the resource efficiency, z^*, is:

$$z^* = \omega^{-1}(\bar{\gamma}\bar{c}). \tag{7.4.11}$$

The intersection of the $\omega(z)$ function with the product $\bar{\gamma}\bar{c}$ gives the optimal z^* maximizing the resource efficiency. The $\omega(z)$ function is plotted in Fig. 7.5.

It is also possible to express the resource efficiency as a function of z:

$$\eta(z) = \left[c_o^r + \frac{c_p}{\bar{\gamma}} \frac{\bar{c}\bar{\gamma} + \pi(z)}{\rho(z)}\right]^{-1}. \tag{7.4.12}$$

Similarly to the rest of the z-functions, the $\eta(z)$ function is valid for LUT-based analysis of power allocation algorithms using channel statistics knowledge only. Figure 7.6 shows the accuracy of the ordered statistics approximation compared to a random channel realization. For instance, we may consider a capacity-achieving power allocation for a given channel statistics.

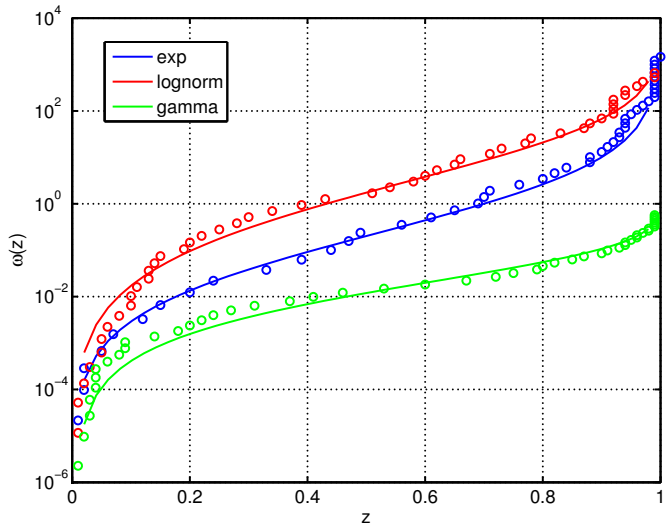

Figure 7.5: $\omega(z)$ function for different distributions. Circles are the average power as a function of the active channels before any approximation with $n = 100$.

This power allocation is determined by a waterlevel constant. From this constant, we obtain z from (7.3.6) and then evaluate the resource efficiency. We could evaluate, for instance, how the resource efficiency would vary if we add or remove channels, without simulating channel realizations.

In Fig. 7.6 we observe that each distribution, albeit having equal mean, have different optimal operating points for maximizing the resource efficiency, indicated by the maximum of the $\eta(z)$ function. The log-normal distribution operates with the smallest waterlevel or lowest number of active channels. The tail of the p.d.f. curve of the log-normal distribution indicates that there is a small probability that some channels are very high. Thus, the most efficient solution is transmitting on a small set of the best channels. The Gamma distribution, on the other hand, transmits on almost all channels. The $\gamma(z)$ function of this distribution (Fig. 7.2) showed that almost all channels have equal value. Thus, a power allocation assuming channels are Gamma distributed will transmit on either almost all channels ($z \to 1$) or only a few of them ($z \to 0$). The $\eta(z)$ function for the exponential distribution shows a smooth shape, with the maximum located in between the log-normal and Gamma distributions.

7.4.1 Properties of the z-functions

Throughout this section and unless stated the contrary, we use the term *increasing* to denote a *strictly increasing* function and *decreasing* to denote a *strictly decreasing* function.

Let us recall the properties of $\gamma(z) \triangleq F^{-1}(1-z)$, given any c.d.f. $F(x)$

- $\gamma(z)$ is increasing with z in its domain $z \in (0,1)$,
- The partial derivative is $\frac{d}{dz}\gamma(z) = -\left[f\left(F^{-1}(1-z)\right)\right]^{-1}$ and exists if $f(x) > 0$ for $x > 0$, and
- $\lim_{z\to 1}\gamma(z) = +\infty$ and $\lim_{z\to 0}\gamma(z) = 0$.

Now we can derive the properties of the z-functions. The first derivatives of $\pi(z)$, $\rho(z)$, $\omega(z)$ and $\eta(z)$ are:

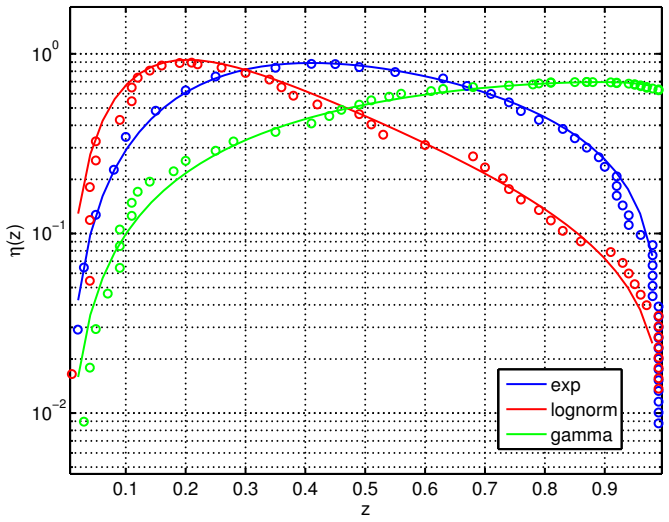

Figure 7.6: $\eta(z)$ function for different distributions. Circles are the average power as a function of the active channels before any approximation with $n = 100$.

$$
\begin{aligned}
\pi'(z) &= -z\frac{\gamma'(z)}{\gamma^2(z)} \\
\rho'(z) &= -z\frac{\gamma'(z)}{\gamma(z)} \\
\omega'(z) &= -\rho(z)\frac{\gamma'(z)}{\gamma^2(z)} \\
\eta'(z) &= -\left[c_o^r + \frac{c_p}{\bar{\gamma}}\frac{\bar{c}\bar{\gamma}+\pi(z)}{\rho(z)}\right]^{-2}\frac{c_p}{\bar{\gamma}}\frac{\pi'(z)\rho(z)-\rho'(z)\left(\pi(z)+\bar{\gamma}\bar{c}\right)}{\rho^2(z)} \\
&= -\left[c_o^r + \frac{c_p}{\bar{\gamma}}\frac{\bar{c}\bar{\gamma}+\pi(z)}{\rho(z)}\right]^{-2}\rho^{-2}(z)\frac{c_p}{\bar{\gamma}}\rho'(z)\left[\omega(z)-\bar{\gamma}\bar{c}\right].
\end{aligned}
\tag{7.4.13}
$$

All the first derivatives are proportional to $-\gamma'(z)$. This explains the fact that the slope of the c.d.f. directly influences the slope of the z-functions.

The $\pi(z)$, $\rho(z)$ and $\omega(z)$ satisfy the following common properties:

1. They are increasing and continuous for $z \in (0,1)$,
2. the limit for $z \to 0$ is 0,
3. the limit for $z \to 1$ is $+\infty$.

The properties of the $\omega^{-1}(y)$ function are easily derived using the inverse function theorem,

$$
\left(f^{-1}\right)'(x) = \frac{1}{f'\left(f^{-1}(x)\right)} \tag{7.4.14}
$$

and the fact that $\omega^{-1}(y) \in (0,1)$ and $\omega'(z) > 0$ for $z \in (0,1)$, it is straightforward to show that

1. $\omega^{-1}(y)$ is increasing and continuous for $y > 0$,
2. $\lim_{y\to 0}\omega^{-1}(y) = 0$,

3. and $\lim_{y\to+\infty} \omega^{-1}(y) = 1$,.

The properties of the $\omega^{-1}(\bar{\gamma}\bar{c})$ function indicate the behaviour of the optimal solution with respect to the problem parameters: the optimal fraction of active channels z^* is increasing with the factor $\bar{\gamma}\bar{c}$ ($\bar{c} = c/n$ and $c = (c_o^0 + c_o^n n)/c_p$). The $\rho(z)$ and $\pi(z)$ functions are increasing with z, implying that the sum-rate and sum-power of the resource-efficient power allocation is an increasing function of the factor $\bar{\gamma}\bar{c}$. This is consistent with the case $n = 1$ derived in Section 6.4. In the case $n = 1$, the optimal rate is an increasing function of $\gamma\frac{c_o}{c_p}$.

The constant c relates the operating costs to the power costs. If the power costs are very high, the solution reduces z^*, which means less rate. The infrastructure (e.g. computing) will be acquired for a longer period for transmitting the same amount of bits, (higher costs in total), albeit this extra cost will be compensated by a low transmission power. Small z^* also means transmitting on the best channels only. Due to the fact that the channels are ordered, if the power costs are high compared to the operation costs or there are many channels (large n), the power required to transmit on the worse channels does not compensate for the additional rate they can provide. Hence, the most efficient solution is transmitting on the best ones only.

On the other hand, if the average channel gain $\bar{\gamma}$ is very high, an efficient solution is $z^* \to 1$, which means high rates. The infrastructure is acquired for shorter time for transmitting the same amount of bits. Because the average channel gain is high, high rates do not imply high transmission power. What determines the rate, power or efficiency, is the ratio $\bar{\gamma}\bar{c}$. Thus, high power costs can be counterbalanced by high channel gain, and vice versa.

If the operational costs are null or negligible, $c_o + c_o^n n \to 0$, the efficient solution is minimizing the power subject to the minimum rate constraint, that is the MM power allocation. Conversely, if the power costs are null or negligible, $c_p \to 0$, the efficient solution is maximizing the sum-rate subject to the maximum power constraint, or RM power allocation (see Section 6.6).

The $\eta(z)$ function, on the other hand

1. is continuous for $z \in (0, 1)$,
2. increasing for $z < \omega^{-1}(\bar{\gamma}\bar{c})$ and decreasing for $z > \omega^{-1}(\bar{\gamma}\bar{c})$,
3. $\lim_{z\to 0} \eta(z) = 0$, and
4. $\lim_{z\to 1} \eta(z) = 0$.

The second property indicates the maximum resource efficiency or the main problem solution.

7.4.2 Optimal Resource Efficiency

In Section 6.6.3, we observed that when the optimal waterlevel is chosen, the resulting resource efficiency is

$$\eta^* = q^* = \frac{1}{c_o^r + c_p q'}. \tag{7.4.15}$$

Using z^*, it is possible to express the optimal resource efficiency as a function of $\bar{\gamma}$ and $\bar{c}$:

$$\eta^*(\bar{\gamma}, \bar{c}) = \left[c_o^r + \frac{c_p}{\bar{\gamma}\gamma(z^*(\bar{\gamma}\bar{c}))}\right]^{-1}, \tag{7.4.16}$$

where the term $z^*(\bar{\gamma}\bar{c}) = \omega^{-1}(\bar{\gamma}\bar{c})$.

The function $\eta^*(\bar{\gamma}, \bar{c})$ is continuous for $z^* > 0$, which implies $\bar{\gamma}\bar{c} > 0$. The range of the function $\eta^*(\bar{\gamma}, \bar{c})$ is $(0, 1/c_o^r)$. In the remaining of this subsection, we study the behaviour of the optimal resource efficiency as a function of the different problem parameters, showing that it is analogous to the case $n = 1$ studied in Section 6.4. More precisely, we will demonstrate that the optimal resource efficiency is:

- increasing with the average CNR $\bar{\gamma}$,
- decreasing with the power costs c_p and
- decreasing with the normalized operational to power costs ratio $\bar{c} = \frac{c_o^0 + c_o^n n}{n c_p}$.

Proposition 7.1. *Given any constant $a > 0$, $\eta^*(\bar{\gamma}, a)$ is monotonically non-decreasing and bounded in the interval $(0, 1/c_o^r)$.*

Proof. We will show that, given any constant $a > 0$, the constant $1/q'$ is a monotonically non-decreasing function of $\bar{\gamma}$, taking values in the interval $(0, \infty)$. The proof follows straightforward.

Let $\bar{c} = a$, then according to (7.4.11)

$$\bar{\gamma} = \frac{\omega(z^*)}{a}. \tag{7.4.17}$$

Using (7.3.6) we have that:

$$\frac{1}{q'(z^*)} = \frac{\omega(z^*)}{a}\gamma(z^*). \tag{7.4.18}$$

Now $1/q'(z^*)$ is a composite function of the form $t(s(x))$ where $s(x) = \omega^{-1}(x)$ and $t(y) = \omega(y)\gamma(y)$. $\frac{d}{dx}g(x) > 0$ for all $x > 0$ (see section 7.4.1). The first derivative of $t(y)$ is

$$\frac{d}{dy}\omega(y)\gamma(y) = -\rho(y)\frac{\frac{d}{dy}\gamma(y)}{\gamma(y)} \geq 0 \tag{7.4.19}$$

because any c.d.f. $F(x)$ is monotone non-decreasing then $\frac{d}{dy}\gamma(y) \leq 0$. By the composition rule, $1/q'$ is a monotonically non-decreasing function of $\bar{\gamma}$.

□

Proposition 7.2. *Given any constants $a > 0$, $b > 0$, $\eta^*\left(a, \frac{b}{c_p}\right)$ is monotonically non-increasing and bounded in the interval $(0, 1/c_o^r)$.*

Proof. We will show that the function $s(c_p) = \frac{a\gamma\left(\omega^{-1}(ab/c_p)\right)}{c_p}$ is non-increasing and $\lim_{c_p \to \infty} s(c_p) = +\infty$, then it follows directly that $\eta^*\left(a, \frac{b}{c_p}\right) = [c_o^r + 1/s(c_p)]^{-1}$ is non-increasing in the interval $(0, 1/c_o^r)$.

We proceed similarly to the previous proof. From (7.4.11) it follows that

$$c_p = \frac{ba}{\omega(z^*)}. \tag{7.4.20}$$

then the composite function

$$s(z^*(c_p)) = \frac{1}{b}\omega(z^*)\gamma(z^*). \tag{7.4.21}$$

Since $\frac{d}{dc_p}z^*(c_p) \leq 0$ (see section 7.4.1) and

$$\frac{d}{d_z^*}s(z^*) = -\rho(y)\frac{\frac{d}{dy}\gamma(y)}{\gamma(y)} \geq 0 \tag{7.4.22}$$

it follows that the composite function satisfies $\frac{d}{dc_p}s(z^*(c_p)) \leq 0$ which completes the proof. □

Proposition 7.3. *Given any constant $a > 0$, $\eta^*(a, \bar{c})$ is strictly decreasing and bounded in the interval $(0, 1/c_o^r)$.*

Proof. $\omega^{-1}(x)$ is strictly increasing with x and $\gamma(z)$ is strictly decreasing with z (see section 7.4.1). It follows from the composition rule that, for any constant $a > 0$, $\eta^*(a, \bar{c})$ is strictly decreasing with $\bar{c} \geq 0$ and since

$$\lim_{\bar{c}\to\infty} \gamma\left(\omega^{-1}(a\bar{c})\right) = 0, \tag{7.4.23}$$

then $\lim_{\bar{c}\to\infty} \eta^*(a, \bar{c}) = 0$. □

According to Proposition 7.3, if n increases, $\bar{c} = \frac{c_o + c_o^n n}{c_p n}$ decreases and the resource efficiency increases. The convergence limit depends on the relation between factors c_o^n and c_o^0:

$$\lim_{n\to\infty} \bar{c} = \begin{cases} 0 & \text{if } c_o^n \approx 0 \\ \frac{c_o^n}{c_p} & \text{otherwise.} \end{cases} \tag{7.4.24}$$

If the cost per extra independent channel is negligible, the optimal strategy is transmitting only on a subset of the best channels ($z^* = \omega^{-1}(0) = 0$). When the cost of using more channels is non-zero, $z^* = \omega^{-1}(c_o^n/c_p) > 0$.

Using the optimal resource efficiency function (7.4.16), we have that for $n \to \infty$:

$$\lim_{n\to\infty} \eta^*(\bar{\gamma}, \bar{c}) = \begin{cases} 1/c_o^r & \text{if } c_o^n \approx 0 \\ \left[c_o^r + \frac{c_p}{\bar{\gamma}\gamma(z^*)}\right]^{-1} & \text{otherwise,} \end{cases} \tag{7.4.25}$$

with $z^* = \omega^{-1}(c_o^n/c_p)$.

That is, the resource efficiency of a multi-channel system increases to the maximum $(1/c_o^r)$ with n, if the cost associated to an extra channel is zero or negligible. Otherwise, the asymptotic resource efficiency converges to a lower value, as shown in Fig. 7.7. In the figure, we have assumed $c_o^r = 0$ and the case $c_o^n = 0$ diverges to infinity.

This conclusion is consistent with the capacity-achieving power allocation at low SNR per dimension [65, Section 5.4]. If n increases and the average SNR remains constant, the best strategy is using only the best channels. For independent channel gains, it is very likely that the maximum is very high as the sample size increases. Order statistics theory indicates that the expected value of the strongest channel increases as $F^{-1}(1 - 1/n)$, which is an increasing function of n.

In the EM power allocation, the number of degrees of freedom are exploited according to the ratio of factors c_o^0, c_o^n to c_p. The practical consequences are that if a system is employing a large number of degrees of freedom (e.g. an SDR Cloud), the most determining factor that influences the power allocation is the ratio of the cost per degree of freedom to the power cost, or c_o^n/c_p. These are the parameters to be considered for maximizing the resource efficiency. The rest of the costs as defined in our linear model have little influence when $n \gg 1$, because the constant costs are amortized by transmitting over more channels (which implies higher sum-rates).

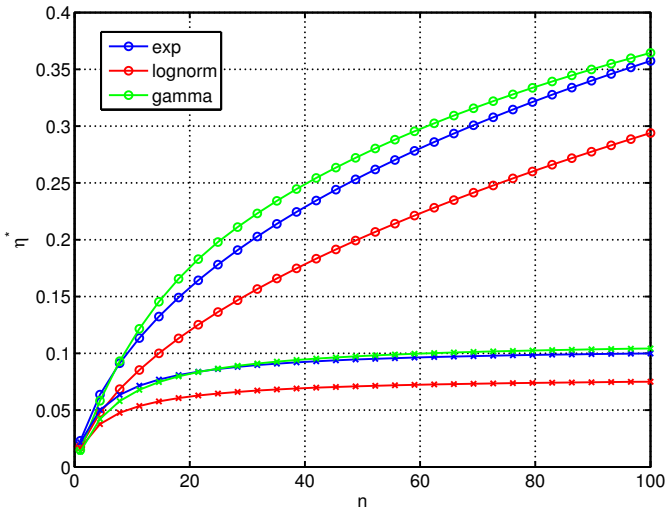

Figure 7.7: Resource efficiency as a function of the number of independent channels n considering that $c_o^r = 0$ and $c_p = 1$. The curves with circles are for $c_o^n = 0$. The curves with crosses are for $c_o^n = 0.1$.

7.4.3 Extreme Cases

The solution $z^* = \omega^{-1}(\bar{\gamma}\bar{c})$ assumes that the number of active channels satisfies $1 < k < n$. In this section, we discuss the solutions for the cases $k = 1$ and $k = n$.

Case $k = 1$.

The properties of the $\omega^{-1}(\bar{\gamma}\bar{c})$ function indicate that if $\bar{\gamma}\bar{c} \to 0$, then $z^* \to 0$. The case $k = 1$ occurs when $z^* \leq 1/n$. The threshold

$$\bar{\gamma}\bar{c} \leq \omega\left(\frac{1}{n}\right) \tag{7.4.26}$$

determines when the optimal solution for maximizing the resource efficiency is transmitting on the best channel only. Fig. 7.8 plots the $\omega(1/n)$ function. The figure is interpreted as follows: given n, the set of $\bar{\gamma}\bar{c}$ values below the curve correspond to the solution $k = 1$. For instance, if $n = 100$, the Gamma, log-normal and exponential distributions will use $k = 1$ channels if $\bar{\gamma}\bar{c} < 2 \cdot 10^{-6}$, $\bar{\gamma}\bar{c} < 5 \cdot 10^{-5}$ and $\bar{\gamma}\bar{c} < 2 \cdot 10^{-4}$, respectively.

