Universität Paderborn, Germany

Reducing Energy Consumption of Radio Access Networks

Matthias Herlich

September 2013

Dissertation

submitted to the

Faculty of Electrical Engineering,
Computer Science, and Mathematics

in partial fulfillment of the requirements for the degree of

Doctor rerum naturalium (Dr. rer. nat.)

Referees:

Prof. Dr. Holger Karl
Prof. Dr. Friedhelm Meyer auf der Heide

Additional committee members:

Prof. Dr. Marco Platzner
Prof. Dr. Heike Wehrheim
Prof. Dr. Hans Kleine Büning

Submitted October 6, 2013
Examination February 18, 2014
Published February 25, 2014
Paderborn, Germany

Abstract

Radio access networks (RANs) have become one of the largest energy consumers of communication technology [LLH+13] and their energy consumption is predicted to increase [FFMB11]. To reduce the energy consumption of RANs different techniques have been proposed. One of the most promising techniques is the use of a low-power sleep mode. However, a sleep mode can also reduce the performance. In this dissertation, I quantify how much energy can be conserved with a sleep mode and which negative effects it has on the performance of RANs. Additionally, I analyze how a sleep mode can be enabled more often and how the performance can be kept high.

First, I quantify the effect of power-cycle durations on energy consumption and latency in an abstract queuing system. This results in a trade-off between energy consumption and latency for a single base station (BS). Second, I show that considering a network as a whole (instead of each BS individually) allows the energy consumption to be reduced even further. After these analyses, which are not specific for RANs, I study RANs for the rest of the dissertation.

RANs need to both detect and execute the requests of users. Because detection and execution of requests have different requirements, I analyze them independently. I quantify how the number of active BSs can be reduced if the detection ranges of BSs are increased by cooperative transmissions. Next, I analyze how more BSs can be deactivated if the remaining active BSs cooperate to transmit data to the users.

However, in addition to increasing the range, cooperative transmissions also radiate more power. This results in higher interference for other users which slows their transmissions down and, thus, increases energy consumption. Therefore, I describe how the radiated power of cooperative transmissions can be reduced if instantaneous channel knowledge is available. Because the implementation in real hardware is impractical for demonstration purposes, I show the results of a simulation that incorporates all effects I studied analytically earlier.

In conclusion, I show that a sleep mode can reduce the energy consumption of RANs if applied correctly. To apply a sleep mode correctly, it is necessary to consider power-cycle durations, power profiles, and the interaction of BSs. When this knowledge is combined the energy consumption of RANs can be reduced with only a slight loss of performance. Because this results in a trade-off between energy consumption and performance, each RAN operator has to decide which trade-off is preferred.

Zusammenfassung

Mobilfunknetze sind zu einem der größten Energieverbraucher der Informations- und Telekommunikationsbranche geworden [LLH+13] und es wird erwartet, dass ihr Verbrauch weiter steigt [FFMB11]. Um den Energieverbrauch von Mobilfunknetzen zu verringern, wurden verschiedene Techniken vorgeschlagen. Eine der vielversprechendsten Techniken ist der Einsatz eines stromsparenden Schlafmodus. Allerdings kann dieser Schlafmodus die Leistungsfähigkeit reduzieren. Meine Dissertation beschreibt, wie viel Energie durch einen Schlafmodus eingespart werden kann und welche Auswirkungen dieser auf die Leistungsfähigkeit hat. Zusätzlich untersuche ich, wie der Schlafmodus möglichst häufig genutzt werden kann und trotzdem die Leistungsfähigkeit erhalten bleibt.

Als erstes quantifiziere ich den Effekt der Ein- und Ausschaltdauer auf die Latenz und den Energieverbrauch in einem abstrakten Warteschlangensystem. Dies resultiert in einer Austauschbeziehung zwischen Energieverbrauch und Latenz für eine einzelne Basisstation. Als zweites demonstriere ich, dass der Energieverbrauch weiter reduziert werden kann, wenn man das gesamte Netz betrachtet (im Gegensatz zu jeder Basisstation einzeln). Nach diesen beiden Analysen, welche nicht speziell auf Mobilfunknetze zugeschnitten sind, betrachte ich im Rest der Dissertation nur noch Mobilfunknetze.

Mobilfunknetze müssen sowohl die Anforderungen der Nutzer detektieren als auch durchführen. Weil Detektion und Durchführung unterschiedliche Anforderungen haben, analysiere ich diese getrennt. Ich quantifiziere, wie die Anzahl der aktiven Basisstationen gesenkt werden kann, wenn ihre Detektierungsreichweite durch Kooperation erhöht wird. Als nächstes stelle ich dar, wie mehr Basisstationen abgeschaltet werden können, wenn die restlichen aktiven Basisstationen kooperieren, um Daten zu den Nutzern zu übertragen.

Allerdings erhöhen kooperative Übertragungen nicht nur die Reichweite, sondern strahlen auch mehr Leistung ab. Dies resultiert in höherer Interferenz bei anderen Nutzern, welche die Übertragungen verlangsamen und somit den Energieverbrauch wieder erhöhen. Deshalb beschreibe ich, wie mit augenblicklichem Kanalwissen die abgestrahlte Leistung verringert werden kann. Weil die Implementierung in echter Hardware für Demonstrationszwecke zu aufwändig ist, zeige ich zuletzt die Ergebnisse einer Simulation, welche alle vorher analytisch betrachteten Effekte zusammenfasst.

Mein Fazit ist, dass ein Schlafmodus den Energieverbrauch eines Mobilfunknetzes reduzieren kann, falls er korrekt verwendet wird. Um einen Schlafmodus korrekt zu verwenden, ist es notwendig Ein- und Ausschaltdauer, Leistungsprofile und die Interaktion von Basisstationen zu berücksichtigen. Wenn dieses Wissen kombiniert wird, kann der Energieverbrauch von Mobilfunknetzen reduziert werden, ohne dass die Leistungsfähigkeit nennenswert beeinträchtigt wird. Weil dies eine Abwägung zwischen Energieverbrauch und Leistungsfähigkeit darstellt, muss jeder Betreiber eines Mobilfunknetzes selbst entscheiden, was er vorzieht.

Table of contents

List of acronyms

BDP	bandwidth-delay product. 46, 50–52
BS	base station. iii, xii–xvi, 1–6, 8–18, 20, 21, 23, 24, 44, 48, 59–70, 72–75, 77–132, 157–159
CC	coherent combining. 13, 14, 99
DU	dense urban. 116, 117, 123–127
ERP	effective radiated power. xv, 95–97, 101–105, 108, 110–113, 119, 120, 130
LTE-A	Long Term Evolution Advanced. 13–15, 118, 131
MILP	mixed integer linear programming. 45, 53, 73
MIMO	multiple-input and multiple-output. 12, 15, 96, 97, 117, 120
MRC	maximal-ratio combining. 13, 14, 62, 76, 77, 99
MWC	minimum wireless coverage. 79
NP	nondeterministic polynomial time. 45, 46, 79
PTAS	polynomial-time approximation scheme. 79
QoS	quality of service. 132
RAN	radio access network. iii, 1–6, 9–12, 14–20, 44–46, 59–62, 73, 79, 81, 90, 94–96, 107, 114–116, 124, 128–132, 157–159
SDU	sparse dense urban. 117, 124, 125
SINR	signal-to-interference-(plus-)noise ratio. 101, 117–121, 125
SNR	signal-to-noise ratio. xvi, 4, 5, 63, 65, 68, 70, 76, 97, 110, 158

UE user equipment. xii, xiv–xvi, 2–6, 8, 9, 11–17, 21, 62, 63, 65, 66, 69, 72, 73, 77–81, 83–85, 88, 91, 94–98, 100, 104, 107–109, 111, 112, 115, 117, 118, 120–122, 124, 125, 129–131, 157–159