Figure 7.8 shows that for the Gamma distribution, $\omega(1/n)$ decays very quickly, which indicates that for large n, the solution $k = 1$ will rarely be the optimal. Indeed, for $n > 150$, $k > 1$ for any $\bar{\gamma}\bar{c}$. For the log-normal and exponential distribution, on the contrary, the solution $k = 1$ is optimal if $\bar{\gamma}\bar{c}$ is relatively small, even if n is large.

Let $\gamma_{1:n}$ be the largest channel gain (i.e. the first ordered statistic), the power allocation can be formulated as the power adaptation problem described in Section 6.4, with $n = 1$ and $\gamma = \gamma_{1:n}$. Since n channels are being reserved despite only one is being used, the cost c_o^n is multiplied by the total number of channels n. For clarity, we will assume that $c_o^r = 0$. The optimal transmission rate R_0^* to allocate to the best channel is the solution to:

$$\begin{aligned} \min_{R_0} \quad & -\frac{R_0}{c_o^0 + c_o^n n + c_p \frac{e^{R_0}-1}{\gamma_{1:n}}} \\ \text{s.t.} \quad & R_0 \geq R_{\min} \\ & R_0 \leq \log\left(1 + P_{\max}\gamma_{1:n}\right). \end{aligned} \tag{7.4.27}$$

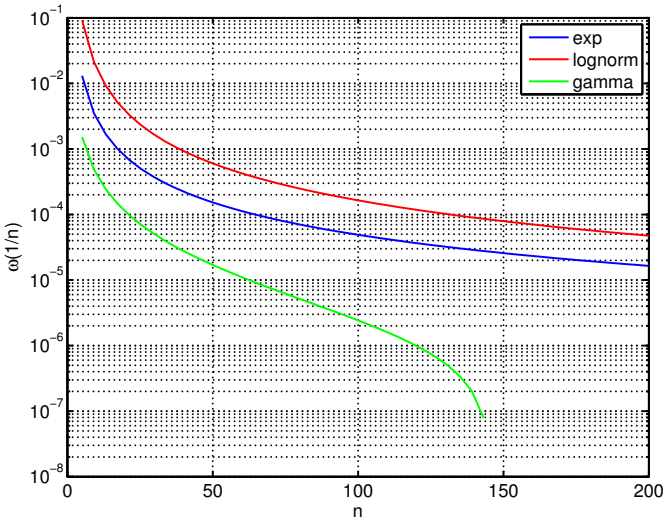

Figure 7.8: Plot of $\omega\left(\frac{1}{n}\right)$ for the three distributions.

The optimal rate was derived in Section 6.4:

$$R_0^* = \left[1 + W\left(\left(\gamma_{1:n}c - 1\right)e^{-1}\right)\right]_{R_{\min}}^{R_{\max}}, \tag{7.4.28}$$

with $R_{\max} = \log(1 + P_{\max}\gamma_{1:n})$ and $c = \frac{c_o^0 + c_o^n n}{c_p}$. In this case, the solution is in closed form without using ordered statistics theory. The first order statistic does not follow the Gaussian approximation of the quantiles and its variance is large (see Appendix C). Thus, using the measured channel gain $\gamma_{1:n}$ instead of the expected value $\gamma(1/n)$ gives more accurate results.

On the other hand, ordered statistics allow computing the average transmission rate:

$$\begin{aligned}\mathbb{E}\left\{R_0^*\right\} &= 1 + \mathbb{E}_{\gamma_{1:n}}\left\{W\left(\left(\gamma_{1:n}c - 1\right)e^{-1}\right)\right\}, \\ &= 1 + \int_0^\infty W\left(\left(yc - 1\right)e^{-1}\right) dF^{(p)}(y)dy\end{aligned} \tag{7.4.29}$$

where $p = 1, 2, 3$ is one of the limiting distributions of the extreme value as described in Appendix C. The expected value of the strongest channel is $\mathbb{E}\{\gamma_{1:n}\} = \bar{\gamma}\gamma(1/n)$. The main branch of the $W(x)$ function is concave, hence we can apply the Jensen's inequality to get an upper bound on the expected rate:

$$\mathbb{E}\left\{R_0^*\right\} \leq 1 + W\left(\left(\gamma(1/n)\bar{\gamma}c - 1\right)e^{-1}\right), \tag{7.4.30}$$

The average optimal resource efficiency can also be upper bounded using the Jensen's inequality and (6.4.14)

$$\mathbb{E}\left\{\eta_0^*\right\} \leq \frac{1}{c_p}\frac{W\left(\left(\gamma(1/n)\bar{\gamma}c - 1\right)e^{-1}\right)}{c - \frac{1}{\bar{\gamma}\gamma(1/n)}}. \tag{7.4.31}$$

Case $k = n$.

It follows from the properties of the inverse of $\omega^{-1}(x)$ that $z \to 1$ for increasing $\bar{\gamma}\bar{c}$. Therefore, if

$$\bar{\gamma}\bar{c} \geq \omega\left(1 - \frac{1}{n}\right), \tag{7.4.32}$$

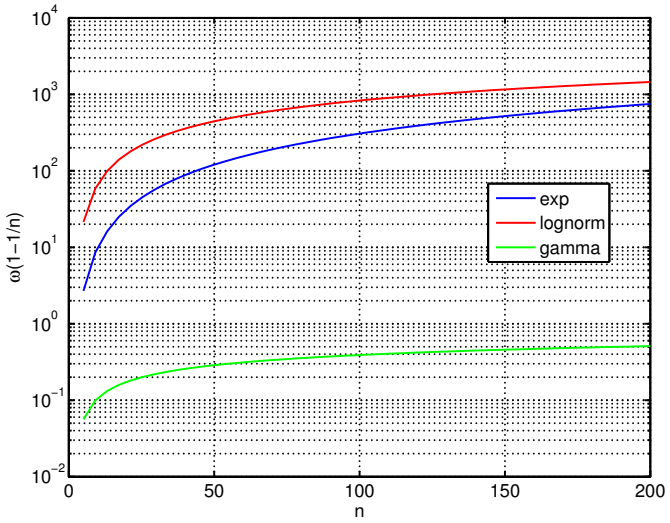

Figure 7.9: Plot of $\omega\left(1-\frac{1}{n}\right)$ for the three distributions

the power allocation solution assigns non-zero power to all channels. Contrary to the case $k=1$, the case $k=n$ occurs frequently if the SNR is high, for instance, when the mobile terminal is close to the base station.

The threshold $\bar{\gamma}\bar{c}$ decreases with the number of channels. Intuitively, the larger the number of channels, the more likely some of them have poor quality. Then, it is efficient to use all channels only if, either the average gain or the costs, are very high. On the other hand, distributions with a smooth p.d.f. require lower average gain, because the shorter difference between the largest and smallest channel gains. Figure 7.9 plots the $\omega(1-1/n)$ function. The figure is interpreted as follows. Given n, the set of $\bar{\gamma}\bar{c}$ values above the curve imply that the optimal solution is $k=n$. Thus, it is observed that for the Gamma distribution, which has most of the values concentrated around the mean, the threshold that determines when $k=n$ is optimal is very low for all range of n values. The log-normal distribution has the highest threshold, because the difference between the highest and lowest channel gain is larger.

In Section 6.6.3, we already derived the optimal waterlevel when all channels where assumed to be active. Let $\hat{h}$ denote the inverse of the harmonic mean of the channels and $\log(\hat{g})$ their geometric mean, the waterlevel q' becomes

$$q' = \frac{\bar{c}-\hat{h}}{W\left[\left(\bar{c}-\hat{h}\right)e^{\hat{g}-1}\right]}. \tag{7.4.33}$$

Using ordered statistics, we can obtain the waterlevel without knowledge of the channel realization, using the fact that $h(1)\approx\frac{\hat{h}}{\bar{\gamma}}$ and $g(1)\approx\hat{g}+\log\bar{\gamma}$, (7.4.33) becomes

$$q' = \frac{\bar{c}\bar{\gamma}-h(1)}{W\left[\left(\bar{c}\bar{\gamma}-h(1)\right)e^{g(1)-1}\right]}. \tag{7.4.34}$$

The average optimal resource efficiency η_1^* can be estimated using ordered statistics, too:

$$\eta_1^* = \frac{\bar{\gamma}}{c_p}\frac{W\left[\left(\bar{c}\bar{\gamma}-h(1)\right)e^{g(1)-1}\right]}{\bar{c}\bar{\gamma}-h(1)}. \tag{7.4.35}$$

7.5 Constrained Solution

In Section 6.6 we showed that if the unconstrained solution violates the maximum power or minimum rate constraints, the EM power allocation equals the RM and MM power allocations, respectively. In this section, we show that these solutions can also be obtained without knowledge of the channel realization, using ordered statistics. We also propose a sub-optimal solution for the case when the peak power constraint is violated.

7.5.1 Maximum Power

Given $z' = \omega^{-1}(\bar{\gamma}\bar{c})$ a maximizer of the resource efficiency, if

$$\frac{n}{\bar{\gamma}}\pi(z') > P_{\max}, \tag{7.5.1}$$

the maximum power constraint is violated and the optimal z^* that maximizes the resource efficiency subject to the maximum power constraint $P_{\max}$ is

$$z^*_{\mathrm{MR}} = \pi^{-1}\left(\frac{\bar{\gamma}}{n}P_{\max}\right). \tag{7.5.2}$$

Remark 7.2. *z^*_{MR} is the solution to the rate maximization water-filling problem under the assumptions of Section 7.3 and the maximum sum-rate is:*

$$R_t = n\rho(z^*_{MR}). \tag{7.5.3}$$

The function $\pi(z)$ has the same properties than the $\omega(z)$ function. Hence z^*_{MR} satisfies the same properties than the unconstrained solution. For $\frac{\bar{\gamma}}{n}P_{\max} \to 0$, $z^*_{\mathrm{MR}} \to 0$ and for $\frac{\bar{\gamma}}{n}P_{\max} \to \infty$, $z^*_{\mathrm{MR}} \to 1$. Therefore, the extreme cases $k = 1$ and $k = n$ apply equivalently.

If

$$\frac{\bar{\gamma}}{n}P_{\max} \leq \pi\left(\frac{1}{n}\right) \tag{7.5.4}$$

only the channel with the maximum gain is used and the transmission rate is:

$$R^*_0 = \log\left(1 + \bar{\gamma}P_{\max}\gamma\left(\frac{1}{n}\right)\right), \tag{7.5.5}$$

and the optimal resource efficiency, assuming $c^r_o = 0$ is

$$\eta^*(P_{\max}) = \frac{1}{c_p}\frac{\log\left(1 + \bar{\gamma}P_{\max}\gamma\left(\frac{1}{n}\right)\right)}{c + P_{\max}}. \tag{7.5.6}$$

On the other hand, if

$$\frac{\bar{\gamma}}{n}P_{\max} \geq \pi\left(1 - \frac{1}{n}\right) \tag{7.5.7}$$

all channels are used and the power allocation is water-filling $x^*_i = \left(q' - \gamma_i^{-1}\right)^+$ with the waterlevel satisfying the maximum power constraint, that is:

$$q' = \frac{P_{\max}}{n} + \frac{h(1)}{\bar{\gamma}}, \tag{7.5.8}$$

and the achieved rate is

$$R_t = n\left[\log\left(\frac{P_{\max}}{n} + \frac{h(1)}{\bar{\gamma}}\right) + \log(\bar{\gamma}) + g(1)\right]. \tag{7.5.9}$$

The optimal resource efficiency when all channels are active, assuming $c^r_o = 0$ is

$$\eta^*(P_{\max}) = \frac{1}{c_p}\frac{\left[\log\left(P_{\max}/n + h(1)/\bar{\gamma}\right) + \log(\bar{\gamma}) + g(1)\right]}{\bar{c} + P_{\max}}. \tag{7.5.10}$$

7.5.2 Minimum Rate

Given $z' = \omega^{-1}\left(\bar{\lambda}\bar{c}\right)$ a maximizer of the resource efficiency, if

$$n\rho(z') < R_{\min}, \tag{7.5.11}$$

the minimum rate constraint is violated and the optimal z^* that maximizes the resource efficiency subject to the minimum rate constraint $R_{\min}$ is the margin maximization solution:

$$z^*_{\mathrm{MM}} = \rho^{-1}\left(\frac{R_{\min}}{n}\right). \tag{7.5.12}$$

Remark 7.3. *z^*_{MM} is the solution to the margin maximization water-filling problem under the assumptions of Section 7.3 and the minimum sum-power is:*

$$P_t = \frac{n}{\bar{\gamma}}\pi(z^*_{MM}). \tag{7.5.13}$$

The function $\rho(z)$ has the same properties of function $\omega(z)$. Hence z^*_{MM} satisfies the same properties than the unconstrained solution. For $\frac{R_{\min}}{n} \to 0$, $z^*_{\mathrm{MM}} \to 0$ and for $\frac{R_{\min}}{n} \to \infty$, $z^*_{\mathrm{MM}} \to 1$. Therefore, the extreme cases $k = 1$ and $k = n$ apply equivalently.

If

$$\frac{R_{\min}}{n} \leq \rho\left(\frac{1}{n}\right) \tag{7.5.14}$$

only the channel with the maximum gain is used and the transmission rate is:

$$R^*_0 = R_{\min}, \tag{7.5.15}$$

and the optimal resource efficiency, assuming $c^r_o = 0$ is

$$\eta^*\left(R_{\min}\right) = \frac{1}{c_p}\frac{R_{\min}}{c + \frac{\exp(R_{\min})-1}{\bar{\gamma}\gamma(1/n)}}. \tag{7.5.16}$$

On the other hand, if

$$\frac{R_{\min}}{n} \geq \rho\left(1 - \frac{1}{n}\right) \tag{7.5.17}$$

all channels are used and the power allocation is water-filling $x^*_i = \left(q' - \gamma_i^{-1}\right)^+$ with the waterlevel satisfying the maximum power constraint, that is:

$$q' = \frac{\exp\left(\frac{R_{\min}}{n} - g(1)\right)}{\sqrt[n]{\bar{\gamma}}}, \tag{7.5.18}$$

and the optimal resource efficiency when all channels are active, assuming $c^r_o = 0$ is

$$\eta^*\left(R_{\min}\right) = \frac{1}{c_p}\frac{R_{\min}}{c + \frac{\exp\left(\frac{R_{\min}}{n} - g(1)\right)}{\sqrt[n]{\bar{\gamma}}} - \frac{h(1)}{\bar{\gamma}}}. \tag{7.5.19}$$

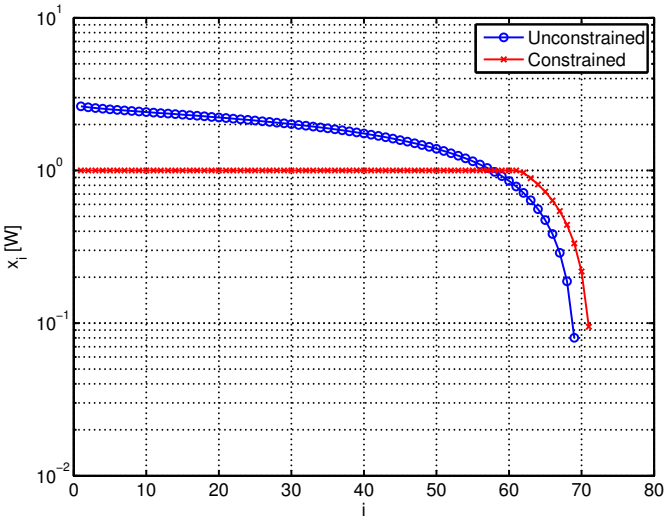

Figure 7.10: Comparison of the power allocation of the unconstrained and constrained solutions for $n = 100$ i.i.d. exponential channels with $\bar{\gamma} = 1$ and $P_{\text{peak}} = 1$ W. The channels are ordered and x_i is the power allocated to the ith largest channel.

7.5.3 Maximum Peak Power

When the peak power constraint is violated, the optimal waterlevel raises and the optimal solution transmits on more channels. Figure 7.10 compares the solution of a constrained and unconstrained power allocation. Because the power allocated to some channels is limited, the constrained solution uses more channels to compensate for the loss of power. That is, if z^*_{co} is the optimal fraction of channels of the constrained solution and z^*_{un} is the optimal fraction of channels of the unconstrained solution, it holds that

$$z^*_{\text{un}} \leq z^*_{\text{co}}, \tag{7.5.20}$$

which follows from the stationary condition of the constrained optimization problem, equation (6.6.8).

A simple but effective approximation consists on assuming $z^*_{\text{co}} = z^*_{\text{un}}$ and applying the peak power constraint to each channel. That is, if x'_i is the unconstrained power solution, $x^*_i = [x'_i]^{P_{\text{peak}}}$. This solution has a loss of efficiency less than 10% for $P_{\text{peak}} > 250$ mW, as shown in Fig. 7.11. The figure fixes $\bar{\gamma} = 1$ and plots the gap to the optimal efficiency as a function of $\bar{c} = c/n$. Note that since z^* is a function of the ratio $\bar{\gamma}\bar{c}$, fixing $\bar{c}$ and varying $\bar{\gamma}$ shows identical performance.

The gap to the optimal efficiency is computed as the ratio between the ordered statistics solution assuming $z^*_{\text{co}} = z^*_{\text{un}}$ and the exact iterative solution as described in Section 6.6, applying the constraint is applied to the vector of powers at each iteration.

The highest loss occurs for small $\bar{c}$, where only a few channels are active. In these cases, the iterative constrained solution allocates power to more channels than the unconstrained one. The relative difference is higher when only a few channels are active, than where many channels are. When the factor $\bar{\gamma}\bar{c}$ increases, even a small peak power constraint has negligible effect to the final resource efficiency.

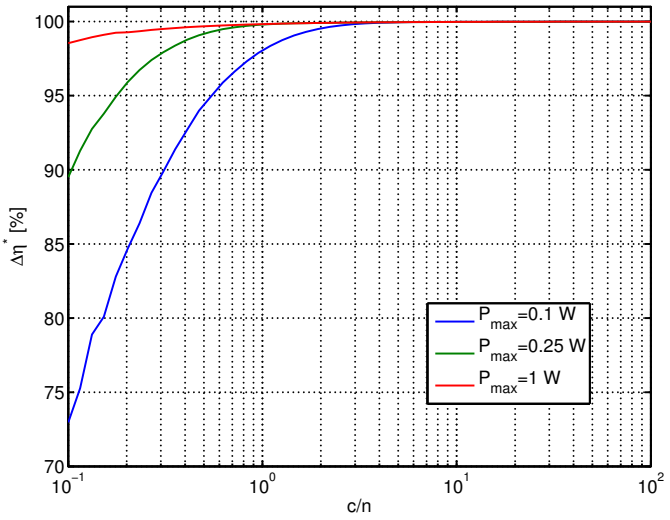

Figure 7.11: Gap to the optimal resource efficiency assuming $z^*_{\text{co}} = z^*_{\text{un}}$ and $x^*_i = [x'_i]^{P_{\text{peak}}}$ with x'_i the solution to the unconstrained optimization problem, for the exponential distribution. The gap $\Delta\eta^*$ is computed as the ratio of the constrained to the unconstrained efficiency.

7.6 Numerical Evaluation

We have presented a sub-optimal solution to the water-filling power allocation problem based on channel statistics rather than the channel realization. In this section, we asses the performance of the proposed solution numerically, with respect to the optimal water-filling power allocation and related work.

From now over, $\Delta\eta^*$ in the y-axis of the figures represents the gap to the optimal solution, computed iteratively without using ordered statistics, as described in Section 6.6. The gap is given in % and is computed as follows

$$\Delta\eta^* = 100 \cdot \frac{\eta^*_{\text{ordered}}}{\eta^*_{\text{iterative}}} \ [\%]. \tag{7.6.1}$$

For instance, if $\Delta\eta^* = 90\%$ means that the efficiency is 90% the efficiency obtained with the iterative solution.

7.6.1 Gap to the Optimal water-filling

Figures 7.12-7.14 plot the difference between the sub-optimal ordered statistics solution, i.e. $z^* = \omega^{-1}(\bar{\gamma}\bar{c})$ with respect to the optimal (iterative) solution for $n = 100$ channels. The results have been computed by fixing $\bar{\gamma} = 1$ and varying $\bar{c}$. Note that varying $\bar{\gamma}$ and fixing $\bar{c}$ gives identical results. The three figures also plot the solutions $k = 1$ and $k = n$.

The results show that the range of $\bar{\gamma}\bar{c}$ values where the ordered statistics solution is accurate enough is wider for the exponential and log-normal distributions than for the Gamma distribution. The explanation is that when channels are Gamma distributed, the optimal solution either chooses $k = 1$ or $k = n$ with high probability, violating the ordered statistics assumptions. The combination of the three solutions, however, covers all range of values with an efficiency higher than 91%, 82% and 96% of the optimal, for the exponential, log-normal and Gamma distributions, respectively. Furthermore, equations (7.4.26) and (7.4.32) indicate, analytically, which of the three solutions need to be selected.

From Fig. 7.8 we observe that if we let $\bar{\gamma} = 1$, $n = 100$, the solution $k = 1$ is optimal for the

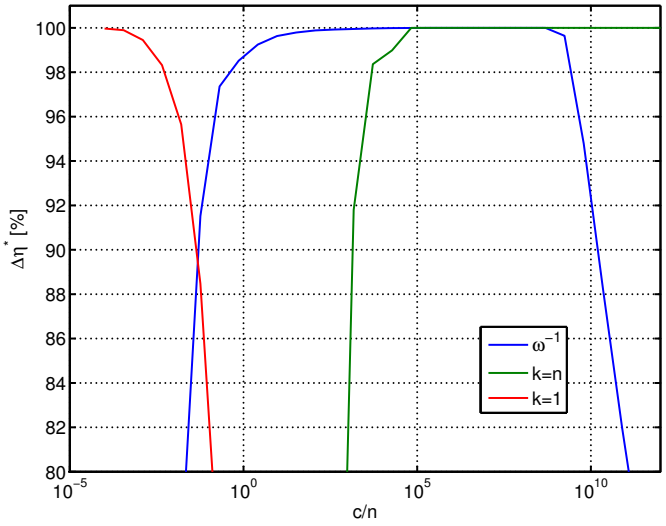

Figure 7.12: Gap to the optimal water-filling solution of the ordered statistics solution as a function of $\bar{c} = c/n$ for $n = 100$ and exponentially distributed channels.

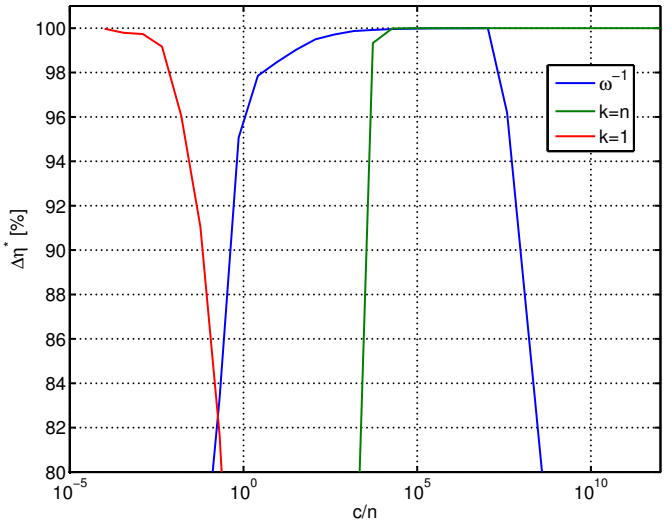

Figure 7.13: Gap to the optimal water-filling solution of the ordered statistics solution as a function of $\bar{c} = c/n$ for $n = 100$ and log-normal distributed channels.

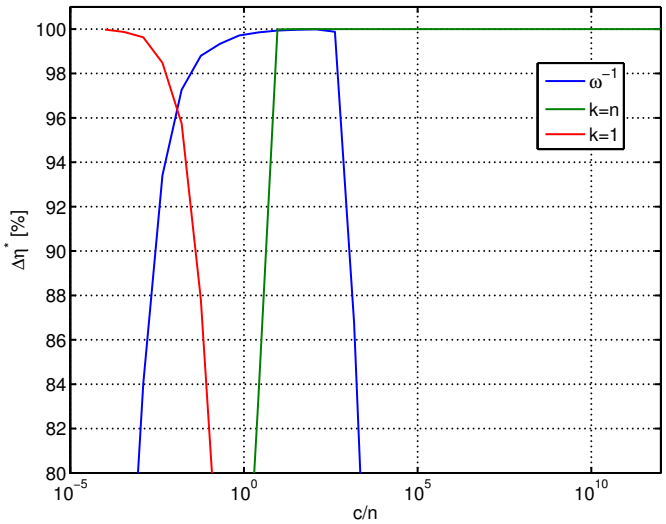

Figure 7.14: Gap to the optimal water-filling solution of the ordered statistics solution as a function of $\bar{c} = c/n$ for $n = 100$ and Gamma distributed channels.

Gamma, log-normal and exponential distributions if $\bar{c} < 2 \cdot 10^{-4}$, $\bar{c} < 5 \cdot 10^{-3}$ and $\bar{c} < 2 \cdot 10^{-2}$, respectively. The results of Figs. 7.12-7.14 show, however, that even if $\bar{c}$ is higher than these thresholds, the accuracy of the ordered statistics solution is not as good as desired. In fact, the accuracy is worse for the log-normal and exponential distributions than for the Gamma distribution. This inaccuracy is due to the fact that the ordered statistics approximation has large variance for the strongest channels. This variance is more significant in the log-normal distribution, because of the heavy tail of its p.d.f.. The ordered statistics waterlevel assumes values of channel gains which are far from the true values and the efficiency decreases.

So far we have assumed simple channel models where all sub-channels are independent and identically distributed. In an OFDM system, for instance, the equivalent channel gains (of each sub-carrier) have some degree of dependence, as a function of the delay spread, and may be differently distributed. Figure 7.15 shows the empirical c.d.f. of the gap to the optimal solution for four standard channel models:

- ITU Veh-A: ITU-R Vehicular A channel model of 6 taps with maximum multipath spread of 2.51 μs,
- TU-6: 3GPP COST 207 Typical Urban models of 6 taps with a maximum multipath spread of 5 μs,
- TU-12: 3GPP COST 207 Typical Urban models of 12 taps with a maximum multi-path spread of 5 μs, and
- BU-12: 3GPP COST 207 Bad Urban 12-taps model with a delay spread of 10 μs.

The figure considers an OFDM system with $n = 512$ sub-carriers. The channel multipath coefficients follow a Rayleigh distribution, the bandwidth is 10 MHz, the noise figure is 7 dB and the path loss is 80 dB. The number of sub-carriers experiencing independent fading depends on the coherence bandwidth or multipath spread. Despite assuming independent channel gains, Fig. 7.15 demonstrates that the efficiency is more than 95% the optimal efficiency for most of the cases. For the ITU Veh-A channel model, which has the shortest delay spread, less than 10% of the times the loss is between 5% and 2%. It is possible to reduce the gap to the optimal if n is reduced and channels of similar gains are grouped.

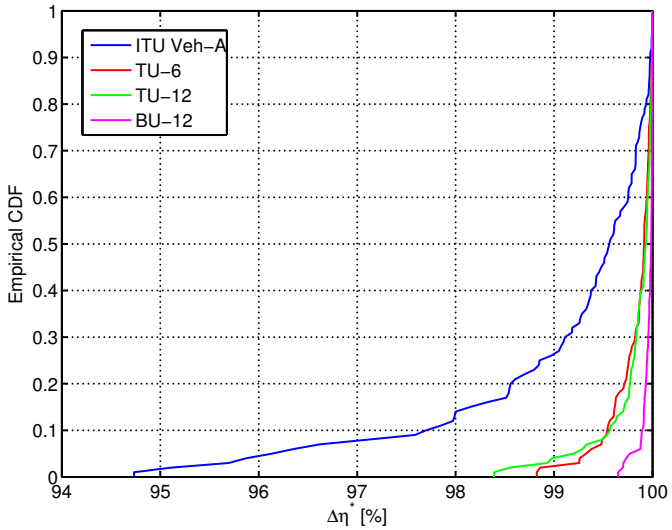

Figure 7.15: Empirical c.d.f. of the difference between the ordered statistics efficiency and the optimal (iterative) water-filling solution for 100 standard channel realizations.

7.6.2 Computational Complexity

The gains in terms of complexity reduction are worth the losses in terms of efficiency when compared to the GABS [102] and Dinkelbach [104] algorithms (Fig. 7.16). These algorithms solve the energy efficiency problem, which has identical formulation to the resource efficiency problem except that $c_o^n = c_o^r = 0$. The Dinkelbach and ordered statistics algorithms compute the waterlevel constant first from equation (6.6.13). Both then use

$$\begin{aligned} x_i^* &= \left(q' - \frac{1}{\gamma_i}\right)^+, \\ r_i^* &= \log\left(q'\gamma_i\right)^+ \end{aligned} \tag{7.6.2}$$

for each $i = 1 \dots n$ to obtain the power and rate for each channel. The GABS algorithm is a gradient descend algorithm. It directly computes the optimal rate vector. The Ordered algorithm obtains the solution $\sim 30x$ faster than the Dinkelbach algorithm and $\sim 400x$ faster than the GABS algorithm.

The GABS and Dinkelbach algorithms obtain the optimal solution, while the ordered statistics algorithm is sub-optimal. However, we have showed in the preceding sections that the loss of efficiency of our sub-optimal solution is negligible. On the other hand, when the number of channels is very large, as in the case of the SDR Cloud, the complexity of optimal methods makes the application of water-filling solution impractical. The complexity of water-filling algorithms explains that these kind of solutions are rarely employed. Our solution, on the other hand, allows these solutions to be implemented in large-scale systems at reasonable cost.

7.7 Practical Considerations

The water-filling solution, despite its good performance has some important drawbacks that difficult its implementation in practical systems. We discuss them in this section.

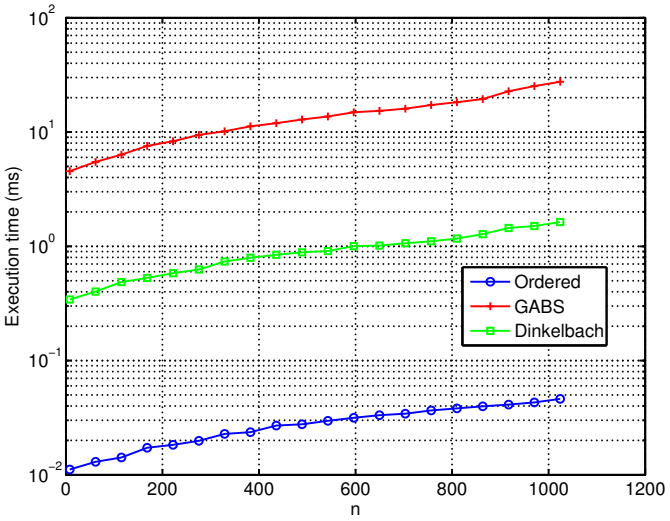

Figure 7.16: Execution time of the ordered statistics based water-filling proposal compared to the GABS and Dinkelbach methods, as a function of the number of channels. The channel gains are independent and exponentially distributed.

7.7.1 Look-up Table Size

We have shown in the preceding sections that the function $\omega^{-1}(\bar{\gamma}\bar{c})$ can be used for obtaining the optimal waterlevel from the channel statistics, rather than the channel realizations. For most channel distributions, however, $\omega^{-1}(\bar{\gamma}\bar{c})$ is complex to compute. On the other hand, $\omega^{-1}(\bar{\gamma}\bar{c})$ can be computed offline for different channel distributions and stored in a LUT.

The size of the LUT determines the precision of z^*. Figure 7.17 plots the gap to the optimal resource efficiency due to the LUT finite size. The loss is less than 2% for a table size of $M = 500$ samples. For smaller table sizes, the loss is appreciable when $\bar{\gamma}\bar{c}$ is large, because z^* is very close to 1. As it has been shown in the preceding section, a better solution can be obtained in this range of values assuming that all channels are active (i.e. $k = n$ extreme case solution).

As $\bar{\gamma}$ or $\bar{c}$ change with time, the optimal z^* is obtained form the table in O(1) time. The power allocation is then given by

$$x_i^* = \left(\frac{1}{\gamma\left(z^*\right)} - \frac{1}{\gamma_i} \right)^+, \tag{7.7.1}$$

for each channel $i = 1 \ldots n$. Computing the transmission power requires at most n arithmetic operations. Once the transmitter knows the power x_i it only remains to compute the maximum achievable rate from the Shannon capacity formula.

7.7.2 Fractional Rates

The water-filling solution produces rates r_i for each channel that have fractional parts or that are very small. Such small or fractional r_i can complicate the encoder and decoder implementation.

In Fig. 7.18 we evaluate the effect of finite encoding granularity [143]. This scheme rounds the optimal rates to the closest supported rate. If r_i^* is the ith channel optimal rate, an encoder supporting a granularity of β rounds the rate to the closest value satisfying $m\beta$, with m an integer. Figure 7.18 shows that for granularities lower than 1/4, the loss of efficiency is less than 5%. On the other hand, if we impose discrete rates ($\beta = 1$ in the

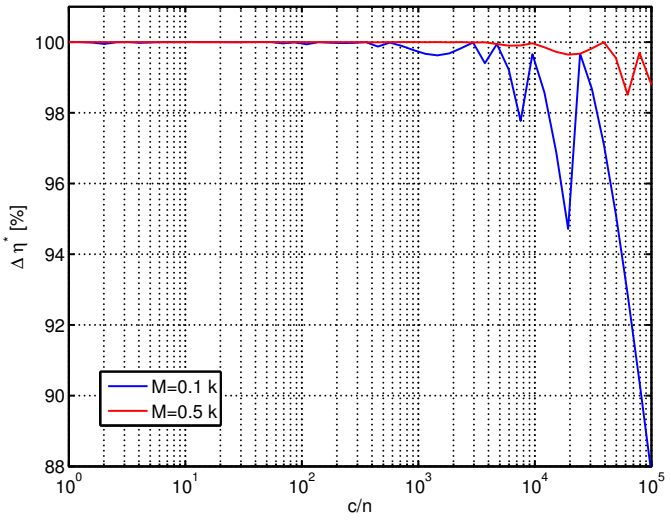

Figure 7.17: Gap to the optimal resource efficiency for the exponential distribution due to finite size M of the $\omega^{-1}(x)$ LUT.

figure), the resource efficiency decreases by $20-30\%$ with respect to the optimal. It is also observed that the loss with respect to the optimal increases with n. Granularities of the order of 0.1-0.3 are feasible to achieve with rate matching, as in the case of the 3GPP UMTS or LTE standard [144].

7.7.3 Constant Rate Power Allocation

Another option for simplifying the transmitter design is to allocate the same number of bits to each channel. We call this problem the constant rate (CR) problem, in contrast to the water-filling problem. If equal power and rate is allocated to all channels, the problem is denoted constant rate and power (CRP).

The CR problem can be also defined by means of ordered channel gains. Given the ordered sequence of channels, the problem consist on finding the k-th largest channel $\gamma_{k:n}$ such that all channels $\gamma_{i:n}$, $i \leq k$ are assigned the same rate. The rest of the channels are switched off. The CRP problem is similarly addressed, with the only difference that the same power is allocated to all active channels as well. The power is chosen so that the capacity of the poorest active channel equals the transmission rate.

Both problems (CR and CRP) can be formulated in terms of the transmission rate, which we denote r, and the number of active channels k. The formulation is equivalent for both problems except that the total transmitted power P_t, is

$$P_t(r) = \begin{cases} \sum_{i=1}^{k} \frac{e^r-1}{\gamma_i} = (e^r-1)\, n \frac{h(z)}{\bar{\gamma}} & \text{for CR} \\ k\frac{e^r-1}{\gamma_k} = (e^r-1)\, \frac{zn}{\bar{\gamma}\gamma(z)} & \text{for CRP.} \end{cases} \tag{7.7.2}$$

The resource efficiency is then

$$\eta = \frac{kr}{c_o^0 + c_o^n n + + c_o^r kr + c_p P_t(r)}. \tag{7.7.3}$$

Define

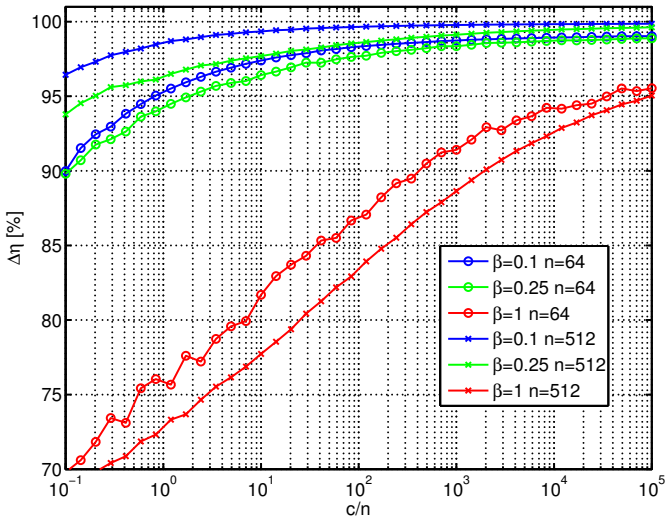

Figure 7.18: Gap to the optimal resource efficiency due finite granularity rate with granularity β and n channels for the exponential distribution.

$$\Psi(z) = \begin{cases} \frac{1}{h(z)} & \text{for CR} \\ \frac{\gamma(z)}{z} & \text{for CRP.} \end{cases} \tag{7.7.4}$$

so that

$$\begin{aligned} \min_{r,z} \quad & -\frac{zr}{\bar{\gamma}\bar{c} + \frac{e^r-1}{\Psi(z)}} \\ \text{s.t.} \quad & znr \geq R_{\min} \\ & \frac{n}{\bar{\gamma}} \frac{e^r-1}{\Psi(z)} \leq P_{\max} \end{aligned} \tag{7.7.5}$$

solves the CR and CRP power allocation problems (considering $c_o^r = 0$ and $\bar{c} = \frac{c_o^0 + c_o^n n}{n c_p}$).

The functions e^r and $1/\Psi(z)$ are monotonically non-decreasing and convex functions of r and z. Hence, the problem is a convex fractional program and the objective function is quasi-convex with respect to z and r separately. The Hessian matrix of the numerator is:

$$H(zr) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{7.7.6}$$

is positive semi-definite. The Hessian matrix of the denominator is:

$$H\left(\frac{e^r-1}{\Psi(z)}\right) = \begin{pmatrix} -\frac{e^r}{\Psi(z)^3}\left[\Psi(z)\Psi''(z) - 2\Psi'(z)^2\right] & -\frac{e^r}{\Psi(z)^2}\Psi'(z) \\ -\frac{e^r}{\Psi(z)^2}\Psi'(z) & \frac{e^r}{\Psi(z)} \end{pmatrix} \tag{7.7.7}$$

is also positive semi-definite, because $\Psi(z)$ is a positive, decreasing and convex function. Therefore, the term $\frac{e^r-1}{\Psi(z)}$ is convex and the problem is jointly quasi-convex w.r.t z and r. Consequently, efficient convex solvers can be applied because any local minima is a global minima.