List of symbols

!	Factorial. 138, 139
∞	Infinity. 25, 36, 49, 50, 69–71, 105, 106, 137–139
A	Area of coverage. 64–67, 70, 71, 80, 83
A_1	Area a BS can cover. 82, 83, 89, 90
A_{1e}	Area that can be covered by *exactly* one BS. 82–87
A_{1ec}	Area that can be covered by *exactly* one BS using cooperative transmissions. 87, 88
A_2	Area that each of two BSs can cover. 82, 83, 85
A_{2e}	Area that can be covered by *exactly* two BSs. 82–86, 88
A_3	Area that each of three BSs can cover. 82–86, 88
A_d	Area that can be covered by *exactly* two cooperating BSs or one other single BS. 85, 87, 88
A_h	Area a BS has to cover if the user equipments (UEs) are assigned to the closest BS. 82–84, 89, 90
$AI(d, r, R)$	Size of intersection area of two circles with distance d and radii r and R. 83, 87
α	Rate of activation of a BS. 24–27, 29, 31, 37
$\mathcal{A}$	Random variable that describes the activation duration of a BS. xii, 11, 23–25, 28–37, 41–43, 118, 122, 126, 127
$\overline{\mathcal{A}}$	Expected value of the activation duration $\mathbb{E}[\mathcal{A}]$. 24, 31, 37, 40, 42–44
$AP(k)$	Accumulate and fire policy with an activation threshold if k. xiii, 24, 26, 33, 35, 38, 41, 121, 123, 124
A_s	Area that one BS has to be active to cover using sparse deployment. 89, 90
A_{se}	Area that can be covered by *exactly* two cooperative BSs in a sparse deployment. 89, 90
avg	Arithmetic mean operator (average). 49–53
C	Configuration. 47–53, 133, 135, 138, 157
c	Maximum number of BSs that can cooperate on a single transmission. xiv, 74–76, 98, 99, 101, 103, 105–107, 112, 113
cap	Capacity of a BS or edge. 47, 48, 51–53
C_L	Latency-minimal configuration. 47–51, 133, 135, 138

CP	Critical point, a point with the lowest received signal strength in the plane. 64, 67–70
$\mathrm{CR_L}$	Competitive ratio for latency. 25, 31–33, 35–37
$\mathrm{CR_P}$	Competitive ratio for power consumption. 25, 34–37
D	Set of demands. 47–53, 133
$\mathrm{d}(a, b)$	Euclidean distance between points a and b. 5, 63, 68–70, 73
D_D	Data rate users demand for data transmissions. xv, xvi, 4, 5, 118
δ	Path-loss exponent describing the loss of signal strength of an electromagnetic wave over distance. 5, 63–66, 68–71, 73, 75, 76, 80, 108
$\mathrm{dr}(\cdot)$	Function that maps from SNR to data rate. 5, 117–120
D_S	Data rate necessary for transmission of signaling traffic. xv, xvi, 4, 5, 63, 72
e	Base of the natural logarithm, also known as Euler's number (≈ 2.71828). 29, 50, 80, 137–139
$\mathrm{EA}(A)$	Function mapping the size of an area A to the probability of activity in this area. 80, 84, 86, 88, 90
$\mathrm{E_G}$	Edges of the graph G. xiii, 47, 48, 52, 53
EP	Energy-minimizing policy. 24, 31, 36
$\mathbb{E}[\cdot]$	Expected value operator. xii–xiv, xvii, 24, 25, 37
$\mathcal{F}$	Factor between average channel gains of BSs (sorted from high to low). 107–110, 113
G	A graph with nodes $\mathrm{V_G}$ and edges $\mathrm{E_G}$. 47
g	Gain of covered area compared to not using cooperation. 65, 66, 68, 70, 71
Γ	Expected channel gain (loss) $\mathbb{E}[\gamma]$. 96, 102, 104, 107
γ	Instantaneous channel gain (loss). xiii, xiv, xvi, 96–100, 102–107, 109, 111, 113
geo	Geometric mean operator. 50, 133, 138
GP	Greedy policy. 24, 25, 31, 33, 34, 36–39, 41–43, 121, 123–127
IC	Infinite cooperation scheme. 68, 70, 71
J	Joule, the unit of energy of the International System of Units (SI). 11, 110
k	Activation threshold of the accumulate and fire policy $\mathrm{AP}(k)$. xii, 24, 26, 27, 33–35, 38, 121, 123
L	Latency. 25, 26, 28, 31, 36, 38–43, 47–53, 123–127, 133, 138, 158
λ	Rate or density of arrivals of a (space-)time Poisson process representing user requests. 4, 6, 24–29, 31, 37–40, 44, 56–58, 80, 93

LC	Limited cooperation scheme. 65, 66, 71
$\mathcal{L}$	Random variable that describes the interarrival duration of requests. xiv, 24, 25, 28, 29
$\overline{\mathcal{L}}$	Expected value of the interarrival duration $\mathbb{E}[\mathcal{L}]$. 24, 29
LP	Latency-minimizing policy with an oracle. 24, 28, 29, 31, 36, 39, 41–43
m	Meter, the unit of length of the International System of Units (SI). 4, 117
MiB	Mebibyte, which is equal to 2^{20} byte. 117
$\mathcal{M}$	Random variable that describes the processing duration of requests. xiv, 24, 25, 30–37
$\overline{\mathcal{M}}$	Expected value of the processing duration $\mathbb{E}[\mathcal{M}]$. 24, 27, 37
μ	Processing rate of requests. 24–29, 31, 37
N	Mean noise power. xvi, 5, 97, 119, 158
NC	No cooperation scheme. 64, 65, 70, 71
O	Outage probability. 75, 97, 100, 103, 109, 111–113
O_{D}	Outage probability under dynamic association. 97, 99, 100, 102, 103
O_{De}	Probability that exactly c out of n BSs need to cooperate to serve a UE. 103, 106
OffP	Always-off policy. 120, 123, 124, 126
ω	Rate of deactivation of a BS. 24–27, 29, 31, 37
OnP	Always-on policy. 24, 28, 31, 36, 48, 120, 122–126
O_{S}	Outage probability under static association. 97, 98, 100
O_{Sp}	Outage probability under static association with power control. 101
$p(\cdot)$	Probability density function of the instantaneous channel gain γ. 96, 98, 99, 105, 106
$\phi(\cdot)$	Function from source and destination of demand to it size. 47–53, 133–137
π	Ratio between circumference and diagonal in every circle (≈ 3.14159). 65, 83
P	Power consumption. 6, 25, 26, 29, 31, 38, 39, 41, 42, 47, 48, 53, 56–58, 74–76, 81, 84–86, 88, 90, 92, 123, 126, 127
$\mathbb{P}[X]$	Probability of event X. 28, 29, 75, 97–99, 103, 106
$\mathbb{P}_{\mathrm{st}}[S]$	Probability to be in state S. 25, 28, 29
PR_{L}	Poisson ratio for latency. 25, 31
PR_{P}	Poisson ratio for power consumption. 25, 31
$\mathcal{P}$	Idle power consumption as fraction of maximal power consumption. 48, 53, 56–58, 118