Contrary to the water-filling optimization problem, now the constraints depend on the two variables and the problem can not be solved analytically. The solution (r^*, z^*) can be computed offline and stored in two LUTs for different channel statistics and the factor $\bar{\gamma}\bar{c}$, similarly to the water-filling case.

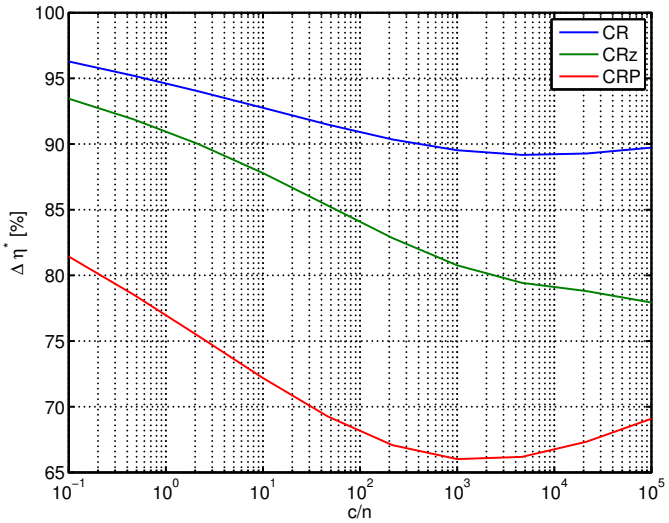

Figure 7.19: Loss of resource efficiency due to constant rate (CR), constant rate with water-filling z^* (CRz) and constant rate and power (CRP) allocations for the exponential distribution.

The implementation can be simplified at the expense of reducing the efficiency by using $z^* = \omega^{-1}(\bar{\gamma}\bar{c})$ the solution of the water-filling power allocation. Then, the problem reduces to find the rate r^* that maximizes the resource efficiency, considering the equal rate or equal power allocation, which is:

$$r^* = 1 + W\left[(\bar{\gamma}\Psi(z^*)\bar{c} - 1)\, e^{-1}\right]. \tag{7.7.8}$$

Figure 7.19 plots the loss of resource efficiency due to the CR and CRP allocation. The CR with z^* the solution of the water-filling problem is denoted CRz in the figure. The CRPz case is omitted for clarity, but as in the CR case it entails an extra loss of efficiency. Figure 7.20, on the other hand, plots the resulting z^* of the water-filling, CR and CRP allocations. It is observed that the CR and CRP allocations concentrate the power in less channels than the water-filling solution. The loss of efficiency is lower when a few channels are used, because the sum of the difference between the CR, CRz or CRP allocations and the optimal allocation sums among less channels.

7.7.4 Transmitting With Partial CSI

The most important drawback of water-filling power allocations is that the channel gains measured at the receiver have to be send back to the transmitter. Ordered statistics allow reducing the feedback to some extend. Transmitting the channel order and $\bar{\gamma}$ suffices for estimating the values of each channel. The counterpart is that, since the ith order channel is a random variable, the mutual information becomes a random variable too, and then there is no guarantee that a certain rate can be achieved, resulting in channel outages. Nevertheless, channel outages can be made arbitrary small because the variance of the ith channel gain is given by the normal approximation (equation (7.1.3)). On the other hand, given a rate and modulation scheme, the transmitted power can be adjusted in order to satisfy the desired BER.

To implement this scheme, the order of the channel gains needs to be computed by the receiver and sent back to the transmitter. This requires $\log_2 n$ bits to index each channel and $k \log_2 n$ to send the k strongest channel indexes (those that are active). The channel model

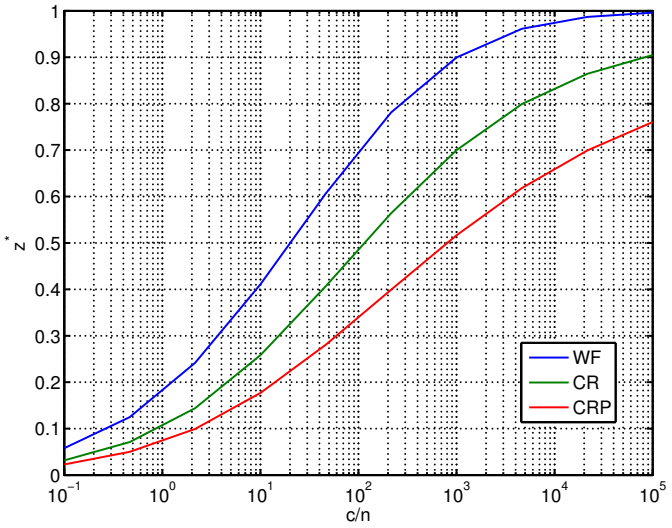

Figure 7.20: Optimal z^* of the water-filling (WF), constant rate (CR) and constant rate and power (CRP) allocations for the exponential distribution.

parameters (e.g. average) need to be transmitted too. In a typical complete CSI scheme, nM bits are required if each channel gain is quantified with M bits. Therefore, ordered statistics do not offer considerable feedback reduction in practice, unless $k \ll n$. The reduction can be significant in scenarios where only a few channels are active and n is very large.

For further reduction of the feedback, some performance needs to be sacrificed. One option is assuming certain non-zero BER. Most of the related work on CSI reduction constraints the average bit error rate to some target. The Ordered Subcarrier Selection Algorithm (OSSA) is a simple scheme for reducing feedback in OFDM [142]. The receiver sends the transmitter a mask of n bits, with the ith bit indicating whether channel i is active or not. The transmitter allocates the same rate and power to all active channels.

The OSSA strategy is equivalent to the CRP strategy defined in the previous subsection. The receiver solves the CRP problem and feedbacks the transmitter a mask indicating the active channels plus the power and rate to transmit. The work in [142], however, only considers the maximum throughput objective. Figure 7.19 shows the resource efficiency loss due to CRP allocation (i.e. partial CSI). Note that, while OSSA is close to the optimal in terms of throughput, the CRP scheme has an efficiency around $65 - 70\%$ of the optimal (see Fig. 7.19). Since perfect CSI is unknown, the system can not operate at capacity and need to assume the worst channel. Power is consequently wasted on the best channels, increasing the power costs and reducing the efficiency.

7.7.5 Solution Algorithm

We have shown throughout this section that ordered statistics allows computing the power allocation in constant time independently of the channel realization. The same procedure is valid for three different objectives

- **EM** or efficiency maximization, maximizes the resource efficiency η,
- **RM** or rate maximization, maximizes the sum-rate R_t,
- **MM** or margin maximization, minimizes the transmission power P_t.

For each objective, the procedure also admits three different transmission modes:

- **WF** or water-filling. Is the optimal solution, adapts power and rate according to each channel gain,
- **CR** or constant rate. Assigns equal rate to all active channels and adapts power only, and
- **CRP** or constant rate and power. Assigns equal rate and power to all active channels.

The procedure is divided in two steps, an offline and an online part [145]:

1. *Offline*: For each channel model, compute $\omega^{-1}(x)$, $\pi^{-1}(x)$ or ρ^{-1} for objectives EM, RM and MM of the WF transmission mode or compute z^* from (7.7.5) for the CR and CRP modes (only the EM objective is described, but the RM and MM follow equivalently).
2. *Online*:
   (a) Obtain γ_i for $i = 1 \dots n$, compute $\bar{\gamma} = \mathbb{E}\{\gamma_i\}$ and estimate the channel model.
   (b) Check if $\bar{\gamma}\bar{c}$ satisfies one of the extreme cases $k = 1$ or $k = n$, otherwise obtain z^* from the tables, according to the problem objective and obtain the power allocation vector x_i^*, $i = 1 \dots n$, while applying the peak power constraint $P_{\max}$.
   (c) If the power constraints are violated, set the problem objective to RM and repeat the step (b), otherwise continue
   (d) If the rate constraints are violated, set the problem objective to MM and repeat the step (b), otherwise finish.
   (e) If both the rate and power constraints are violated, the problem is infeasible.

7.8 Example: Exponential Distribution

In Chapter 4 we found that the exponential distribution appears in many different scenarios, e.g. OFDM, time-varying channel with NLOS and is also a good approximation of the eigenvalues of a MD-MIMO channel matrix. In this section, we derive the z-functions for the exponential distribution.

Consider a multi-channel system with $n \gg 1$ i.i.d. exponentially-distributed channel gains with average 1. The $\gamma(z)$ function is

$$\gamma_{\exp}(z) = -\log z. \tag{7.8.1}$$

Figure 7.21 plots the ordered sequence of 100 realizations of 50 exponential random variables as well as the expression (7.8.1). Note that the variance is very small for $i > 30$. The deterministic assumption, however, completely fails for the largest channel gain. The limiting distribution of the maximum order statistic is Gumbell distributed (see Appendix C). We however neglect the randomness of the largest channel and assume that it is deterministic and has a value of $\gamma_{\exp}(1/n) = \log n$.

The $g_{\exp}(z)$ function is

$$\begin{aligned} g_{\exp}(z) &= \int_0^z \log\left(-\log y\right) dy \\ &= z \log\left(-\log z\right) - \mathrm{li}\left(z\right), \end{aligned} \tag{7.8.2}$$

and the $h_{\exp}(z)$ function is

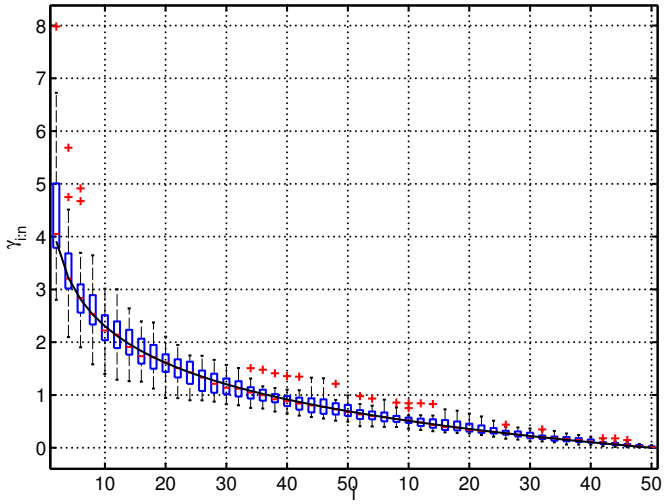

Figure 7.21: Ordered values of 100 realizations of 50 i.i.d. exponential random variables with $\bar{\gamma} = 1$. The central box represents the central 50% of the data. Its lower and upper boundary lines are at the 25% and 75% quantile of the data. The whiskers extend to the most extreme data points not considered outliers and red crosses represent outliers. The solid line is the theoretical ordered statistic given by (7.8.1).

$$\begin{aligned} h_{\exp}(z) &= -\int_0^z \frac{1}{\log y} dy \\ &= -\mathrm{li}\,(z)\,. \end{aligned} \tag{7.8.3}$$

$\mathrm{li}(z)$ is the logarithmic integral and has a singularity at $z = 1$. However, the case $z = 1$ is the case $k = n$, which does not satisfy the assumption 7.4.

The average power per channel or $\pi(z)$ function, as defined in (7.4.6) becomes

$$\begin{aligned} \pi_{\exp}(z) &= \frac{z}{\gamma_{\exp}(z)} - h_{\exp}(z) \\ &= -\frac{z}{\log(z)} + \mathrm{li}\,(z) \\ &= \mathrm{li}\,(z) - \frac{z}{\log z}. \end{aligned} \tag{7.8.4}$$

The average rate per channel $\rho(z)$, as defined in (7.4.8) becomes

$$\begin{aligned} \rho_{\exp}(z) &= g(z) - z \log \gamma_{\exp}(z) \\ &= z \log \left(\log \frac{1}{z} \right) - \mathrm{li}\,(z) - z \log \left(-\log(z) \right) \\ &= -\mathrm{li}(z). \end{aligned} \tag{7.8.5}$$

The $\omega(z)$ function for the exponential case is:

$$\begin{aligned} \omega_{\exp}(z) &= \frac{1}{\gamma_{\exp}(z)} \rho_{\exp}(z) - \pi_{\exp}(z) \\ &= \frac{\mathrm{li}(z)}{\log(z)} - \left(\mathrm{li}\,(z) - \frac{z}{\log z} \right) \\ &= \mathrm{li}(z) \left(\frac{1}{\log z} - 1 \right) + \frac{z}{\log z}. \end{aligned} \tag{7.8.6}$$

The optimal resource efficiency η^* is

$$\eta^*_{\text{exp}}\left(\bar{\gamma}\bar{c}\right) = -\frac{\bar{\gamma}}{c_p}\log\left(\omega^{-1}_{\text{exp}}\left(\bar{\gamma}\bar{c}\right)\right), \tag{7.8.7}$$

where we have assumed $c^r_o = 0$ for mathematical clarity.

7.8.1 Parametric Approximation

The exponential distribution has a single parameter, the mean. By Assumption 7.1, the mean of $F(x)$, the distribution used to compute the z-functions, is 1. Thus, it is possible to parametrize the z-functions, including $\omega^{-1}(x)$ and reduce the expressions complexity.

The function $z = \omega^{-1}(x)$ has a *sigmoid* or logistic shape when the argument is in the logarithmic domain. A sigmoid function is a function having an "S" shape. It can be parametrized as follows:

$$s(x) = \frac{1}{1 + a \cdot e^{-bx}}. \tag{7.8.8}$$

Applying the logarithm to x, we obtain

$$z^*_{\log}\left(\bar{\gamma}\bar{c}\right) = \left(1 + a(\bar{\gamma}\bar{c})^{-b}\right)^{-1}. \tag{7.8.9}$$

This simple expression also leads to a simple expression of the optimal resource efficiency. Inserting (7.8.9) in (7.8.7) yields:

$$\eta^*_{\log} = \frac{\gamma}{c_p}\log\left(1 + \frac{a}{(\bar{\gamma}\bar{c})^b}\right). \tag{7.8.10}$$

We have performed non-linear least squares model fitting and obtained the optimal parameters:

$$\begin{aligned} \hat{a} &= 0.4136 \\ \hat{b} &= 0.5366. \end{aligned} \tag{7.8.11}$$

Figure 7.22 plots the $\omega^{-1}(x)$ function and the logistic approximation. The performance of the logistic approximation is good enough for computing the water-filling solution without need of using or storing a LUT. The approximation is accurate except for $\bar{\gamma}\bar{c} \to \infty$, where $z^* \to 1$. In this case, however, the solution $k = n$ can also be computed in closed form. Thus, the combination of the $k = 1$ and $k = n$ cases with the logistic approximation eliminates the need of using iterative algorithms or LUTs, provided that the channel gains are exponentially distributed. This results are shown in Fig. 7.23, where we have repeated those shown on Fig. 7.12. Note that when the logistic curve starts to decay, approximately at $c/n > 10^4$, it crosses with the $k = n$ solution. For small c/n values, the logistic approximation is very close to the efficiency obtained by the original ω^{-1} function.

Besides being able to compute the water-filling in closed form, these kind of expressions allow to analyse or optimize the resource efficiency as a function of the system parameters. The same exercise can be done with other distributions, fixing the other parameters.

7.9 Power Allocation in the SDR Cloud

In Chapter 4 we discussed the statistical properties of the equivalent channel gain a user observes in the cell-less network architecture enabled by the SDR cloud, when an SDMA

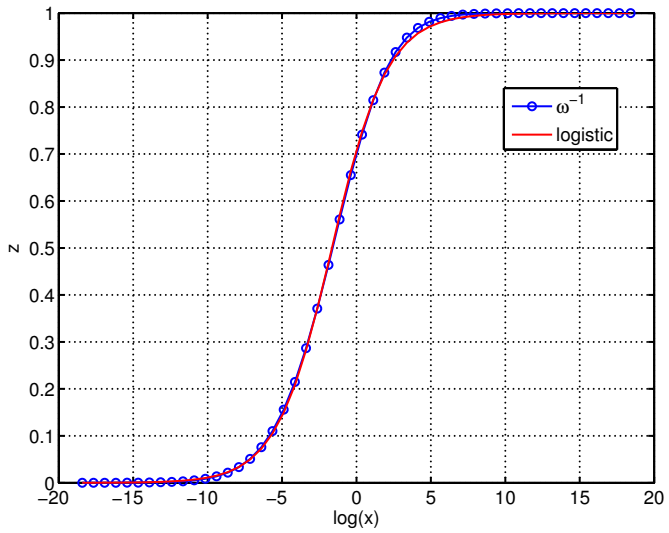

Figure 7.22: Plot of the $z = \omega^{-1}(\log x)$ and the logistic approximation

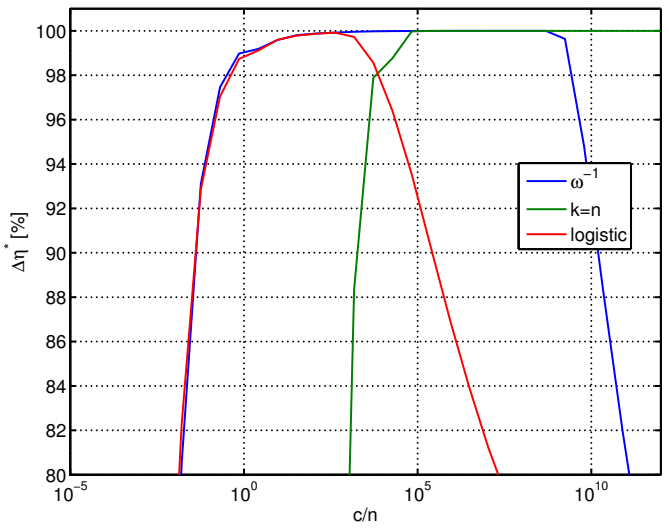

Figure 7.23: Gap to the optimal resource efficiency of the logistic approximation (see Fig. 7.12 for simulation parameters).

technique is employed to multiplex users. The best resource efficiency is achieved when the transmission rate approaches capacity. The maximum mutual information with SDMA is determined by the eigenvalues of $\mathbf{H}\mathbf{H}^H$ with $\mathbf{H}$ the equivalent MIMO matrix. The eigenvalues of $\mathbf{H}\mathbf{H}^H$ are approximately exponentially distributed when the entries of $\mathbf{H}$ are a combination of path loss, shadowing and multipath Rayleigh fading. If the BSs employs ZF precoding (in the downlink) or equalization (in the uplink), the equivalent channel gains are also approximately exponential. Let B and K denote the number of BS and users, respectively, the average channel gains are

$$\bar{\gamma}^{\mathrm{SVD}} = B\frac{\nu(d)}{N_0} = B\Gamma(d), \tag{7.9.1}$$

$$\bar{\gamma}^{\mathrm{ZF}} = (B-K+1)\,\frac{\nu(d)}{N_0} = (B-K+1)\,\Gamma(d), \tag{7.9.2}$$

for the SVD and ZF cases, respectively. N_0 is the noise power and ν is the variance of the entries of $\mathbf{H}$, which is a function of the *supercell* diameter d. The function $\Gamma(d)$ is the average CNR a user sees with all BS.

A supercell is the set of BS whose signals are processed jointly. If the distance between a user and a BS is uniformly distributed in the interval (d_0, d), ν becomes a function of d and is approximated as

$$\nu(d) = \Upsilon\frac{1}{d}, \tag{7.9.3}$$

where

$$\Upsilon = c_0 \cdot 10^{\log(10)\cdot(\sigma/10)^2/2}\,\frac{d_0^{1-\alpha}}{\alpha-1}. \tag{7.9.4}$$

α is the signal path loss exponent, c_0 is the loss at the reference distance of 1 m and σ is the log-normal shadowing variance.

SDMA transmission allows choosing the power allocation that optimizes the resource efficiency in the SDR cloud. The number of independent channels is $n = K$. $B > K$ channels have to be reserved from the SDR cloud, though. Since there are B signal flows being processed at B antennas, the cost per channel should be proportional to B instead of n. We can redefine $c_o^n = \frac{B}{n}c_o^b$ where c_o^b is the cost per antenna to obtain the desired model.