Q	Expected number of active BSs per area. 81, 84, 91–93, 124, 125
R	Effective radiated power (ERP) of all BSs combined. 97, 110–113
r	Effective radiated power (ERP) of a single BS, which is equivalent to the mean transmit power. 96, 97, 101, 103–105
R_{D}	Radiated power under the dynamic association scheme. 103
r_{D}	Range of a BS in which it can provide data traffic with a data rate D_{D}. 159
R_{Dp}	Radiated power under the dynamic association scheme with power control. 105, 106
$\mathbb{R}$	Set of real numbers. 47
R_{S}	Radiated power under the static association scheme. 101
r_{S}	Range of a BS in which it can provide signaling traffic with a data rate D_{S}. 63–67, 69–73, 159
R_{Sp}	Radiated power under static association scheme with power control. 101
s	Second, the unit of time of the International System of Units (SI). 4, 118, 123–127
$\mathrm{S_A}$	BS state of being **a**ctive. 23, 25–27, 157
S^{AS}	Aggregated stretch determined by arithmetic mean of stretches. 49, 51, 53, 55–57, 134–137
$\mathrm{S_D}$	BS state of **d**eactivating. 23, 25–27, 157
S^{GS}	Aggregated stretch determined by geometric mean of stretches, equal to S^{SG}. xvi, 50, 51, 133, 136–138
$\sigma(e)$	Function mapping an edge e to the amount of data flow over it. 47, 48, 52, 53
skew	Skew operator, returns ratio of highest to lowest element. 51, 135–137
S^{MS}	Aggregated stretch determined by maximum of stretches. 48–50, 53, 135–137
S_n	Symmetric group, that is, the set of all permutations of the natural numbers $(1, \ldots, n)$. xvi, 99, 105, 106
SNR	Signal-to-noise ratio. xiii, 4, 5, 14, 73, 75, 97
SP(A)	Set cover policy using average activity in areas. 122–127
SP(C)	Set cover policy using currently active UEs. 122, 123
$\mathrm{S_S}$	BS state of **s**leeping. 23, 25–27, 159
S^{SA}	Aggregated stretch determined by stretch of arithmetic mean. 49, 51, 53, 55–57, 134–137

S^{SG}	Aggregated stretch determined by stretch of geometric mean, equal to S^{GS}. xv, 50, 133
S^{SM}	Aggregated stretch determined by stretch of maximum. 48, 49, 51, 53, 134–137
$\mathrm{S_U}$	BS state of activating (starting **u**p). 23, 25–27, 157
T	Threshold of signal-to-noise ratio (SNR) necessary for a UE to reliably decode transmissions. 4, 5, 13, 14, 63, 75
τ	A permutation out of the symmetric group S_n. 99, 105, 106
T_{D}	Threshold of SNR necessary for a UE to reliably decode transmissions at the demanded data rate D_{D}. xvi, 5, 97
T_γ	Threshold of the channel gain (loss) for a successful transmission, which is equal to $T_{\mathrm{D}}N$. 97–100, 102–107, 109, 111, 113
T_{S}	Threshold of SNR necessary for a UE to reliably decode transmissions at the rate necessary for signaling D_{S}. 5, 65, 68, 70, 73
U	Utilization of a BS or edge. 48, 51–53
UC	Unlimited cooperation scheme. 66–68, 71
$\mathrm{V_G}$	Vertices (nodes) of the graph G. xiii, 47, 53
W	Watt, the unit of power of the International System of Units (SI). 1, 118, 123–127
ξ	Spacing of BSs, also known as inter-site distance (ISD). 64–70, 82–85, 87, 89, 92, 159
ξ_{max}	Maximum spacing for the non-cooperative transmissions. 82, 83, 86, 91
ξ_{maxD}	Maximum considered spacing for the dense cooperative deployment. 85, 86
ξ_{maxSL}	Maximum spacing for the sparse deployment with limited cooperation. 88, 89
ξ_{maxSU}	Maximum spacing for the sparse deployment with unlimited cooperation. 88, 89
ξ_{min}	Minimum considered spacing for the non-cooperative transmissions. 82, 83
ξ_{minD}	Minimum considered spacing for the dense cooperative deployment. 85, 86
ξ_{minS}	Minumum considered spacing for the sparse deployment with cooperation. 88, 89
$\mathcal{Z}$	Random variable that describes the deactivation duration of BSs. xvii, 11, 23–25, 28, 29, 31–37, 41–43, 118, 122, 126, 127

$\overline{\mathcal{Z}}$	Expected value of the deactivation duration $\mathbb{E}[\mathcal{Z}]$. 24, 31, 37, 40, 42–44

List of figures

List of tables

1. Introduction

Information and communication technology is responsible for 2% of the world carbon-dioxide equivalent emissions. About 75% of these come from the electrical energy consumed during the use of the equipment [WO08]. The other 25% are consumed during production of the equipment.

It is sometimes stated that the carbon dioxide equivalent emissions of the information and communications technology have surpassed that of aviation. This comparison is not fair because it includes the cost of hardware manufacturing for communication technology, but not for aviation [MML+10, MBL13]. But as the communication sector grows faster than the aviation sector [Ols01, Cis13], the communication sector will overtake the aviation sector in dioxide equivalent emissions, *if* current trends (in growth and dioxide equivalent emissions) keep the same.

Additionally, reducing the energy consumption is not only an environmental concern, but also a cost factor. The electricity cost of the information and communication sector is already at an all time high (both absolute and relative to total electrical energy consumption) [LLH+13]. The price of electricity is predicted to further increase [U.S13, Aus12]. This will result in further increasing costs for operating electrical equipment.

One of the largest consumers of electrical energy in the information and communication sector are network operators [LLH+13]. The operators of radio access networks (RANs) face high power consumption [KAK+11] and high growth of user demands [Cis13]. The global energy consumption of all RANs is predicted to be 99 TWh in the year 2020 [FFMB11] resulting from a strong increase in end user demand in the next years [Rui11]. This is equivalent to 11.3 GW, which is close to the output of 12 nuclear power generating units (at a mean output of 975 MW [U.S12]). Future ubiquitous RANs that are able to support the growing number of users require to reduce their energy consumption.

The energy consumption of RANs is to a large degree determined by their base stations (BSs) [FFMB11]. The power consumption of BSs ranges from 10 W for small BSs to over 1 kW for large BSs [AGD+11]. In the year 2007 there were approximately 3.3 million BSs operating worldwide [FFMB11].

Methods to reduce their power consumption can be applied at different levels and using different techniques, such as more efficient power amplifiers and transmission technologies. In this dissertation, I compare different ways to reduce the energy consumption of entire RANs and not just individual BSs. The techniques I focus on are the use of a sleep mode and deploying BSs with different distances between each other. I focus on analytical approaches instead of simulations. While simulations can be more detailed (and thus realistic) the analytic approach provides more insights into the fundamental dependencies and relationships. I use these techniques to reduce the energy consumption

of RANs while considering how long users have to wait for their request to be processed. I create models which focus on different aspects and analytically evaluate them individually. Finally, I combine all aspects in a simulation to determine their interaction.

In this chapter, I first describe what RANs are and define what they do (section 1.1). To compare different RANs, I describe two performance metrics (energy consumption and latency) in section 1.2. Next, section 1.3 describes the signaling and data traffic split which allows me to analyze the need to provide data traffic to users and the need to reach every possible location with signaling traffic in two different models. As I use cooperative transmissions for both signaling and data traffic, I provide an overview of cooperative transmissions next (section 1.4). In section 1.5, I describe which alternative approaches to conserving energy in RANs exist and what distinguishes them from my approach. This section only summarizes the general approaches and is supplemented by a related work section in each of the following chapters. Finlly, I provide an overview of the contribution of the individual chapters and their approaches in section 1.6.

1.1. Radio access networks

In this section, I will provide an overview of what RANs are, what features they provide and how they usually fulfill them. Moreover, I describe how I model a RAN and which assumptions I make. These assumptions hold for the rest of the dissertation unless declared otherwise in the individual chapters. The model explained here forms the common basis for all chapters and is expanded in the individual chapters where necessary.