It is clear that for large number of users, algorithmic solutions are inefficient if the channel varies frequently. Furthermore, if SDMA is combined with OFDM, we have a different and independent channel matrix $\mathbf{H}_j$ for each sub-carrier $j = 1 \ldots N_c$. The number of independent channels is then $n = N_c\dot{K}$. In this scenario, the ordered statistics solution offers significant advantages.

Let us consider the case $N_c = 1$ and $n = K$. It was shown in section 7.4.2 that the optimal resource efficiency increases with n, because more channels increase the likelihood of having *very good* ones. The average channel gain $\bar{\gamma}$, when the channel gains are the eigenvalues of $\mathbf{H}\mathbf{H}^H$, is independent of K (equation (7.9.1)). Therefore, the theoretical maximum resource efficiency, the one that employs a capacity-achieving SDMA technique, increases with the number of users K. On the other hand, the average channel gain for ZF beamforming decreases with K, as shown by equation (7.9.2). Figure 7.24 shows the resource efficiency computed iteratively (i.e. the optimal solution). When the number of users is less than $B/2$, the ZF scheme achieves an efficiency almost equal to the maximum efficiency (SVD in the figure). For large K, however, $\bar{\gamma}$ is too small and the efficiency decreases. We can observe that the ZF beamforming scheme achieves a maximum resource efficiency for certain ratio K/B.

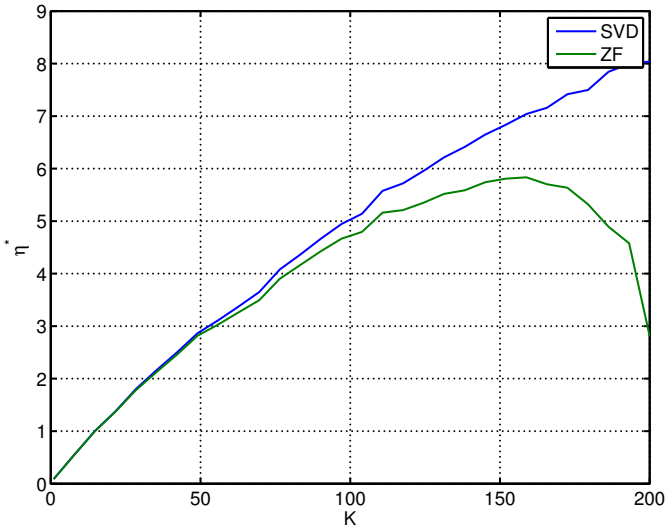

Figure 7.24: Optimal resource efficiency η^* of a capacity-achieving SDMA (SVD in the figure) compared to the ZF beamforming efficiency (ZF in the figure). The cost constants are $c_o^0 = 1$, $c_o^n = 0$, $c_o^r = 0$ and $B = 200$.

One of the main advantages of the MD-MIMO architecture is that the transmission power can be made arbitrary small increasing the number of BS per user. Intuitively, it is very likely that a randomly located user finds an antenna geometrically close. Moreover, different signals from different antennas are combined, producing a power gain. Mathematically, $\Gamma(d)$ is decreasing with d and γ^{SVD} or γ^{ZF} increasing with B.

In terms of the resource efficient solution, the consequence of this high average CNR scenario is that, unless the power costs are much larger than the rest of the costs ($\bar{c} \ll 1$), all channels will be used and the extreme case $k = n$ (see Section 7.4.3) is optimal. We know that if

$$\bar{c} \geq \frac{\omega\left(1 - \frac{1}{n}\right)}{\bar{\gamma}}, \tag{7.9.5}$$

the optimal solution is transmitting on all channels. For instance, assuming capacity-achieving SDMA, if $n = K = 100$, $B = 200$ and the supercell radius is 2 km, (7.9.5) becomes

$$c = \frac{c_o^0 + 200 c_o^b}{c_p} \geq 200 \frac{306.63}{3.6 \cdot 10^5} = 0.167. \tag{7.9.6}$$

Thus, unless the power costs are 10 times greater than the operational costs, all users will be selected. This is a reasonable assumption in an SDR cloud. Operational costs can be made arbitrary small by applying economy of scale principles, or improving the infrastructure computational power, for instance. The power costs, on the other hand, are directly proportional to the energy market cost.

Figures 7.25 and 7.26 plot the gap to the optimal resource efficiency for the SVD and ZF multiplexing schemes, assuming that $c = 10^{-3}$. Users' distance with the antennas is uniformly distributed in the interval $(20, 2000)$ meters. The log-normal shadowing has variance 6 dB, the path loss exponent is 3, the noise bandwidth is 10 MHz and the noise figure is 7 dB. The carrier frequency is 2 GHz. The results show that a combination of the logistic approximation and the $k = n$ solution is accurate in the entire range of operation (i.e. $K \in (1, B)$). In the ZF case (Fig. 7.26), the assumption that the channels are exponentially distributed does not

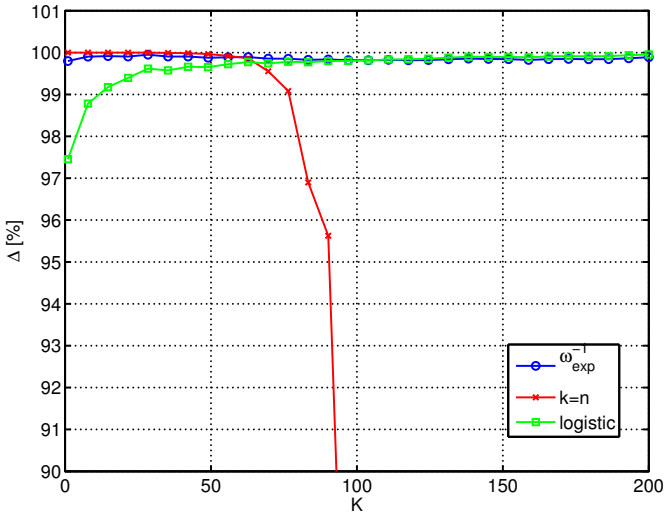

Figure 7.25: Gap to the optimal efficiency of a power allocation given by $z^* = \omega^{-1}(\bar{\gamma}\bar{c})$, by the logistic approximation and by the assumption that all channels are active ($k = n$), when the channel gains are the eigenvalues of $\mathbf{H}\mathbf{H}^H$. $c = 10^{-3}$ and $B = 200$.

hold in the case $K = B$, as it is shown in Fig. 4.6. Thus, the ordered statistics or the logistic solutions entail a loss of performance of 10% with respect to the iterative solution.

7.9.1 SDR Cloud Parameter Optimization

The simplicity of the logistic approximation allows studying the resource efficiency as a function of the system parameters. Using equation (7.8.10), the optimal efficiency becomes

$$\begin{aligned} \eta^*_{\text{SVD}} &= \frac{B\Gamma(d)}{c_p} \log\left(1 + a\left(B\Gamma(d)\frac{c_o^0 + c_o^b B}{K c_p}\right)^{-b}\right) \\ \eta^*_{\text{ZF}} &= \frac{(B-K+1)\Gamma(d)}{c_p} \log\left(1 + a\left((B-K+1)\Gamma(d)\frac{c_o^0 + c_o^b B}{K c_p}\right)^{-b}\right). \end{aligned} \tag{7.9.7}$$

More BS per user increase the average CNR with both techniques, reducing the power required to achieve a given rate. More BSs also imply more cost though, because the network operator pays the infrastructure operator for using the BS resources. Under some circumstances, a maximum exists for $B \geq K$, so that one can find the optimal number of BS per user that maximises the resource efficiency. Figure 7.27 plots the resource efficiency as a function of the ratio B/K. If the cost per BS is $c_o^b = 1/200$, the maximum efficiency is achieved when there are approximately 3 and 5 times more BS than users for the capacity-achieving (SVD) and ZF schemes, respectively.

These results are interpreted as follows. Consider a supercell with $B = 500$ BS and ZF precoding. The optimal working point is having $K = 100$ users. If there are less users requesting service, some BS should be turned off and returned to the cloud because they add a cost that is not paid off. On the other hand, if more users request service, more BS should be dynamically acquired, while maintaining the ratio B/K constant. This ratio, is independent of the realization, it is also a function of the statistical properties of the channel matrix $\mathbf{H}$.

Note that the elasticity of the SDR cloud architecture enables the network to operate at the maximum levels of efficiency. On the other hand, if for some reason, the network operator

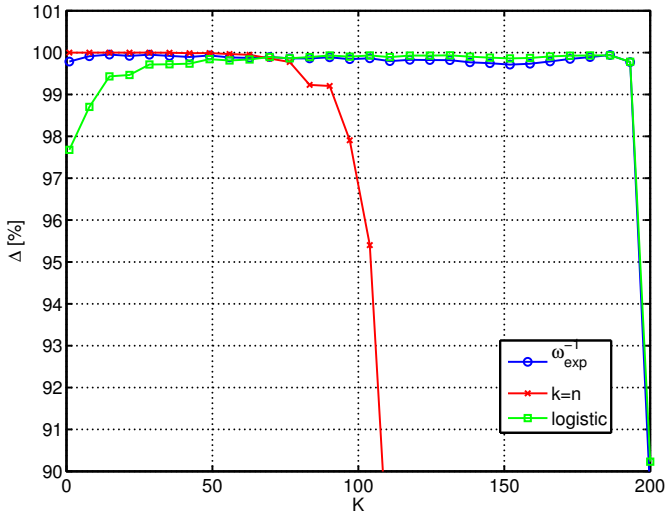

Figure 7.26: Gap to the optimal efficiency of a power allocation given by $z^* = \omega^{-1}(\bar{\gamma}\bar{c})$, by the logistic approximation and by the assumption that all channels are active ($k = n$), when the channel gains follow from ZF precoding. $c = 10^{-3}$ and $B = 200$.

can not acquire more BS, new users should be denied, because there is no power allocation possible that achieves a higher efficiency. In other words, serving a new user provides no benefit to the operator, only increases the costs.

The optimal ratio B/K can not be derived analytically, given the functions (7.9.7). The derivative is easy to obtain, though, and can be used to plot the optimal B against the desired parameter, using a numerical solver. The main advantage of the ordered statistics solution, more precisely the logistic approximation, is that the finding the root of the derivative is simpler than optimizing the efficiency by simulations.

Let $c_o^b B \gg c_o^0$ and $c = \frac{c_o^b}{c_p}$, the second derivative of η^*_{SVD} with respect to B is negative if

$$B \leq \theta \sqrt{\frac{K}{c\Gamma(d)}}, \tag{7.9.8}$$

where

$$\theta = \left(\frac{2b-1}{2a}\right)^{-1/2b} \approx 5. \tag{7.9.9}$$

Then $\eta^*_{\text{SVD}}(B)$ is convex. If (7.9.8) holds for any $B \geq K$, we know the optimal number of BS is greater than the number of users. If (7.9.8) does not hold, $\eta^*(B)$ is concave and the maximum is at $B = K$. Thus, if

$$\frac{c_p}{c_o^b} \leq \frac{1}{\theta^2} \frac{\Gamma(d)}{K}, \tag{7.9.10}$$

the optimal number of BS is $B = K$. The cost of operating a BS is very high or the power cost is low, there is no benefit in using more spatial degrees of freedom. It is more efficient to increase the transmission power. The number of BS are reduced to the minimum required to give service to the K users. On the other hand, if the power costs are high and the BS costs are low, the efficient strategy is the converse: turn down the transmission power and increase the number of active BS, exploiting the power gain that the extra degrees of freedom offer.

We plot in Fig. 7.28 the optimal ratio B/K as a function of the cost per BS c_o^b. When c_o^b is very small, the optimal strategy is $B \gg K$. The SVD and ZF schemes need the same

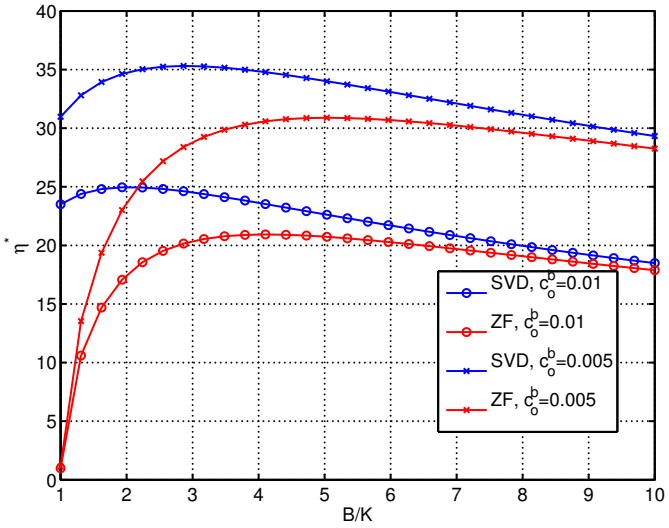

Figure 7.27: Resource efficiency as a function of the ratio B/K for different costs c_o^b. The scenario is the same as in Figs. 7.25-7.26.

number of BS per user. The resource efficiency of ZF precoding is almost optimal, as shown in Figs. 7.24 and 7.29. As c_o^b approaches 1, the ZF scheme requires approx $25-50\%$ more BSs than the SVD to achieve the optimum efficiency. The optimum efficiency of the ZF scheme is 35% lower than the SVD, however (Fig. 7.29). The SVD scheme achieves a power gain independently of K. Hence, as the cost per BS becomes non-negligible, the optimal solution is $B = K$. The ZF scheme never operates in this region, because the power gain becomes zero. Thus, the efficient solution uses more BS, which increases the gap to the efficiency shown by the capacity-achieving multiplexing scheme (Fig. 7.29).

7.10 Summary

A novel technique based on ordered statistics was proposed for solving water-filling power allocation problems. The solution is obtained in constant time, that is, independently of the number of channels n. This ensures a highly scalable solution for large systems. The gap to the optimal solution is negligible under typical conditions, while the computational complexity is reduced by 2-4 orders of magnitude when compared to state-of-the-art techniques for energy-efficiency maximization. The same methodology has been applied to three different objectives: efficiency, rate or margin maximization.

The efficiency maximizer waterlevel is a function of $\bar{\gamma}c/n$, where $c = \frac{c_o^0 + c_o^n n}{c_p}$ is the ratio of operational to power costs. The rate maximizing waterlevel, on the other hand, is determined by $\bar{\gamma}P_{\max}/n$. The cost-channel-gain per channel, thus, play a role similar than the SNR per channel in RM problems. While in RM, the aim is finding the optimal *distribution* of the available power, in the EMA problem, the aim is finding the optimal *amortization* of the consumed costs c.

The optimal solution is characterized by a single function, $\omega(z)$, which depends on the channel statistics and is independent of n, the average gain $\bar{\gamma}$ and the costs $\bar{c}$. When efficiency is maximized, the number of active channels is increasing with the factor $\bar{\gamma}\bar{c}$. The presented solution allows evaluating the behaviour of the optimal solution with respect to the problem parameters. We have demonstrated that this behaviour is consistent with the single channel model ($n = 1$). For instance, the optimal rate is an increasing function of the ratio $\bar{\gamma}\bar{c}$, which

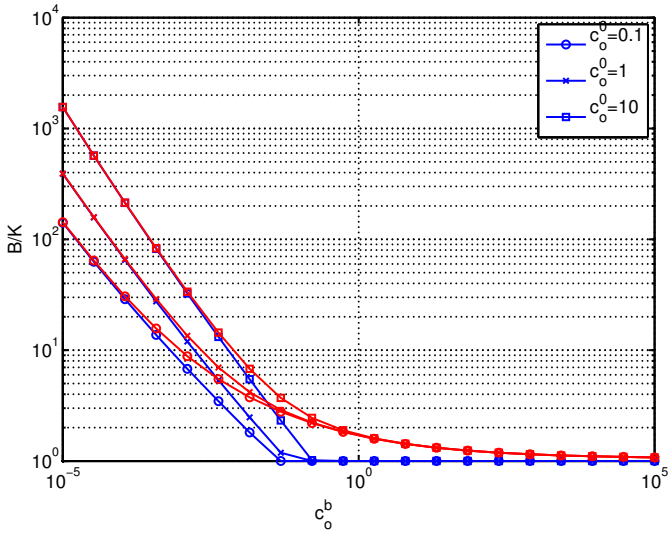

Figure 7.28: Optimal ratio B/K as a function of the cost per BS, c_o^b, when $c_p = 10^5$ and c_o^0 takes different values. The blue lines correspond to the capacity-achieving SVD multiplexing and the red lines to ZF precoding.

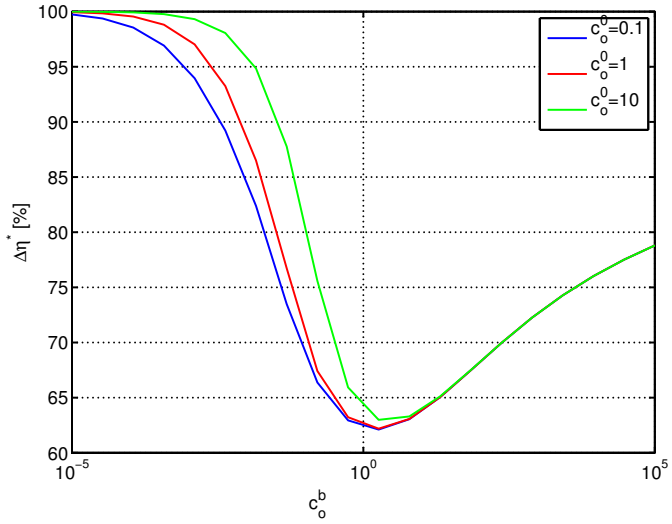

Figure 7.29: Comparison (ratio) of the ZF precoding multiplexing with respect to the capacity-achieving SVD multiplexing, when the system operates in the optimal ratio B/K as a function of c_o^b ($c_p = 10^5$).

implies that:

- if the operational costs are zero or negligible, the efficient strategy is a MM power allocation or
- if the power costs are zero or negligible, the efficient strategy is a RM power allocation.

The performance of the optimal efficiency is also consistent with the case $n = 1$, in particular:

- decreases with all cost constants c_o^r, c_o^n, c_o^0 and c_p,
- increases with the average CNR $\bar{\gamma}$ with limit $1/c_o^r$,
- if $c_o^r = 0$, increases with n without limit if $c_o^n = 0$ or converges to a constant if $c_o^n > 0$.

The last conclusion is of important interest because explains the behaviour of the best achievable efficiency as a function of the available degrees of freedom. For instance, it explains why the efficiency of SDMA is increasing with the number of users or multi-user diversity. The gains obtained by this diversity can be quantified using the tools introduced in this chapter.

The water-filling solution produces rates which are difficult to implement in practice. Simple rounding solutions, easy to achieve with a combination of a mother code and rate matching, have negligible effect to the resource efficiency. We have described and analysed other practical solutions. In particular, transmitting a constant rate in all channels or reducing the required CSI. These techniques are described in the patent application [145].

We simulated the power allocation problem in an SDR cloud, considering SDMA capacity and ZF precoding. We conclude that the assumption that the eigenvalues and ZF precoding channel gains are exponentially distributed is accurate enough and practical. The exponential distribution is characterized by its mean only. The z-functions can be expressed in closed form, as it was done in [146] for energy efficiency maximization. The exponential distribution also admits a parametric approximation of the $\omega^{-1}(x)$ function, obtaining the power allocation without any LUT or iterative method. The simple form of the approximation offers extra insights on the optimal solution and facilitates the optimization of system parameters.

As an example of the latter, we investigated the optimal number of BS per user, ratio B/K, as a function of the cost per BS. The results show that, under certain conditions, an optimal point $B > K$ exists that maximizes the resource efficiency. In order to be efficient, a network operator should keep this ratio constant, acquiring or leasing BS from the cloud as users arrive or leave. This dynamic scenario shows the benefits of the elasticity provided by the SDR cloud.

Our results also showed that ZF precoding is almost optimal if the cost per BS is small. As this cost increases, the loss of power gain of ZF precoding reduces the efficiency, when compared to the capacity-achieving SDMA scheme, because more BS per user have to be used, which implies more cost. Therefore, we conclude that, when designing a linear precoding technique, the power gain is very important if the cost per BS is non-negligible.

Chapter 8

Concluding Remarks

The SDR cloud is a promising solution for the "1000x data challenge" expected for the next decade. The platform offers the possibility to implement an MD-MIMO cell-less architecture, boosting the network capacity. At the same time, radio and computing virtualization allows a cost efficient deployment and operation of the infrastructure, creating new business models and opportunities. SDR clouds are a topic of current research. Some of the main challenges have been identified and solutions have been proposed. The main focus of this thesis have been the interactions between the virtualized radio and computing infrastructure.

Computing resource management for large distributed systems has a long research record. The problem of serving wireless users has never been tackled before, though. The probability that a user can be accepted by the cloud and the maximum session establishment time imposes a limit to the capacity a computing resource manager can deal with. We have identified this tradeoff and derived the capacity. Based on this analysis, a distributed computing resource management architecture was proposed for a large data center serving wireless users.

From the wireless point of view, the fact that the infrastructure is not owned, but consumed in a *pay-per-use* fashion, alters the objective of traditional radio resource management algorithms or techniques. Based on a simple pay-per-use cost model, we formulated the resource efficiency maximization power allocation (EMPA) problem. The solution to this problem indicates how power and rate should be allocated to users, in order to deliver the maximum amount of bits per consumed resource unit. The solution indicates that higher rates should be allocated to users with the best channels. The solution is algorithmically complex, though, and a novel technique based on ordered statistics has been proposed. This technique can be applied to any kind of water-filling power allocation problem. It consists on computing the water-level based on the channel statistics rather than the channel realization. The water-level is obtained in constant time, independently of the number of channels. This low complexity is of special interest in SDR clouds, where the number of channels is very large.

On the other hand, the presented solution allows analysing how the optimal solution behaves as a function of the problem parameters. For instance, it has been shown that the cost of using the infrastructure plays a role similar to the maximum power constraint or the minimum rate constraint in the rate or margin maximization problems. The rate maximization problem, is determined by the product of the power constraint and the average channel gain per channel, or equivalently the average signal-to-noise ratio per channel. The resource efficiency maximization problem is a function of the product of the infrastructure costs and the average gain. Therefore, while the rate maximization *distributes* power among channels, the efficiency maximization *amortizes* the cost transmitting more or less rate in certain channels.

The solution based on ordered statistics also indicates that the optimal resource efficiency

increases with the number of channels without bound if the cost of using one channel is negligible. If this cost is non-zero, the efficiency is also increasing but converges to a known constant and adding more channels does not improve the efficiency. In an SDR cloud system, this conclusion indicates that increasing the number of channels of the system, i.e. number of BS and users, increases the efficiency. The intuition suggests that more users increase the likelihood that some of them observe a good channel with a BS (multi-user diversity). On the other hand, more BS means more costs. Under some circumstances, an optimal number of BS per user exists, such that the efficiency is maximized. Thus, in a dynamic system where users enter and leave the system dynamically, the optimal strategy is acquire and lease antennas dynamically, while keeping the optimal ratio constant. This kind of dynamism is impossible to achieve with today's static infrastructure, requiring a scalable and elastic architecture as offered by the SDR cloud.

8.1 Future Work

We have identified some challenges related to the radio and computing resource management for SDR clouds. Furthermore, the tools developed throughout the thesis laid the foundations for further research on these topics. Among others, we identify the following open problems:

- The distributed computing resource management approach presented in Chapter 3 is optimal but very complex. Sub-optimal but less complex methods should be studied and evaluated.
- The predicted limitations of the computing resource allocator in an SDR cloud and the introduced solutions, should be evaluated in a small/medium-scale testbed.
- Linear precoding methods should be defined and evaluated for MD-MIMO systems. In this sense, the statistics of the channel gains resulting from these techniques should be characterized. The tools presented in this dissertation enable assessing the performance of a linear precoder with respect to capacity. In particular, it is of special interest in an SDR cloud to maintain the power gain as the number of users approach the number of BS.
- The optimal ratio of antennas per user in the SDR cloud was derived as an example. Future work should address other design parameters in the SDR cloud.
- We have considered interference-free independent channels. The resource efficiency in channels with interference should be investigated.
- User fairness or per-user constraints introduction in the power allocation formulation need to be considered.
- Finally, it should be investigated whether other class of problems, e.g. channel allocation, can be solved more efficiently using ordered statistics.

Appendices

Appendix A

The Lambert-W Function

The Lambert function $W(x)$ was originally defined in the L. Euler's paper [147]. It is defined as the multivalued inverse of the function $f(x) = xe^w$. That is, any function that is a solution to the transcendental equation

$$x = W(xe^x), \tag{A.0.1}$$

or

$$x = W(x)e^{W(x)} \tag{A.0.2}$$

is a Lambert $W(x)$ function. For some values of x, equation (A.0.2) has more than one root. Each different solution is called a branch of W. If x is real, the relation W is defined only for $x \geq -e^{-1}$. For $-e^{-1} \leq x \leq 0$ is double-valued and for $x \geq 0$ is single-valued. This is represented by two different branches (Fig. A.1). The main branch is denoted $W_0(x)$ or simply $W(x)$. It is defined for $x \in [-e^{-1}, +\infty)$ and takes values $W(x) \in [-1, +\infty)$. The lower-branch is denoted $W_{-1}(x)$ and is defined for $x \in [-e^{-1}, 0)$ and takes values $W_{-1} \in (-\infty, -1]$.

The function is implemented in many mathematical software: as *LambertW* in Maple and *lambertw* in MATLAB, for instance. It is also provided by the GNU Scientific Library GSL.

The first derivative of the Lambert W function is [148]:

$$\frac{dW(e^x)}{dx} = \frac{W(e^x)}{(1 + W(e^x))}. \tag{A.0.3}$$

The Lambert-W function allows formulating solutions in explicit form to equations of the type:

$$e^x x = a, \tag{A.0.4}$$

by applying $W()$ to both sides of the equality:

$$W(e^x x) = W(a) \tag{A.0.5}$$

$$x = W(a). \tag{A.0.6}$$

Then, the procedure for solving the equation is to manipulate it such that it can be expressed in the form $e^{f(x)} f(x) = a$.

The Taylor series of $W(x)$ around $x = 0$ has been known since Euler's paper [147]. The series can be derived usign the Lagrance Inversion Formula:

$$W(x) = \sum_{n=1}^{\infty} \frac{(-n)^{n-1}}{n!} = x - x^2 + \frac{3}{2}x^3 - \frac{8}{3}x^4 + \cdots \tag{A.0.7}$$

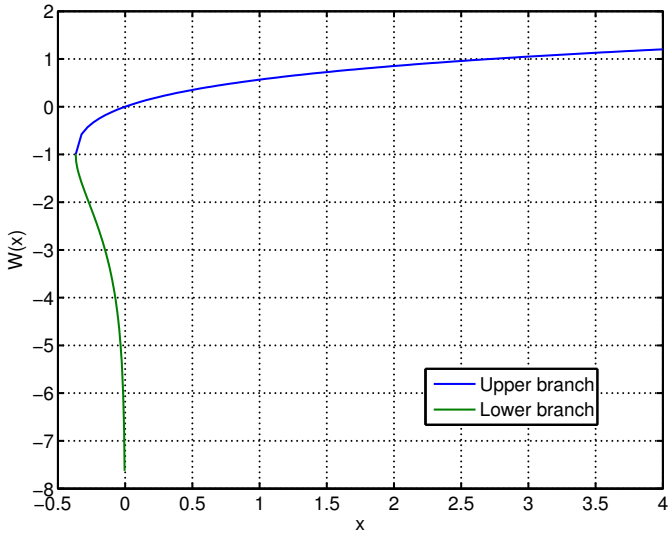

Figure A.1: Plot of the upper and lower branches of the Lambert-W function.

Figure A.2 plots the series approximation for the first terms. The approximation $W(x) = x$ is more or less accurate in the interval $(-e^{-1}, e^{-1})$.

The series expansion for $x = \infty$ is given in [149]:

$$W(x) = \log x - \log\log x + \sum_{n=1}^{\infty} \left(\frac{-1}{\log x}\right)^n \sum_{m=1}^{n} (-1)^m \left[\binom{n}{n-m+1}\right] \frac{(\log\log x)^m}{m!} \quad \text{(A.0.8)}$$

Figure A.3 plots the $W(x)$ and the series expansion for large x. The approximation $W(x) \approx \log(x)$ is not accurate for numerical computations but sufficient for limit or asymptotic analysis.

In [150], it is shown that

$$\log x - \log\log x < W(x) < \log x, \quad \text{(A.0.9)}$$

where the left hand side holds true for $x > 41.9$ and the right hand side holds true for $x > e$.

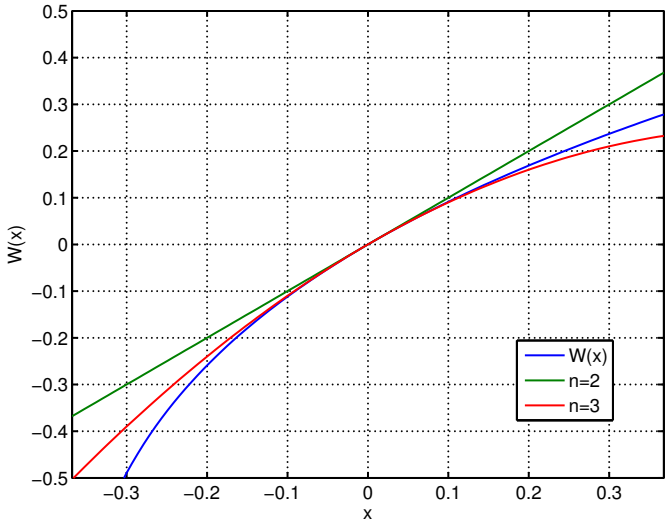

Figure A.2: Taylor approximation of the Lambert-W function around $x = 0$, with $n = 2$ and $n = 3$ terms.

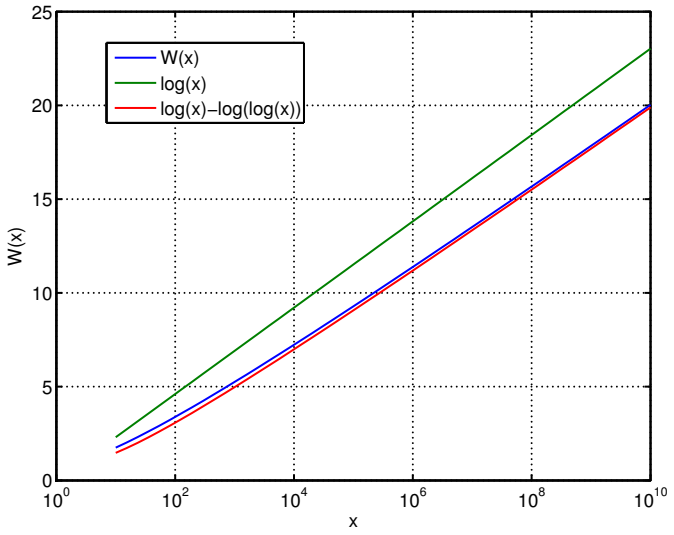

Figure A.3: Series expansion approximation of the Lambert-W function for $x \to \infty$, with 2 terms ($\log(x)$) and 3 terms ($\log(x) - \log\log(x)$).

Appendix B

Fractional Optimization

Fractional programs are nonlinear programs where the objective function is a ratio of two real-valued functions. For simplicity, only differentiable fractional programs, i.e. where both the numerator and the denominator are differentiable, are considered in this thesis. A general nonlinear fractional program has the form

$$\max_{x \in \mathcal{S}} \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}, \tag{B.0.1}$$

where $\mathcal{S} \in \mathrm{R}, f_1, f_2 : \mathcal{S} \to \mathrm{R}$ and $f_2(\mathbf{x}) > 0$. Problem (B.0.1) is called a concave-convex fractional program if f_1 is concave, f_2 is convex, and $\mathcal{S}$ is a convex set; additionally $f_1(\mathbf{x}) \geq 0$ is required, unless f_2 is affine. When f_1 and f_2 are differentiable, the objective function is pseudoconcave [151], implying that any stationary point is a global maximum and that the KKT conditions are sufficient if a constraint qualification is fulfilled. Because of this, (B.0.1) can be solved direclty by various convex programming algorithms [151]. However, when f_1 is concave and f_2 is convex, the fractional program can be transformed to an equivalent parametric convex program, which may be solved more efficiently in certain cases [122]:

$$F(\nu) \triangleq \max_{x \in \mathcal{S}, \nu \in \mathrm{R}} \quad f_1(\mathbf{x}) - \nu f_2(\mathbf{x}). \tag{B.0.2}$$

It is possible to show that $F(\nu)$ is convex, continuous and strictly decreasing in ν [128]. Furthermore, let q^* be the optimum value of the objective function of the original problem (B.0.1), then the following statements are equivalent [151]:

$$\begin{aligned} F(\nu) > 0 &\Leftrightarrow \nu < q^* \\ F(\nu) = 0 &\Leftrightarrow \nu = q^* \\ F(\nu) > 0 &\Leftrightarrow \nu > q^*. \end{aligned} \tag{B.0.3}$$

Thus, solving problem (B.0.1) is equivalent to finding the root of the nonlinear function $F(\nu)$.

Appendix C

Order Statistics Theory

Order statistics theory covers many different cases: independent or dependent variables, identically or non-identically distributed, and so forth. This section briefly introduces how the theory applies to solve the water level equation. Most of the contents of this appendix are found in the book [130].

Definition C.1. *The order statistics of a random samples $X_1, \ldots, X_n$ are the sample values placed in ascending order of magnitude. They are denoted by $X_{(1)}, \ldots, X_{(n)}$ or by $X_{1:n}, \ldots, X_{n:n}$. Then, order statistics are random variables that satisfy $X_{(1)} \leq X_{(2)} \leq X_{(n)}$.*

C.1 Distribution of Order Statistics

Theorem C.1. *Let $X_1, \ldots, X_n$ be a random sample from a discrete distribution with p.d.f. $f_X(x_i) = p_i$,where $x_1 < x_2 < \cdots$ are the possible values of X in ascending order. Define*

$$\begin{aligned} P_0 &= 0 \\ P_1 &= p_1 \\ P_2 &= p_1 + p_2 \vdots P_i = p_1 + p_2 + \cdots + p_i \vdots \end{aligned} \tag{C.1.1}$$

Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics from the sample. Then

$$P\left(X_{(j)} \leq x_i\right) = \sum_{k=j}^{n} \left(\binom{n}{k} \right) P_i^k \left(1 - P_i\right)^{n-k} \tag{C.1.2}$$

Proof. Fix i, and let Y be a random variable that counts the number of $X_1, \ldots, X_n$ that are less than or equal to x_i. For each $X_1, \ldots, X_n$, call the event $\{X_j \leq x_i\}$ a success and $\{X_j > x_i\}$ a failure. Then, Y is the number of success in n trials. Thus, $Y \sim$ binomial (n, P_i). The event $\{X_j > x_i\}$ is equivalent to $\{Y \geq j\}$; that is, at least j of the sample values are less than or equal to x_i. □

Theorem C.2. *Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of a random sample, $X_1, \ldots, X_n$, from a continuous population with c.d.f. $F_X(x)$ and p.d.f. $f_X(x)$. Then the p.d.f. of $X_{(j)}$ is*

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!\,(n-j)!} f_X(x) \left(F_X(x)\right)^{j-1} \left(1 - F_X(x)\right)^{n-j}. \tag{C.1.3}$$

Proof. See [130]. □

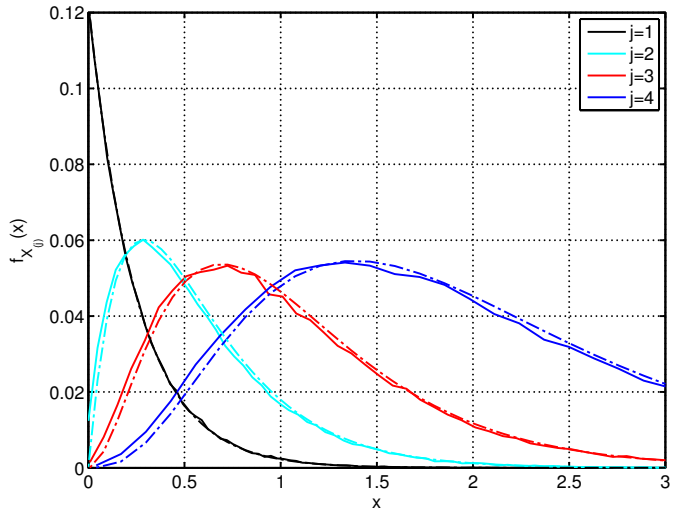

Figure C.1: Comparison of the empirical density function (dashed line) and the p.d.f. (line) given by (C.1.3) of $X_{(j)}$ for $n = 4$ exponentially distributed random variables.

Remark C.1. *The nth order statistic, or the sample maximum $X_{(n)}$ has the p.d.f.*

$$f_{X_{(n)}}(x) = n\left(F_X(x)\right)^{n-1} f_X(x) \tag{C.1.4}$$

Figure (C.1) plots the p.d.f. of $X_{(j)}$ given by (C.1.3) for a sequence of $n = 4$ exponentially distributed random variables. The distribution is compared with the empirical density function computed from 10,000 realizations. The expected p.d.f. correctly matches the empirical distribution, although the accuracy is better for small j. On the other hand, we can observe that the variance of the jth order statistic is higher for the largest order statistic.

C.1.1 Example: uniform order statistic

Let $X_1, \ldots, X_n$ be i.i.d. uniform$(0, 1)$, so $f_X(x) = 1$ for $x \in (0, 1)$ and $F_X(x) = x$ for $x \in (0, 1)$. Thus, the p.d.f. of the jth order statistic is

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!\,(n-j)!} x^{j-1}\,(1-x)^{n-j}\,, \tag{C.1.5}$$

for $x \in (0, 1)$. Hence, $X_{(j)} \sim \text{Beta}(1, n - j + 1)$. From this, we can deduce that

$$\mathbb{E}\{X_{(j)}\} = \frac{j}{n+1} \tag{C.1.6}$$

and

$$\mathbb{E}\{\left(X_{(j)} - \mathbb{E}\{X_{(j)}\right)^2\} = \frac{j(n-j+1)}{(n+1)^2(n+2)}. \tag{C.1.7}$$

C.2 Asymptotic Normality

For the uniform distribution, as $n \to \infty$, the pth sample quantile is asymptotically normally distributed, since it is approximated by

$$X_{(\lceil np \rceil)} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right). \tag{C.2.1}$$

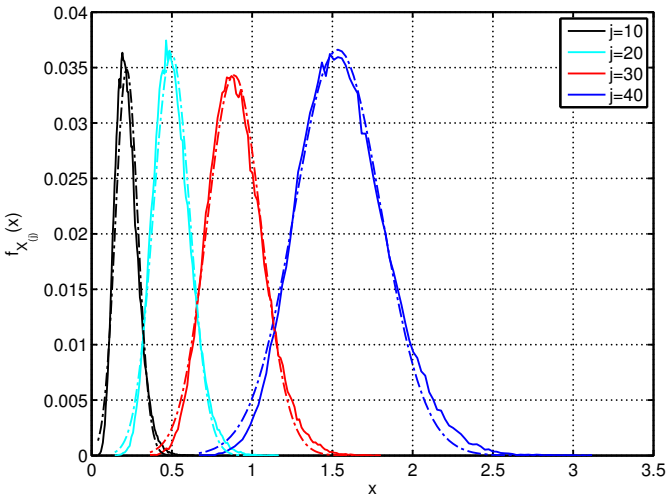

Figure C.2: Comparison of the empirical density function (dashed line) and the p.d.f. (line) given by (C.2.2) of $X_{(j)}$ for $n = 50$ exponentially distributed random variables.