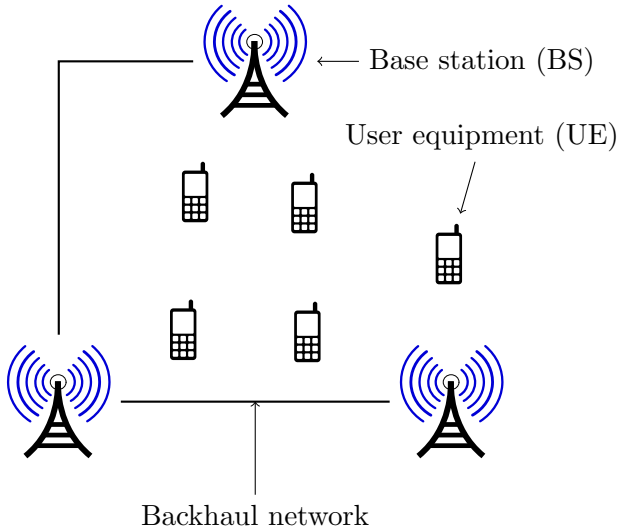

Figure 1.1.: UEs connect to close BSs, which provide connectivity to the internet and between UEs using a backhaul network.

I define a *radio access network (RAN)* as a telecommunication system that allows users

to wirelessly connect to the internet. Users can do this using, for example, mobile phones, notebooks, and tablets, which are collectively called *user equipment (UE)*. I denote the plural as UEs as is standard in the wireless communication community, even though the work "equipment" is uncountable. A RAN consists of stationary radio transceivers called *base station (BS)*. The BSs are connected with each other and to the internet by a backhaul network. The backhaul network is a mix of wired (e.g., copper or fiber optic connections) and wireless (e.g., microwave) links. Figure 1.1 illustrates the interactions of UEs, BSs, and the backhaul network.

UEs can, for example, be used to make phone calls, read eMails, and surf the world-wide web. For my work there is no need to distinguish between these different uses as, from the point of view of the RAN, they all are data transfers (with different sizes).

When a UE uses a RAN it executes the following steps. It creates a *request*, which I define as the need of a UE to exchange data. To grant this request the RAN decides to which BS the UE has to connect. I understand a *connection* as being the logical association of a UE with a BS, which allows them to exchange data. I define *transmission* as the act of exchanging data between UE and BS.

The most important aspect to determine the quality of radio transmissions is the path loss of wirelessly transmitted signals over distance. Because of path loss, wireless communication between UEs and BSs is only possible if the communicating partners are close together. Therefore, it is necessary to spread the BSs of a RAN over the area to which they have to provide their service (called *covering* an area).

A typical UE neither sends nor receives data (called *idle*) most of the time [FMK+10, Pau11]. Hence, there is no need for the BSs to maintain connections to the UEs at all times. Typically, a UE sends or receives data (called *active*) for a short time and becomes idle again thereafter. This activity period can be triggered both from the UE (e.g., an outgoing call) or from the network (e.g., an incoming call). Hence, both UE and BS must be able to initiate the connection between UE and BS. This is usually implemented by both partners listening for transmissions from the other partner.

Table 1.1.: All combinations of initiator and prevalent direction of traffic are possible, but video streaming is prominent.

Prevalent data direction	Initiated by UE	Initiated by the BS
More uplink data	Sending eMail	Remote access
Symmetric	Outgoing (video) call	Incoming (video) call
More downlink data	**Watching video stream**	Receiving eMail

While a transmission can be initiated both by a UE or by the BS, this does not necessarily mean that the initiator is also the source of the resulting transmission. Table 1.1 provides an overview with examples of the different possibilities to initiate and transfer data. Because video streaming is becoming the predominant consumer of data rate [Cis13], I focus on this. Using video streaming, traffic is initiated by the UE but is

mainly transmitted from the BS to the UE.

Initiating connections and transmitting data directly corresponds to the two distinct duties for each BS: (1) receive UE transmission requests (called *signaling traffic*) and (2) execute the transmissions (called *data traffic*). This distinction is important because I consider providing them from different BSs (section 1.3) and analyze both duties separately.

When the network wants to initiate a connection to a UE the network has to know which BS can connect to the UE. This can for example be achieved by a broadcast from all BSs the UE has been connected to recently [Sau10]. The BS which gets a response then establishes a connection to the UE. I simplify this and assume that a RAN always knows which BS is in range of a UE.

When the UE wants to initiate a connection, it sends a request, which is received by a nearby BS and the connection is established. When the connection is established the data is transmitted regardless of the traffic direction.

I separate both duties in the following way: the RAN has to (1) provide a small signaling data rate D_{S} to every possible UE location and (2) fulfill transmission requests of UEs at a higher data rate D_{D}. While the signaling traffic has to be provided to every possible UE location all the time, data traffic has only to be provided when demanded. I model these demands as being generated by a space-time Poisson process [Kin93]. That is, each request for a transmission occurs at a point in space and time. All request arrivals are independent of each other. One possibility to create such a space-time Poisson process is to create a Poisson process over time with exponentially distributed inter-arrival times and assign each arrival to a location. The location is determined by randomly selecting a position in the considered area, where every position is equally likely. In case of a space-time Poisson process the arrival rate λ is measured in requests per area and time (e.g., $1\,\mathrm{request/km^2/s}$).

I assume both UEs and BSs reside in a 2-dimensional plane. I consider both as points without size. In reality idle UEs move and continuously or periodically have to be provided with signaling traffic. From their current positions they then generate requests, randomly distributed in time. However, I do not model idle UEs and, thus, define requests as appearing randomly.

In chapters where I do not model time, I simplify this model to a spatial Poisson process. A spatial Poisson process places a Poisson-distributed number of requests in a given area. All requests are placed independently and with equal likelihood for each location in the area. In case of a spatial Poisson process the arrival rate λ is measured in requests per area (e.g., $1\,\mathrm{request/km^2}$).

Poisson processes have been shown to adequately model the timing of arriving transmission requests in RANs [WMBW09]. Another empirical study [Pau11] shows that the spatial correlation between load of BSs is low, which indicates that a Poisson modeling of user locations is acceptable.

Providing a given data rate to a UE (both for signaling and data traffic) requires the signal-to-noise ratio (SNR) to be higher or equal than a certain threshold T. Hence, a transmission is successful if

$$\mathrm{SNR} \geq T. \tag{1.1}$$

I assume the mean noise power is constant. This assumption results in a lower bound on the received signal strength. The required data rates for signaling D_{S} and data traffic D_{D} are different and, thus, also the required SNR to reliably transmit them. Therefore, I define the threshold of SNR for a transmission to be reliable as T_{S} for signaling traffic and T_{D} for data traffic. This mean I define the data rate by ergodic capacity instead of outage capacity (for a definition and comparison see section 4.7). This also means I consider the long term mean data rate and not the instantaneous data rate, which is influenced by fading. The conversion of the demanded data rate D_{D} to the threshold T_{D} depends for example on the modulation of the the transmission scheme, but is not in the scope of this dissertation. I assume a function dr to map the mean SNR to the resulting data rate. However, in all chapters except chapter 7 I do not use the function. In the other chapters I assume the required thresholds T_{D} and T_{S} to be given directly.

The signal strength mostly depends on the distance to the BS. While I only consider path loss analytically, I consider further effects such as fading and shadowing in the simulation in chapter 7. I model the loss of signal strength by the log-distance path-loss model [Sey05]. This results in the mean SNR to be calculated by

$$\mathrm{SNR} = \mathrm{d}(\mathrm{BS}, \mathrm{UE})^{\delta}/N, \tag{1.2}$$

where $\mathrm{d}(a, b)$ is the Euclidean distance of a and b, N is the mean noise power, and δ is the path-loss exponent. As the log-distance path-loss model needs a reference distance, but it only scales the result I assume it is 1. I define the highest distance at which the threshold T is still reached as *range* of a BS.

Some BSs have several antenna sectors, which can transmit independently. This increases the maximum capacity by a factor of 3 or 6, respectively. The increase in capacity is only important under high load. Because I consider low load scenarios and signaling traffic, I consider all BSs as omnidirectional. That is they have only a single antenna transmitting in all directions. The only exception to this is chapter 7, where I also consider high loads and, thus, also sectorized antennas. Because interference is also more important in high load scenarios I also include it in chapter 7. For all other chapters I assume the channels are noise-limited and ignore interference.

1.2. Load, energy consumption, and latency

The most important performance metrics I analyze are the energy consumption and the latency of demands. The energy consumption is the sum of the energy consumed by all BSs. The *latency* is the time that passes between a UE's request for a transmission and the end of the corresponding transmission. Both energy consumption and latency depend on the ratio of work it currently executes to the work it can maximally execute.

As BSs in industrial countries are all connected to the power grid (as opposed to some solar or Diesel-powered BSs in developing countries) [MLOH10], I consider only the total consumed energy and do not differentiate where the energy is consumed or how it is generated. Therefore, I define the *energy consumption* of a RAN as the total electrical energy consumed by all BSs . Also, I do not differentiate *when* energy is

consumed (e.g., to compensate for changing electricity prices). I do not consider the energy consumption of the UEs because they are not under control of the RAN operator. Moreover, most of the energy consumption UEs are responsible for is consumed during their production and not during their operation [HHA+11]. In contrast to this, most of the energy consumption BSs are responsible for is during their operation and not during their production [HHA+11]. Although the backhaul is under control of the operator I do not consider its energy consumption because it is small compared to the energy consumption of the BSs [FFMB11].

1.2.1. Power profiles

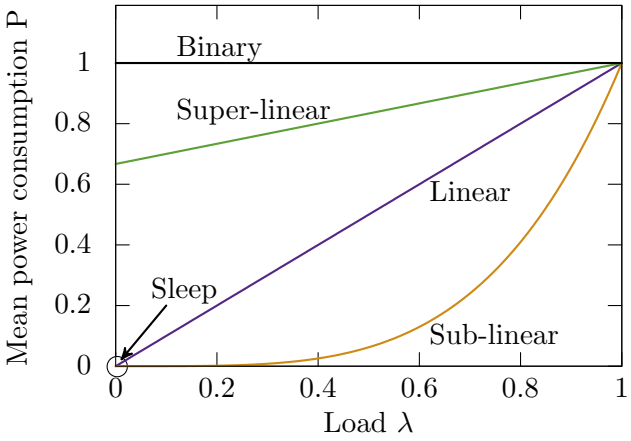

Figure 1.2.: These different categories of power profiles are typical for electrical devices.

The energy consumption of electrical devices is usually not constant but changes with the tasks it currently handles. An important parameter to determine energy consumption is the load. The *load* of a device is the ratio of work it currently executes to the work it can maximally execute. Different types of devices have different characteristic functions mapping load to power consumption. I call this function the *power profile* of a device. Figure 1.2 illustrates four different categories of power profiles: binary, super-linear, linear, and sub-linear. In a *linear* power profile the power consumption is always proportional to the load. In a *sub-linear* power profile the power consumption is lower than using the linear profile. In a *super-linear* power profile the power consumption is higher than using the linear profile. The *binary* power profile is a special case of the super-linear power profile which has constant power consumption if not in sleep mode.

On the smallest scales of time and space (i.e., the smallest component) digital equipment is usually either fully loaded or idle. On larger scales (of time and space) the power profile is composed from smaller scales. Methods to reduce the energy consumption on one scale result in a different power profile on a larger scale. Depending on how the load of a partially loaded component is distributed to its sub-components the composed

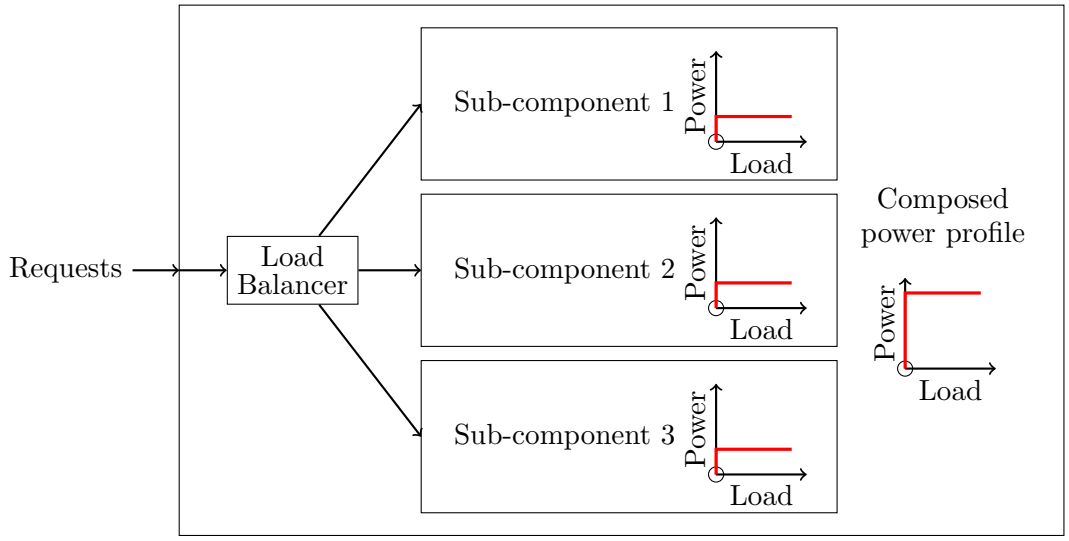

Figure 1.3.: The composed power profile of a component depends on the load distribution to its sub-components and their power profile. Load balancing recreates the power profile of the sub-components.

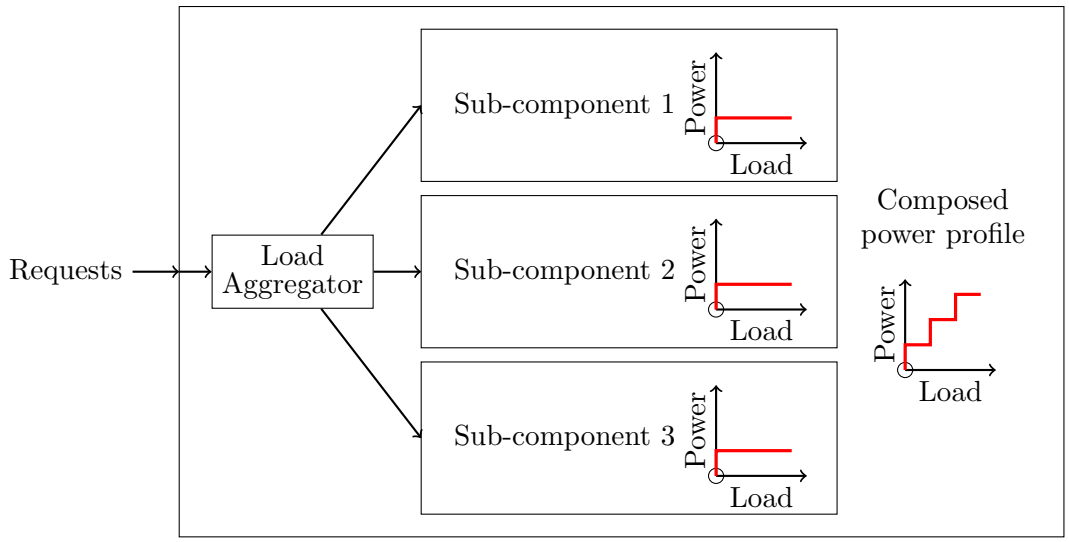

Figure 1.4.: The composed power profile of a component depends of the load distribution to its sub-components and their power profile. Load aggregation approaches a linear power profile if the number of sub-components approaches infinity.

power profile will look different.