For a general distribution, if the sample size n is sufficiently large and the random variables are i.i.d. with f and F the p.d.f. and c.d.f., it is possible to show [130, Ch. 9] that asymptotically:

$$X_{(\lceil np \rceil)} \sim \mathcal{N}\left(F_X^{-1}(p), \frac{p(1-p)}{n\left[f_X(F_X^{-1}(p))\right]^2}\right). \tag{C.2.2}$$

Figure C.2 plots the empirical density function and the p.d.f. given by (C.2.2) of the jth order statistic for $n = 50$ exponential random variables. We can observe that the jth order statistic is more deterministic (less variance) for small j. Figure C.3 plot the same distributions for a sample size of $n = 10$. The asymptotic normality approximation is inaccurate for small sample sizes in terms of p.d.f., specially for large j.

C.3 Extreme Value Theory

The nth order statistic is a special case called the extreme value. The asymptotic behaviour of the extreme value does not possess a limiting distribution, in general. It has been established that the limiting distribution of $X_{(n)}$, i.e. $\lim_{n\to\infty} F_{X_{(n)}}(x)$, if it exists, must be one of the following three types [130]

- Fréchet distribution with c.d.f.

$$F^{(1)}(x) = \exp\left(-x^{-\alpha}\right), \; x > 0, \alpha > 0; \tag{C.3.1}$$

- Weibull distribution with c.d.f.

$$F^{(2)}(x) = \begin{cases} \exp\left(-(-x)^{\alpha}\right), & x \le 0, \alpha > 0 \\ 1, & x > 0 \end{cases}; \tag{C.3.2}$$

- Gumbel distribution with c.d.f.

$$F^{(3)}(x) = \exp\left(-e^{-x}\right). \tag{C.3.3}$$

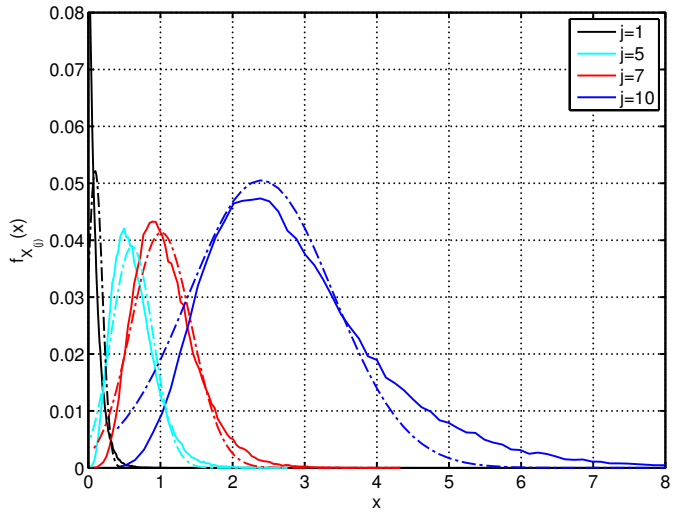

Figure C.3: Comparison of the empirical density function (dashed line) and the p.d.f. (line) given by (C.2.2) of $X_{(j)}$ for $n = 10$ exponentially distributed random variables.

The type of the limiting distribution depends on the properties of the distribution functions of the unordered random variables. It can be shown that if the distribution functions of X satisfy

$$\lim_{x\to+\infty} \frac{x f_X(x)}{1 - F_X(x)} = \alpha, \tag{C.3.4}$$

for $f_X(x)$ the p.d.f. of the random variables X and some constant $\alpha > 0$, then the limiting distribution of $X_{(n)}$ will be of Fréchet type. If, instead, the following condition is satisfied by the distribution of X

$$\lim_{x\to+\infty} \frac{1 - F_X(x)}{f_X(x)} = \alpha \tag{C.3.5}$$

where the constant $\alpha > 0$, then the limiting distribution of $X_{(n)}$ will be of Gumbel type.

For example, if X follow i.i.d. exponential distribution with p.d.f. $f_X(x) = e^{-x}$, condition (C.3.5) holds and, as such, the limiting distribution of $X_{(n)}$ is of Gumbel type.

Appendix D

List of Publications

D.1 Patents

Gomez-Miguelez, I. and A. Gelonch, "Metodo para optimizar la asignacion o distribucion de potencia de un transmisor en un sistema de comunicacion," Spanish Patent Application Number P201330948

D.2 Journals and Book Chapters

- Chapter 2: Enabling Technologies
  - V. Marojevic, **Gomez-Miguelez, I.**, and A. Gelonch, "Cognitive resource management: For all wireless access layers," *Vehicular Technology Magazine, IEEE*, vol. 7, no. 2, pp. 100–106, 2012
  - **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "ALOE: An open-source sdr execution environment with cognitive computing resource management capabilities," *Communications Magazine, IEEE*, vol. 49, September 2010
  - J. Torres, G. Reig, **Gomez-Miguelez, I.**, and X. Pegenaute, *Una vision del Cloud Computing desde una aula de la UPC*. Lulu Enterprises Incorporated, 2009
- Chapter 3: The SDR cloud
  - **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "Deployment and management of sdr cloud computing resources: problem definition and fundamental limits," *EURASIP Journal on Wireless Communications and Networking*, vol. 2013, no. 1, pp. 1–11, 2013
  - **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "Resource management for software-defined radio clouds," *IEEE Micro*, vol. 32, no. 1, pp. 44–53, 2012
  - V. Marojevic, **Gomez-Miguelez, I.**, P. Gilabert, G. Montoro, and A. Gelonch, "Resource management implications and strategies for sdr clouds," *Analog Integrated Circuits and Signal Processing*, vol. 73, no. 2, pp. 473–482, 2012
  - V. Marojevic, **Gomez-Miguelez, I.**, and A. Gelonch, "Tools for Analyzing Computing Resource Management Strategies and Algorithms for SDR Clouds," *Frequenz*, vol. 6, pp. 241–249, Sep. 2012
- Chapter 6: Efficiency Maximization Power Allocation

  - **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "Link adaptation for energy-efficient transmission with receiver csi," *Communications Letters, IEEE*, vol. 16, no. 9, pp. 1412–1415, 2012
  - **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "Processing-to-amplifier power ratio for energy efficient communications," *Electronics letters*, vol. 48, no. 12, pp. 732–734, 2012

- Chapter 7: Ordered Statistics Based EMPA
  - **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "Energy-efficient water-filling with order statistics," accepted for publication in the IEEE Transactions on Vehicular Technology

- Others:
  - **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "SDR receiver computational model for application in power control," *Analog Integrated Circuits and Signal Processing*, pp. 1–8, 2011
  - J. Salazar, **Gomez-Miguelez, I.**, and A. Gelonch, "Adaptive resource management and flexible radios for wimax," *Journal of Telecommunications and Information Technology*, no. 4, pp. 101–107, 2009

D.3 Conferences

- Chapter 2: Enabling Technologies
  - **Gomez-Miguelez, I.**, V. Marojevic, J. Bracke, and A. Gelonch, "Performance and overhead analysis of the ALOE middleware for sdr," in *Military Communications Conference, 2010 - MILCOM 2010*, 31 2010-nov. 3 2010, pp. 1134 –1139
  - **Gomez-Miguelez, I.**, H. Wang, A. Nafkha, V. Marojevic, C. Moy, P. Leray, and A. Gelonch, "Middleware extension for partial reconfiguration management in cognitive radios," in *Workshop on Software Radios. 6th Workshop on Software Radios (WSR 2010). Karlsruhe: 2010*, 2010, pp. 1–7
  - **Gomez-Miguelez, I.**, M. Camatel, J. Bracke, V. Marojevic, A. Gelonch, F. Vacca, and G. Masera, "ALOE-based flexible LDPC decoder," in *Digital System Design: Architectures, Methods and Tools (DSD), 2010 13th Euromicro Conference on.* IEEE, 2010, pp. 314–320
  - **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "Automatic computing resource awareness in resource managers for cognitive radios," in *Cognitive Information Processing (CIP), 2010 2nd International Workshop on.* IEEE, 2010, pp. 122–127
  - **Gomez-Miguelez, I.**, V. Marojevic, J. Salazar, and A. Gelonch, "A lightweight operating environment for next generation cognitive radios," in *Digital System Design Architectures, Methods and Tools, 2008. DSD '08. 11th EUROMICRO Conference on*, sept. 2008, pp. 47 –52
  - V. Marojevic, J. Salazar, **Gomez-Miguelez, I.**, and A. Gelonch, "Computing system modeling for flexible radios," in *Proc. Karlsruhe Worksh. Softw. Radio WSR08*, 2008

- Chapter 3: The SDR cloud

  - V. Marojevic, **Gomez-Miguelez, I.**, P. Gilabert, G. Montoro, and A. Gelonch, "Resource management strategies for sdr clouds," in *Wireless Innovation Forum. SDR'11 Technical Conference and Product Exposition, Washington D.C. 2011*, 2011, pp. 162–168

- Others:

  - **Gomez-Miguelez, I.**, V. Marojevic, J. Salazar, and A. Gelonch, "Analyzing the effect of power control algorithms on the receiver's computing resource consumption," in *Wireless Innovation Forum. SDR'10 Technical Conference and Product Exposition, Washington D.C. 2010*, 2010, pp. 1–6
  - J. Salazar, **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "Computing and radio resource management interactions in flexible radio environments," in *Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on.* IEEE, 2007, pp. 1–5

References

[1] "ITU-R Report M.2243: Assessment of the global mobile broadband deployments and forecasts for International Mobile Telecommunications," ITU-R, Technical Report, 01 2011.

[2] "UMTS Forum Report 44: Mobile traffic forecasts 2010-2020 report," UMTS Forum Report, Technical Report, 01 2011.

[3] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2017," Cisco, White Paper, 02 2013.

[4] "2020: Beyond 4G Radio Evolution for the Gigabit Experience," Nokia Siemens Networks, White Paper, 01 2011.

[5] "C-RAN: The Road Towards Green RAN," China Mobile Research Institute, Technical Report, 10 2011.

[6] D. Gesbert, S. Hanly, H. Huang, S. S. Shitz, O. Simeone, and W. Yu, "Multi-cell mimo cooperative networks: a new look at interference," *IEEE J.Sel. A. Commun.*, vol. 28, no. 9, pp. 1380–1408, Dec. 2010.

[7] B. Soret, H. Wang, K. Pedersen, and C. Rosa, "Multicell cooperation for lte-advanced heterogeneous network scenarios," *Wireless Communications, IEEE*, vol. 20, no. 1, pp. 27–34, 2013.

[8] X.-H. You, D.-M. Wang, B. Sheng, X.-Q. Gao, X.-S. Zhao, and M. Chen, "Cooperative distributed antenna systems for mobile communications," *Wireless Commun.*, vol. 17, no. 3, pp. 35–43, Jun. 2010.

[9] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, and S. Mohanty, "Next generation/dynamic spectrum access/cognitive radio wireless networks: a survey," *Comput. Netw.*, vol. 50, no. 13, pp. 2127–2159, Sep. 2006.

[10] **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "Resource management for software-defined radio clouds," *IEEE Micro*, vol. 32, no. 1, pp. 44–53, 2012.

[11] Y. Lin, L. Shao, Z. Zhu, Q. Wang, and R. K. Sabhikhi, "Wireless network cloud: Architecture and system requirements," *IBM Journal of Research and Development*, vol. 54, no. 1, pp. 4:1–4:12, 2010.

[12] M. W. Jonathan Segel, "LightRadio: White paper 1," Alcatel-Lucent, Tech. Rep.

[13] NSN, "Liquid Radio: Let traffic waves flow most efficiently," Nokia Siemens Networks, Tech. Rep.

[14] J. Salazar, **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "Computing and radio resource management interactions in flexible radio environments," in *Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on.* IEEE, 2007, pp. 1–5.

[15] J. Salazar, **Gomez-Miguelez, I.**, and A. Gelonch, "Adaptive resource management and flexible radios for wimax," *Journal of Telecommunications and Information Technology*, no. 4, pp. 101–107, 2009.

[16] J. Torres, G. Reig, **Gomez-Miguelez, I.**, and X. Pegenaute, *Una vision del Cloud Computing desde una aula de la UPC.* Lulu Enterprises Incorporated, 2009.

[17] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 1, pp. 68–73, Dec. 2008.

[18] J. Mitola, "The software radio architecture," *Comm. Mag.*, vol. 33, no. 5, pp. 26–38, May 1995.

[19] ——, "Software radios: Survey, critical evaluation and future directions," *IEEE Aerospace and Electronic Systems Magazine*, vol. 8, no. 4, pp. 25–36, Apr. 1993.

[20] E. Buracchini, "The software radio concept," *Comm. Mag.*, vol. 38, no. 9, pp. 138–143, Sep. 2000.

[21] R. Baines, "The dsp bottleneck," *Communications Magazine, IEEE*, vol. 33, no. 5, pp. 46–54, 1995.

[22] J. Mitola, "Technical challenges in the globalization of software radio," *Communications Magazine, IEEE*, vol. 37, no. 2, pp. 84–89, 1999.

[23] A. Salkintzis, H. Nie, and P. Mathiopoulos, "Adc and dsp challenges in the development of software radio base stations," *Personal Communications, IEEE*, vol. 6, no. 4, pp. 47–55, 1999.

[24] W. H. W. Tuttlebee, "Software-defined radio: facets of a developing technology," *Personal Communications, IEEE*, vol. 6, no. 2, pp. 38–44, 1999.

[25] K. Lange, G. Blanke, and R. Rifaat, "A software solution for chip rate processing in cdma wireless infrastructure," *Communications Magazine, IEEE*, vol. 40, no. 2, pp. 163–167, 2002.

[26] **Gomez-Miguelez, I.**, V. Marojevic, J. Salazar, and A. Gelonch, "A lightweight operating environment for next generation cognitive radios," in *Digital System Design Architectures, Methods and Tools, 2008. DSD '08. 11th EUROMICRO Conference on*, sept. 2008, pp. 47 –52.

[27] **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "ALOE: An open-source sdr execution environment with cognitive computing resource management capabilities," *Communications Magazine, IEEE*, vol. 49, September 2010.

[28] **Gomez-Miguelez, I.**, V. Marojevic, J. Bracke, and A. Gelonch, "Performance and overhead analysis of the ALOE middleware for sdr," in *Military Communications Conference, 2010 - MILCOM 2010*, 31 2010-nov. 3 2010, pp. 1134 –1139.

[29] V. Marojevic, **Gomez-Miguelez, I.**, and A. Gelonch, "Cognitive resource management: For all wireless access layers," *Vehicular Technology Magazine, IEEE*, vol. 7, no. 2, pp. 100–106, 2012.

[30] X. Revés, A. Gelonch, V. Marojevic, and R. Ferrús, "Software radios: unifying the reconfiguration process over heterogeneous platforms," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 2626–2640, January 2005.

[31] **Gomez-Miguelez, I.**, H. Wang, A. Nafkha, V. Marojevic, C. Moy, P. Leray, and A. Gelonch, "Middleware extension for partial reconfiguration management in cognitive radios," in *Workshop on Software Radios. 6th Workshop on Software Radios (WSR 2010). Karlsruhe: 2010*, 2010, pp. 1–7.

[32] **Gomez-Miguelez, I.**, M. Camatel, J. Bracke, V. Marojevic, A. Gelonch, F. Vacca, and G. Masera, "ALOE-based flexible LDPC decoder," in *Digital System Design: Architectures, Methods and Tools (DSD), 2010 13th Euromicro Conference on.* IEEE, 2010, pp. 314–320.

[33] **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "Automatic computing resource awareness in resource managers for cognitive radios," in *Cognitive Information Processing (CIP), 2010 2nd International Workshop on.* IEEE, 2010, pp. 122–127.

[34] 3rd Generation Partnership Project, "3GPP specification: 36.913 Rel-9; Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN)," 3rd Generation Partnership Project, Tech. Rep., 2009.

[35] ——, "3GPP specification: 36.913 Rel-10; Requirements for further advancements for Evolved Universal Terrestrial Radio Access (E-UTRA) (LTE-Advanced)," 3rd Generation Partnership Project, Tech. Rep., 2011.

[36] M. Hill and M. Marty, "Amdahl's law in the multicore era," *Computer*, vol. 41, no. 7, pp. 33 –38, july 2008.

[37] V. Marojevic, X. R. Balleste, and A. Gelonch, "A computing resource management framework for software-defined radios," *IEEE Transactions on Computers*, vol. 57, pp. 1399–1412, 2008.

[38] C. van Berkel, "Multi-core for mobile phones," in *Design, Automation Test in Europe Conference Exhibition, 2009. DATE '09.*, april 2009, pp. 1260 –1265.

[39] A. Bastoni, B. B. Brandenburg, and J. H. Anderson, "An empirical comparison of global, partitioned, and clustered multiprocessor edf schedulers," in *Proceedings of the 2010 31st IEEE Real-Time Systems Symposium*, ser. RTSS '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 14–24.

[40] B. B. Brandenburg, J. M. Calandrino, and J. H. Anderson, "On the scalability of real-time scheduling algorithms on multicore platforms: A case study," in *Proceedings of the 2008 Real-Time Systems Symposium*, ser. RTSS '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 157–169.

[41] K. Ramamritham, J. Stankovic, and W. Zhao, "Distributed scheduling of tasks with deadlines and resource requirements," *Computers, IEEE Transactions on*, vol. 38, no. 8, pp. 1110 –1123, aug 1989.

[42] Y.-K. Kwok and I. Ahmad, "Static scheduling algorithms for allocating directed task graphs to multiprocessors," *ACM Comput. Surv.*, vol. 31, no. 4, pp. 406–471, Dec. 1999.

[43] Y.-C. Lee and A. Zomaya, "A novel state transition method for metaheuristic-based scheduling in heterogeneous computing systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 19, no. 9, pp. 1215–1223, Sep. 2008.

[44] A. Khokhar, V. Prasanna, M. Shaaban, and C.-L. Wang, "Heterogeneous computing: challenges and opportunities," *Computer*, vol. 26, no. 6, pp. 18 –27, june 1993.

[45] D. Thain and M. Livny, "Building reliable clients and servers," in *The Grid: Blueprint for a New Computing Infrastructure*, I. Foster and C. Kesselman, Eds. Morgan Kaufmann, 2003.

[46] L. Smarr and C. E. Catlett, "Metacomputing," *Commun. ACM*, vol. 35, no. 6, pp. 44–52, Jun. 1992.

[47] S. H. Bokhari, "On the mapping problem," *IEEE Trans. Comput.*, vol. 30, no. 3, pp. 207–214, Mar. 1981.

[48] A. Iosup and D. Epema, "Grid computing workloads," *Internet Computing, IEEE*, vol. 15, no. 2, pp. 19 –26, march-april 2011.

[49] S. Sakr, A. Liu, D. Batista, and M. Alomari, "A survey of large scale data management approaches in cloud environments," *Communications Surveys Tutorials, IEEE*, vol. 13, no. 3, pp. 311 –336, quarter 2011.

[50] N. Doulamis, A. Doulamis, E. Varvarigos, and T. Varvarigou, "Fair scheduling algorithms in grids," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 18, no. 11, pp. 1630 –1648, nov. 2007.

[51] X. Liu, C. Qiao, D. Yu, and T. Jiang, "Application-specific resource provisioning for wide-area distributed computing," *Network, IEEE*, vol. 24, no. 4, pp. 25 –34, july-august 2010.

[52] R. Entezari-Maleki and A. Movaghar, "A probabilistic task scheduling method for grid environments," *Future Gener. Comput. Syst.*, vol. 28, no. 3, pp. 513–524, Mar. 2012.

[53] Z. Zhu, P. Gupta, Q. Wang, S. Kalyanaraman, Y. Lin, H. Franke, and S. Sarangi, "Virtual base station pool: towards a wireless network cloud for radio access networks," in *Proceedings of the 8th ACM International Conference on Computing Frontiers*, ser. CF '11. New York, NY, USA: ACM, 2011, pp. 34:1–34:10.

[54] G. D. Micheli, *Synthesis and Optimization of Digital Circuits*, 1st ed. McGraw-Hill Higher Education, 1994.

[55] **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "Deployment and management of sdr cloud computing resources: problem definition and fundamental limits," *EURASIP Journal on Wireless Communications and Networking*, vol. 2013, no. 1, pp. 1–11, 2013.

[56] J. Guo, F. Liu, and Z. Zhu, "Estimate the call duration distribution parameters in gsm system based on k-l divergence method," in *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on*, sept. 2007, pp. 2988 –2991.

[57] "Flexible wireless communications systems and networks (flexnets) web site," http://flexnets.upc.edu/.

[58] D. Gross and C. Harris, *Fundamentals of queueing theory*, 3rd ed., ser. A Wiley-Interscience publication. New York, NY [u.a.]: Wiley, 1998.

[59] A. Jagers and E. A. D. van, "On the continued erlang loss function," *Operations Research Letters*, vol. 5, no. 1, pp. 43–46, 1986.

[60] Y. Rapp", ""planning of junction network in a multi-exchange area"," in *ITC-4*, 1964, p. 4.

[61] V. Marojevic, **Gomez-Miguelez, I.**, P. Gilabert, G. Montoro, and A. Gelonch, "Resource management implications and strategies for sdr clouds," *Analog Integrated Circuits and Signal Processing*, vol. 73, no. 2, pp. 473–482, 2012.

[62] V. Marojevic, **Gomez-Miguelez, I.**, and A. Gelonch, "Tools for Analyzing Computing Resource Management Strategies and Algorithms for SDR Clouds," *Frequenz*, vol. 6, pp. 241–249, Sep. 2012.

[63] V. Marojevic, **Gomez-Miguelez, I.**, P. Gilabert, G. Montoro, and A. Gelonch, "Resource management strategies for sdr clouds," in *Wireless Innovation Forum. SDR'11 Technical Conference and Product Exposition, Washington D.C. 2011*, 2011, pp. 162–168.

[64] E. Telatar, "Capacity of multi-antenna gaussian channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, 1999.

[65] D. Tse and P. Viswanath, *Fundamentals of wireless communication.* New York, NY, USA: Cambridge University Press, 2005.

[66] M. Costa, "Writing on dirty paper (corresp.)," *IEEE Trans. Inf. Theor.*, vol. 29, no. 3, pp. 439–441, Sep. 2006.

[67] H. Harashima and H. Miyakawa, "Matched-transmission technique for channels with intersymbol interference," *Communications, IEEE Transactions on*, vol. 20, no. 4, pp. 774–780, 1972.

[68] M. Tomlinson, "New automatic equaliser employing modulo arithmetic," *Electronics Letters*, vol. 7, no. 5, pp. 138–139, 1971.

[69] C. Peel, B. Hochwald, and A. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part i: channel inversion and regularization," *Communications, IEEE Transactions on*, vol. 53, no. 1, pp. 195–202, 2005.

[70] T. K. Y. Lo, "Maximum ratio transmission," *Communications, IEEE Transactions on*, vol. 47, no. 10, pp. 1458–1461, 1999.

[71] M. Joham, W. Utschick, and J. Nossek, "Linear transmit processing in mimo communications systems," *Signal Processing, IEEE Transactions on*, vol. 53, no. 8, pp. 2700–2712, 2005.

[72] T. Yoo and A. Goldsmith, "Optimality of zero-forcing beamforming with multiuser diversity," in *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*, vol. 1, 2005, pp. 542–546 Vol. 1.

[73] S. Wagner, R. Couillet, D. Slock, and M. Debbah, "Large system analysis of zero-forcing precoding in miso broadcast channels with limited feedback," in *Signal Processing Advances in Wireless Communications (SPAWC), 2010 IEEE Eleventh International Workshop on*, 2010, pp. 1–5.

[74] C. Peel, B. Hochwald, and A. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part i: channel inversion and regularization," *Communications, IEEE Transactions on*, vol. 53, no. 1, pp. 195–202, 2005.

[75] R. A. Horn and C. R. Johnson, Eds., *Matrix analysis*. New York, NY, USA: Cambridge University Press, 1986.

[76] Q. Spencer and M. Haardt, "Capacity and downlink transmission algorithms for a multi-user mimo channel," in *Signals, Systems and Computers, 2002. Conference Record of the Thirty-Sixth Asilomar Conference on*, vol. 2, 2002, pp. 1384–1388 vol.2.

[77] L.-U. Choi and R. Murch, "A transmit preprocessing technique for multiuser mimo systems using a decomposition approach," *Wireless Communications, IEEE Transactions on*, vol. 3, no. 1, pp. 20–24, 2004.

[78] S. Kaviani and W. Krzymien, "On the optimality of multiuser zero-forcing precoding in mimo broadcast channels," in *Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th*, 2009, pp. 1–5.

[79] G. A. G. Heliotis, I.P. Chochliouros, "Fibre optic networks for distributed, extendible heterogeneous radio architectures and service provisioning: The case of the futon programme," *The Journal of The Institute of Telecommunications Professionals*, vol. 2, no. 3, pp. 113–118, 2008.

[80] F. Diehm, P. Marsch, and G. Fettweis, "The futon prototype: Proof of concept for coordinated multi-point in conjunction with a novel integrated wireless/optical architecture," in *Wireless Communications and Networking Conference Workshops (WCNCW), 2010 IEEE*, 2010, pp. 1–4.

[81] F. Hansen and F. Meno, "Mobile fading - rayleigh and lognormal superimposed," *Vehicular Technology, IEEE Transactions on*, vol. 26, no. 4, pp. 332–335, 1977.

[82] A. Abdi and M. Kaveh, "K distribution: an appropriate substitute for rayleigh-lognormal distribution in fading-shadowing wireless channels," *Electronics Letters*, vol. 34, no. 9, pp. 851–852, 1998.

[83] R. Xu, M. Chen, C. Tian, X. Lu, and C. Diao, "Statistical distributions of ofdm signals on multi-path fading channel," in *Wireless Communications and Signal Processing (WCSP), 2011 International Conference on*, nov. 2011, pp. 1 –6.

[84] A. Goldsmith, S. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of mimo channels," *Selected Areas in Communications, IEEE Journal on*, vol. 21, no. 5, pp. 684–702, 2003.

[85] A. Zanella and M. Chiani, "The pdf of the lth largest eigenvalue of central wishart matrices and its application to the performance analysis of mimo systems," in *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE*, 30 2008-dec. 4 2008, pp. 1 –6.

[86] A. T. James, "Distributions of Matrix Variates and Latent Roots Derived from Normal Samples," *Ann. Math. Statist*, vol. 35, no. 2, pp. 475–501, 1964.

[87] N. R. Goodman, "Statistical Analysis Based on a Certain Multivariate Complex Gaussian Distribution (An Introduction)," *The Annals of Mathematical Statistics*, vol. 34, pp. 152–177, 1963.

[88] A. T. James, "Distributions of Matrix Variates and Latent Roots Derived from Normal Samples," *Ann. Math. Statist*, vol. 35, no. 2, pp. 475–501, 1964.

[89] L. Ordonez, D. Palomar, and J. Fonollosa, "Ordered eigenvalues of a general class of hermitian random matrices with application to the performance analysis of mimo systems," *Signal Processing, IEEE Transactions on*, vol. 57, no. 2, pp. 672–689, 2009.

[90] S. Jin, X. Gao, and M. McKay, "Ordered eigenvalues of complex noncentral wishart matrices and performance analysis of svd mimo systems," in *Information Theory, 2006 IEEE International Symposium on*, 2006, pp. 1564–1568.

[91] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*, ser. Random Matrix Methods for Wireless Communications. Cambridge University Press, 2011.

[92] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Mathematics of the USSR-Sbornik*, vol. 1, no. 4, pp. 457–483, Oct. 2007.

[93] W. Zhang, G. Abreu, M. Inamori, and Y. Sanada, "Spectrum sensing algorithms via finite random matrices," *Communications, IEEE Transactions on*, vol. 60, no. 1, pp. 164–175, 2012.

[94] J. Hoydis, A. Muller, R. Couillet, and M. Debbah, "Analysis of multicell cooperation with random user locations via deterministic equivalents," in *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), 2012 10th International Symposium on*, 2012, pp. 374–379.

[95] F. Heliot, R. Hoshyar, and R. Tafazolli, "An accurate closed-form approximation of the distributed mimo outage probability," *Wireless Communications, IEEE Transactions on*, vol. 10, no. 1, pp. 5–11, 2011.

[96] S. Chatzinotas, M. Imran, and C. Tzaras, "Information theoretic uplink capacity of the linear cellular array," in *Telecommunications, 2008. AICT '08. Fourth Advanced International Conference on*, 2008, pp. 249–254.

[97] D. Gore, R. Heath, and A. Paulraj, "Transmit selection in spatial multiplexing systems," *Communications Letters, IEEE*, vol. 6, no. 11, pp. 491–493, 2002.

[98] P. Li, D. Paul, R. Narasimhan, and J. Cioffi, "On the distribution of sinr for the mmse mimo receiver and performance analysis," *IEEE Trans. Inform. Theory*, vol. 52, p. 2006, 2006.

[99] **Gomez-Miguelez, I.**, V. Marojevic, J. Salazar, and A. Gelonch, "Analyzing the effect of power control algorithms on the receiver's computing resource consumption," in *Wireless Innovation Forum. SDR'10 Technical Conference and Product Exposition, Washington D.C. 2010*, 2010, pp. 1–6.

[100] **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "SDR receiver computational model for application in power control," *Analog Integrated Circuits and Signal Processing*, pp. 1–8, 2011.

[101] V. Marojevic, J. Salazar, **Gomez-Miguelez, I.**, and A. Gelonch, "Computing system modeling for flexible radios," in *Proc. Karlsruhe Worksh. Softw. Radio WSR08*, 2008.

[102] G. Miao, N. Himayat, and G. Li, "Energy-efficient link adaptation in frequency-selective channels," *Communications, IEEE Transactions on*, vol. 58, no. 2, pp. 545 –554, february 2010.

[103] R. Prabhu and B. Daneshrad, "An energy-efficient water-filling algorithm for ofdm systems," in *Communications (ICC), 2010 IEEE International Conference on*, may 2010, pp. 1 –5.

[104] C. Isheden, Z. Chong, E. Jorswieck, and G. Fettweis, "Framework for link-level energy efficiency optimization with informed transmitter," *Wireless Communications, IEEE Transactions on*, vol. 11, no. 8, pp. 2946 –2957, august 2012.

[105] D. Halperin, B. Greenstein, A. Sheth, and D. Wetherall, "Demystifying 802.11n power consumption," in *Proceedings of the 2010 international conference on Power aware computing and systems*, ser. HotPower'10. Berkeley, CA, USA: USENIX Association, 2010, pp. 1–.

[106] O. Arnold, F. Richter, G. Fettweis, and O. Blume, "Power consumption modeling of different base station types in heterogeneous cellular networks," in *Future Network and Mobile Summit, 2010*, 2010, pp. 1–8.

[107] J. Lorincz, T. Garma, and G. Petrovic, "Measurements and modelling of base station power consumption under real traffic loads," *Sensors*, vol. 12, no. 4, pp. 4281–4310, 2012.

[108] C. Isheden and G. Fettweis, "Energy-efficient multi-carrier link adaptation with sum rate-dependent circuit power," in *GLOBECOM 2010, 2010 IEEE Global Telecommunications Conference*, dec. 2010, pp. 1 –6.

[109] Y. Chen, S. Zhang, S. Xu, and G. Li, "Fundamental trade-offs on green wireless networks," *Communications Magazine, IEEE*, vol. 49, no. 6, pp. 30–37, 2011.

[110] P. Grant, "Green radio - the case for more efficient cellular basestations," in *Keynote Speech in IEEE Int. Globecom'10*, 2010.

[111] "Earth: Energy aware radio and network technologies project," http://www.ict-earth.eu, accessed: 07-07-2013.

[112] "Opera-net: Optimising power efficiency in mobile radio networks project," http://opera-net.org/.

[113] "Wireless@kth, ewin: Energy-efficient wireless networking," http://www.wireless.kth.se/research/projects/19-ewin.

[114] D. Feng, C. Jiang, G. Lim, J. Cimini, L.J., G. Feng, and G. Li, "A survey of energy-efficient wireless communications," *Communications Surveys Tutorials, IEEE*, vol. 15, no. 1, pp. 167–178, 2013.

[115] R. G. Gallager, "Energy limited channels: Coding, multiaccess, and spread spectrum," MIT, Tech. Rep., 1987.

[116] S. Verdu, "Spectral efficiency in the wideband regime," *Information Theory, IEEE Transactions on*, vol. 48, no. 6, pp. 1319 –1343, jun 2002.

[117] S. Cui, A. J. Goldsmith, and A. Bahai, "Energy-constrained modulation optimization," *IEEE Transactions on Wireless Communications*, vol. 4, pp. 2349–2360, 2005.

[118] C. Isheden and G. Fettweis, "Energy-efficient link adaptation with transmitter csi," in *Wireless Communications and Networking Conference (WCNC), 2011 IEEE*, march 2011, pp. 1381 –1386.

[119] **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "Processing-to-amplifier power ratio for energy efficient communications," *Electronics letters*, vol. 48, no. 12, pp. 732–734, 2012.

[120] C. Isheden and G. Fettweis, "Energy-efficient link adaptation on a rayleigh fading channel with receiver csi," in *Communications (ICC), 2011 IEEE International Conference on*, june 2011, pp. 1 –5.

[121] **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "Link adaptation for energy-efficient transmission with receiver csi," *Communications Letters, IEEE*, vol. 16, no. 9, pp. 1412–1415, 2012.

[122] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, Mar. 2004.

[123] R. Prabhu and B. Daneshrad, "Energy minimization of a qam system with fading," *Wireless Communications, IEEE Transactions on*, vol. 7, no. 12, pp. 4837 –4842, december 2008.

[124] R. T. Rockafellar, *Convex Analysis (Princeton Mathematical Series)*. Princeton Univ Pr.

[125] M. Bagnoli, M. Bagnoli, T. Bergstrom, and T. Bergstrom, "Log-concave probability and its applications," *Econom. Theory*, vol. 26, 1989.

[126] S. Loyka, F. Gagnon, and V. Kostina, "Error rates of capacity-achieving codes are convex," in *ISIT*, 2010, pp. 325–329.

[127] C. Isheden and G. Fettweis, "Energy-efficient link adaptation with shadow fading," in *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, may 2011, pp. 1 –5.

[128] W. Dinkelbach, "On Nonlinear Fractional Programming," *Management Science*, vol. 13, no. 7, pp. 492–498, 1967.

[129] H.-C. Yang and M.-S. Alouini, *Order Statistics in Wireless Communications: Diversity, Adaptation, and Scheduling in MIMO and OFDM Systems*, 1st ed. New York, NY, USA: Cambridge University Press, 2011.

[130] H. A. David, *Order Statistics*, 2nd ed. Wiley, 1981.

[131] B. S. Krongold, K. Ramchandran, and D. Jones, "Computationally efficient optimal power allocation algorithm for multicarrier communication systems," in *Communications, 1998. ICC 98. Conference Record. 1998 IEEE International Conference on*, vol. 2, 1998, pp. 1018–1022 vol.2.

[132] J. Jang, K.-B. Lee, and Y.-H. Lee, "Transmit power and bit allocations for ofdm systems in a fading channel," in *Global Telecommunications Conference, 2003. GLOBECOM '03. IEEE*, vol. 2, 2003, pp. 858–862 Vol.2.

[133] A. Leke and J. Cioffi, "A maximum rate loading algorithm for discrete multitone modulation systems," in *Global Telecommunications Conference, 1997. GLOBECOM '97., IEEE*, vol. 3, 1997, pp. 1514–1518 vol.3.

[134] N. Papandreou and T. Antonakopoulos, "A new computationally efficient discrete bit-loading algorithm for dmt applications," *Communications, IEEE Transactions on*, vol. 53, no. 5, pp. 785–789, 2005.

[135] W. Al-Hanafy and S. Weiss, "Reduced complexity schemes to greedy power allocation for multicarrier systems," in *Microwave Radar and Wireless Communications (MIKON), 2010 18th International Conference on*, 2010, pp. 1–4.

[136] A. Mahmood and J. C. Belfiore, "Improved 3-db subgroup based algorithm for optimal discrete bit-loading," in *Sarnoff Symposium, 2008 IEEE*, 2008, pp. 1–5.

[137] G. Wunder, T. Michel, and C. Zhou, "Delay-limited transmission in ofdm systems: performance bounds and impact of system parameters," *Wireless Communications, IEEE Transactions on*, vol. 8, no. 7, pp. 3747 –3757, july 2009.

[138] J. Liu, J. Chen, A. Host-Madsen, and M. Fossorier, "Capacity-approaching multiple coding for mimo rayleigh fading systems with transmit channel state information," *Vehicular Technology, IEEE Transactions on*, vol. 56, no. 5, pp. 2965 –2975, sept. 2007.

[139] Y.-C. Liang, R. Zhang, and J. Cioffi, "Subchannel grouping and statistical waterfilling for vector block-fading channels," *Communications, IEEE Transactions on*, vol. 54, no. 6, pp. 1131 –1142, june 2006.

[140] N. Ermolova and B. Makarevitch, "Performance of practical subcarrier allocation schemes for ofdma," in *Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007. IEEE 18th International Symposium on*, sept. 2007, pp. 1 –4.

[141] X. He, S. Y. Wang, W. Zhang, and X. Y. Gan, "Ordered statistics based rate allocation scheme for closed loop mimo-ofdm system," in *Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, 2005. WIOPT 2005. Third International Symposium on*, april 2005, pp. 389 – 395.

[142] D. Dardari, "Ordered subcarrier selection algorithm for ofdm-based high-speed wlans," *Wireless Communications, IEEE Transactions on*, vol. 3, no. 5, pp. 1452 – 1458, sept. 2004.

[143] J. M. Cioffi, "Chapter 4: Multi-channel modulation." Stanford University, 2011.

[144] 3rd Generation Partnership Project, "3GPP specification: 36.212; Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding," 3rd Generation Partnership Project, Tech. Rep., Dec. 2009.

[145] **Gomez-Miguelez, I.** and A. Gelonch, "Metodo para optimizar la asignacion o distribucion de potencia de un transmisor en un sistema de comunicacion," Spanish Patent Application Number P201330948.

[146] **Gomez-Miguelez, I.**, V. Marojevic, and A. Gelonch, "Energy-efficient water-filling with order statistics," accepted for publication in the IEEE Transactions on Vehicular Technology.

[147] L. Euler, "De serie Lambertina Plurimisque eius insignibus proprietatibus," *Acta Acad. Scient. Petropol*, vol. 2, pp. 29–51, 1783.

[148] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the lambertw function," *Advances in Computational Mathematics*, vol. 5, pp. 329–359, 1996.

[149] R. M. Corless, D. J. Jeffrey, and D. E. Knuth, "A sequence of series for the lambert w function," in *Proceedings of the 1997 international symposium on Symbolic and algebraic computation*, ser. ISSAC '97. New York, NY, USA: ACM, 1997, pp. 197–204.

[150] M. Hassani, "Approximation of the Lambert W function," *RGMIA Research Report Collection*, vol. 8, 2005.

[151] S. Schaible and T. Ibaraki, "Fractional programming," *European Journal of Operational Research*, vol. 12, no. 4, pp. 325–338, April 1983.