If the load of a large component is equally distributed over all its sub-components the power profile of the large component will be the same as of its sub-components. For a component with a sub-linear power profile the power consumption is low for most loads and, hence, reproducing it is a good idea. But deadlines, unknown future demands, and unsplittable demands make it hard or impossible to perfectly balance load [KCLMT12].

If the power profile is super-linear, load balancing also recreates the power profiles of the sub-components. A component with a super-linear power profile will have high power consumption for most loads. However, in this case there is an alternative with less power consumption: load aggregation. Load can be aggregated in space or in time and ideally uses each sub-component at full load or not at all. In case it is not used at all it can be placed in a *sleep* mode which consumes less energy than running idle. However, in sleep mode it cannot serve any traffic (signaling or data) to UEs.

Aggregating the load results in a power profile that is linear between the power consumed in sleep mode and full power under full load. This is good for super-linear power profiles, but bad for sub-linear power profiles. Figures 1.3 and 1.4 illustrate an example for load balancing and aggregation over three sub-components (i.e., in space). Figure 1.5 illustrates load distribution over three time slots. This example assumes load can be freely moved in time and space, which is not true in most systems. Real systems will have deadlines which prevent load from being moved in time and the sub-components will have different functionalities which prevent load from being moved in space. Therefore, it is important to consider which negative effects the movement of load has. Moving load in time can, for example, be quantified by latency.

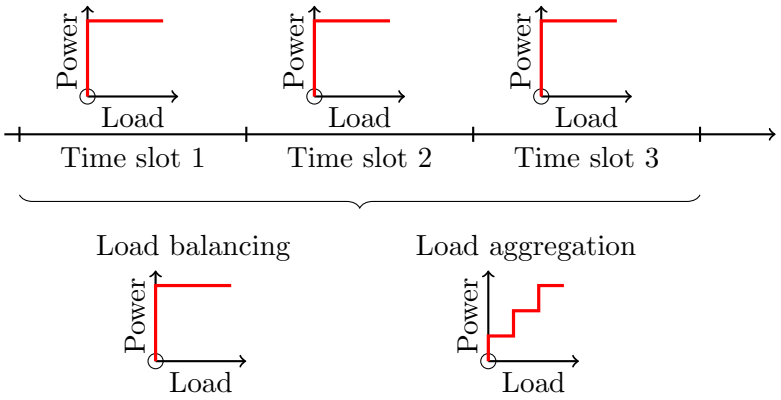

Figure 1.5.: The aggregated power profile of a large time slot depends on the load distribution to smaller time slots. Load balancing recreates the power profile of the smaller time slots. Load aggregation approaches a linear power profile.

To apply methods to conserve energy, it is necessary to first determine the power profile of the considered components. The power profile of a typical BS has been close to binary in the past [AGD+11, ARFB10, CH08, LGP12]. While the power profile of

future BSs will get closer to the linear profile, they will be far from achieving it [Fis07]. But a power profile closer to linear and higher-level energy conservation methods can be combined for even less energy consumption than each of them individually (see chapter 3). Additionally, the absolute energy savings by the techniques described in this dissertation will increase over time because the number of deployed BSs is expected to increase [FFMB11].

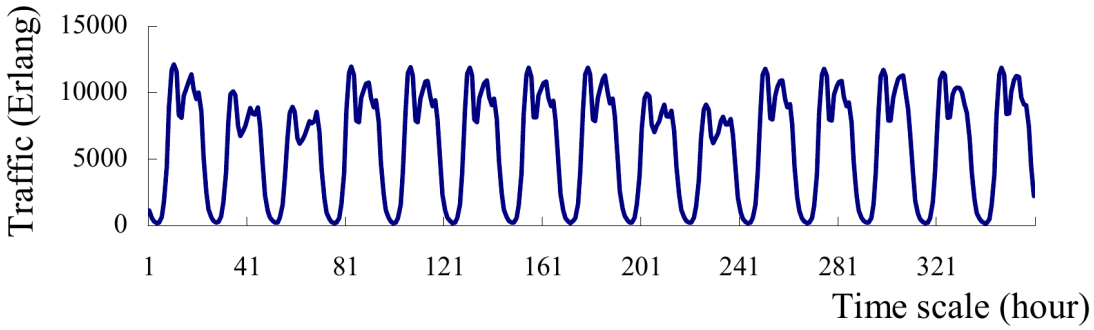

Figure 1.6.: The mobile traffic pattern repeats daily with a slight variation during weekends. The plot is based on measured data from China Mobile of Tianjin; both data and image are from Shu et al. [SYLY03].

To be most energy-efficient a system with a super-linear power profile should always run under full load. But because the traffic demands of the users of a RAN change over time, it is only possible to run a RAN under full load if demands are moved in time or dropped. Both are only acceptable in very limited amounts. The load changes over time in different forms:

- short-term (minutes and shorter) random fluctuations,
- medium-term (hours to days) periodic changes due to daily cycle (see figure 1.6), and
- long-term (months and longer) changes due to changes in usage behavior and population density.

The power profile of a complete RAN can be determined from the power profiles of the BSs and the load distribution scheme. By summarizing the power profiles of all BSs, it is possible to create a power profile of an entire RAN. If UEs are assigned to BSs without any concern for energy consumption, the load will be equally distributed to all BSs. Then all BSs will be under the same load as the RAN. Therefore, the power profile of the RAN will look the same as the power profile of each BS. But it is possible to change the power profile of the RAN to be closer to the linear power profile. The greatest difference between the super-linear power profile of a BS and the linear power profile is usually at low load [AGD+11]. Therefore, the greatest potential for energy

saving is achievable at low load. Because the fraction of time spent under low load is not negligible [SYLY03] I focus on low load. However, it is important to keep in mind that the RAN must still be able to handle high load when necessary.

1.2.2. Energy-latency trade-off

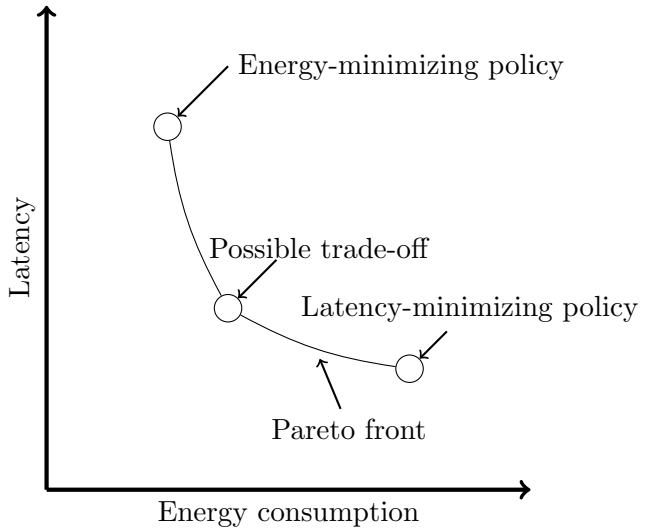

Figure 1.7.: Depending on the requirements each policy at the Pareto front is a viable candidate for the best policy.

Under high load either requests have to be dropped or the latency will increase. An alternative is to reduce reduce the load before this happens. However, I do not consider reducing the total traffic because those efforts are mostly application-specific. If the total amount of traffic is fixed, load can be moved in two ways: (1) move it in time and (1) move it in space. Most traffic cannot be freely moved in time but the loss of performance can be measured in latency. For a single type of technology this usually results in a trade-off between energy consumption and latency. Figure 1.7 shows how the trade-off looks like conceptually.

One possibility to move traffic in space is to move users. However, it is impractical to move users. But, due to the wireless nature of the transmission, most users can communicate with several nearby BSs. Hence, the load of one BS can be shifted to other BSs. On the one hand, this usually results in a worse channel, which results in slower transmissions and higher latency. On the other hand, it allows the deactivation of now idle BSs which in turn results in a lower energy consumption.

To reduce the energy consumption of a BS, it needs to be deactivated [ABH11]. Activation and deactivation consume time. During this time the BS also consumes energy. This delay and additional energy consumption are two reasons which make it non-trivial to decide when and which BSs to deactivate.

I assume the activation time $\mathcal{A}$ to be either constant or exponentially distributed. I make the same assumption for the deactivation time $\mathcal{Z}$. I call the time needed for activation and deactivation *power-cycle duration* and assume the BS consumes full power during that time. Model and analysis work the same way if activation and deactivation consumptions are different, but this makes results more complex.

Except for chapter 7, I only consider one type of BS in each chapter. Therefore, I normalize the maximum power consumption of a BS to 1. Additionally, I assume a binary power profile and the power consumption in sleep mode to be 0.

1.2.3. Consumption vs. efficiency

If the total traffic can be changed, reducing the energy consumption is not a useful goal because deactivating the system and ignoring all traffic would reduce the energy consumption to zero. For this reason, in cases where the traffic can be changed usually the energy efficiency of a system is considered. The *energy efficiency* can be defined as consumed (electrical) energy per unit of work done (e.g., J/bit). An alternative definition is to use the inverse: the amount of work done per unit of consumed energy (e.g., bit/J). Both definitions are technically equivalent, but differ in their ease of use. Which definition is easier to use depends on the scenario, but in general it is more convenient to have the quantity which is fixed in the denominator. This makes comparisons between the energy efficiency of systems easier because the gains are proportional to the increase in efficiency (in contrast to inversely proportional) [Kah11].

Because I consider the amount of traffic handled by the RAN to be fixed (with a small exception in chapter 7), changing the energy consumption of a RAN is equivalent to changing the energy efficiency. This is also equivalent to a reduction of the mean consumed power of the RAN. Because all energy and power consumption metrics discussed in this section are equivalent for my model I choose the one that is easiest to calculate: energy consumption.

1.3. Splitting signaling and data traffic

A BS can be deactivated only if another BS is in range of each UE and every location can still be reach with signaling traffic. When every BS is needed to provide signaling traffic to the UEs, no BSs can be put into sleep mode.

A new approach [CSF12] places two different types of BSs: *macro* BSs to provide signaling traffic and *pico* BSs to provide data traffic. Macro BSs have high ranges and only need to provide low data rates to the UEs. Pico BSs have small ranges but can provide high data rates to the users.

By splitting the BSs in two groups which fulfill two different duties, it is possible to deactivate the pico BSs if they do not transmit data and activate them again via the backhaul network when necessary.

When signaling and data traffic is served by different types of BSs it becomes easier to serve different amounts of data requests in an area by simply deploying more or fewer pico BSs. This is especially important when providing high data rates to hotspots.

Hotspots are areas in which the rate of request generation is considerably higher than in the surrounding area. Because the traffic in hotspots changes it is important to deactivate the pico BSs, which have been placed to provide high data rates, when the demands are low (e.g., at night).

To be able to conserve energy with a pico BS sleep mode a RAN has to be able to (1) determine the channel quality between a UE and nearby BSs, (2) connect a UE to another BS, possible in mid-transmission (called *handover*), and (3) deactivate idle pico BSs. The hardest part is to estimate the channel quality between a UE and a *deactivated* pico BS. This could for example be done by periodic transmission of a beacon from deactivated BSs.

These requirements can be fulfilled by modifying traditional handover schemes from one BS to the next. But redesigning the structure of RANs allows the macro BSs to consume low energy, to have long range, and to be designed to provide only the data rates needed for signaling. I assume the RAN operates using split signaling and data traffic. While BSs with other ranges exist [CHS08] I consider only macro and pico BSs for simplicity.

The split of signaling and data traffic between macro and pico BSs allows the energy consumption of signaling and data traffic to be analyzed independently from each other (chapters 4 and 5).

1.4. Cooperative transmissions

When each UE can only be served by a single BS there is no freedom to decide which BSs to activate and which to deactivate. Increasing the number of available pico BSs a UE can be assigned to can be achieved by deploying more BSs. This not only increases the setup costs of the RAN, but also its effect on the energy consumption is unclear because the maximum consumed power also increases (I analyze the effect of increasing the number of pico BSs on the energy consumption in chapters 5 and 7).

Another possibility of increasing the freedom to assign UEs to BSs is to increase the transmit power. This might not only be counterproductive for energy consumption, but is also largely prevented by law and inter-cell interference.

Another possibility to increase the range of BSs is to use cooperative transmissions from several BSs. The idea of *cooperative transmissions* is to transmit the same signal from more than one BS and combine the signal at the receiver. The combined signal is stronger than any of the individual signals and can thus reach a UE that is not in range of any of the individual BSs. Figure 1.8 illustrates how such a range extension works. Cooperation is sometimes also referred to as network multiple-input and multiple-output (MIMO) or distributed MIMO. In contrast to classical MIMO, where the antennas are close together, the antennas used for cooperation are further apart.

When the ranges of BSs are increased, fewer BSs are needed to cover an area. For every power profile which does not consume zero power when idle, reducing the number of BSs also reduces the idle power consumption of the RAN. Larger ranges increase the flexibility to serve traffic from more distant BSs. Therefore, it becomes possible to

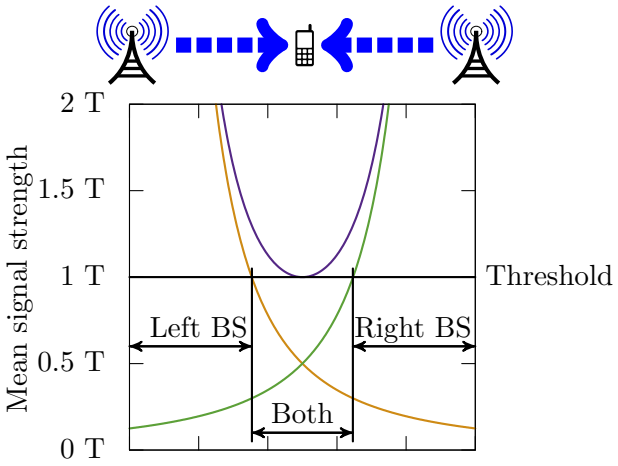

Figure 1.8.: Using cooperation the signal of two BSs is combined. Thus, they can provide service in an area where none of them individually could.

aggregate load on fewer BSs more often. If the load is aggregated to fewer BSs more BSs can be deactivated and thus less energy is consumed.

For downlink transmissions different implementations of cooperative transmissions are possible. One way to achieve this is to synchronously transmit from the cooperating BSs so that the signals constructively interfere at the receiver. This is called coherent combining (CC) [Gol05]. Another possibility is to transmit both signals at different times, record them in the receiver, and combine them inside the receiver. This is called maximal-ratio combining (MRC) [Pro01].

For uplink transmissions the receiving BSs can communicate the received signal to a single location (using the backhaul network) where the transmission is then decoded. One possibility to implement this in Long Term Evolution Advanced (LTE-A) is joint detection [MF11].

If two BSs cooperate, they both send the transmission to the UE, which combines the two transmissions into one. This allows the UE to decode the transmission, even if it could not decode a single transmission on its own. A similar method can be used for the uplink.

The requirements for cooperative transmissions differ for the different implementations and can be quite high (especially on the data rates and the latency of the backhaul network [BSC+12]). I ignore these requirements and the difficulties in their implementation and use cooperative transmissions as a tool and determine what can be gained by their use. I determine the energy gain of cooperation for both signaling (chapter 4) and data transmissions (chapter 5).

In chapter 4 I analyze how cooperation reduces the number of macro BSs needed to cover an area and how many macro BSs need to cooperate to achieve this. Additionally, I determine how cooperation can be used to decrease the fraction of time a pico BS is

needed to transmit data to UEs.

I focus on scenarios in which the network load is low, for example at night and in rural areas. In such scenarios the demanded data rates of UEs can easily be fulfilled, but the need to cover an area prevents macro BSs from being deactivated.

The ideas presented here are applicable to state of the art RANs, for example Universal Mobile Telecommunications System (UMTS) with macro diversity [SFH04] and LTE-A with coordinated multipoint transmission/reception (CoMP) [PDF+08].

I model cooperation as the possibility of a UE to decode a transmission if the sum of the received SNR(i) of BS i is at least T:

$$\sum_{i \in S} \mathrm{SNR}(i) \geq T, \tag{1.3}$$

where S is the set of cooperating BSs. This is a valid model for coherent combining (CC) [Gol05] and maximal-ratio combining (MRC) [Pro01].

As the shapes of the areas covered by these definitions are hard to determine geometrically (see section 4.7) I define a more restrictive, but geometrically easier to use type of cooperation. I call the type of cooperation defined by equation 1.3 *unlimited cooperation*. In contrast to this I define *limited cooperation* as

$$\forall i \in S : \mathrm{SNR}(i) \geq \frac{T}{|S|}, \tag{1.4}$$

where $|S|$ is the number of cooperating BSs.

1.5. General related work

In this section, I give an overview of other work which reduces the energy consumption of RANs. In this section, I summarize work which is related to all chapters. More specific related work sections are in every chapter.

1.5.1. Energy consumption on the BS level

Mancuso and Alouf [MA11] describe how the energy efficiency of each BS can be increased individually by using more energy-efficient antennas and power amplifiers. Song et al. [SDA+11] describe how deactivating carriers in a single BS can reduce their power consumption. Son and Krishnamachari [SK12] describe how individual components of BSs can be slowed down to conserve energy. While these papers describe how to reduce the energy consumption of an individual BS, I analyze how the energy consumption can be reduced on the network level.

An alternate approach by Fehske and Marsch [FMF10, MFF10] determines the energy efficiency of cooperative and dense deployments in terms of cost per bit. In contrast to my work they do not consider sleep modes and only consider fully loaded RANs. Other alternatives to considering the total energy consumption is to increase the energy efficiency of transmissions [CZXL11] and increasing the achieved data rates [PF11] without increasing energy consumption.

Son and Krishnamachari [SK12] describe how the speed of components of BSs can be scaled to apply dynamic speed scaling algorithms [Alb09]. In contrast to their work, I do not need to modify the internals of BSs and just use the sleep mode to conserve energy. The results of Son and Krishnamacharis work are an example of lowering power profiles, which I consider in chapter 3 .

1.5.2. Other wireless techniques

While cooperative transmissions are sent from antennas of different BSs, MIMO transmissions use similar principles but from different antennas of the same BS. While both are available in LTE-A [She10], I only consider cooperation between antennas of different BSs. An alternative I do not consider is cooperative transmission from at least two UEs to a single BS [NH04]. This is not only a problem because UEs are under user control, but most importantly it is impractical during low load because only few cooperation partners will be available. Another alternative is relaying of transmission, which in contrast to cooperation does not combine both signals at the receiver, from BS or UE [FSC11, HCM12, LWS$^+$10, NYZR06].

Other applications of cooperation [NH04] usually consider the additional problem of getting the transmission to the cooperating partner, but because BSs are usually connected by high-bandwidth, low-latency wired networks I assume they are able to do so reliably.

1.5.3. Energy consumption on the network level

My goal is to reduce the energy consumption of a single RAN of one technology. Alternatives include reducing the energy consumption of a network of mixed technologies [Ism11] and reducing the total energy consumption of several RANs by cooperation between operators [Ser12]. Marsan and Meo [MM11] describe how roaming UE between different operators can conserve energy. I do not consider fairness [KLZ09] to be of high priority because I only consider low load scenarios where the demands of most UEs can be fully met most of the time.

I do not explicitly consider the backhaul network [TZJ11]. While approaches exist that consider the limitations of the backhaul for cooperation [BSC$^+$12], adding backhaul to my models would increase the complexity. Because the BSs consume more energy than the backhaul network in RANs [VH11], I ignore the energy consumption of the backhaul.

1.5.4. Prediction

Paul et al. [Pau11] show empirically that the load of RAN is periodic and thus easy to predict, but the load of individual BSs is not. I only use predictions of aggregated future traffic to determine which type of traffic will be most dominant (video) [Cis13] to create a model that can represent this type of traffic.

It is possible to determine the intensity of a Poisson process for an area and track it over time [FN04, Lee91, Rat13, RB03]. While such techniques are necessary to determine

locations of hotspots and the change of network load over time I assume to know the location of hotspots.

1.5.5. Moving and reducing load

I consider the traffic the UEs generate to be only slightly movable in time and not to have (known) deadlines. I determine the mean latency of requests from the UEs. Alternative formulations of the problem consider the traffic to be fully movable in time [Sir02] and can thus be moved into low load periods [GBW95]. Other alternatives include deadlines for requests [CLMW07, CCLL07, Li09].

Another possibility to reduce the energy consumption of RANs is to reduce the demands it has to serve. These range from better content dissemination [BBEE08] in Content-Delivery networks [BPV08] to information-centric networking [KAK10]. Lee et al. [LRH10, LRKH11] consider the effects of information-centric networking on energy consumption, while I assume the demands to be fixed.

1.5.6. Effects on the power grid

My goal is to reduce the energy consumed by RANs. Other methods which are not necessarily specific to RANs or even computer networking try to consume power in areas where it is cheaper [QWB+09]. While this does not reduce the energy consumption it reduces the cost and environmental impact of the energy consumption. Hashimoto et al. [HYT05] describe how to build BSs which are independent of external energy sources. Felter et al. [FRKR05] describe how to reduce peak power consumption because reducing the variance in power consumption leads to a more efficient power grid [FMXY12].

While I consider how to run a communication network with less power, it is also possible to increase the efficiency of the power grid with the help of a computer network [Far10]. Last but not least computer networks can help to reduce the energy consumption and greenhouse gas emissions of other industrial or residential sectors [WO08]. The gained efficiency in other sectors is projected to be even larger than the gain in RANs [WO08], but for these gains to be useful RANs need to run energy efficiently themselves.

1.6. Contributions and chapter overview

In this section, I provide an overview of the contributions of the following chapters. Additionally I reference the publications they are based on.

Chapter 2 I begin by considering an individual BS as the server in a queuing system. Because this model does not depend on any specifics of a RAN, the results can be applied to a large class of other systems. Using this queuing model I quantify the statement that considering an individual BS with a sleep mode only reduces the energy consumption if power cycles are short.

- M. Herlich and H. Karl. Average and Competitive Analysis of Latency and Power consumption of a Queuing System with a Sleep Mode. In *Proceedings of the International Conference on Future Energy Systems: Where Energy, Computing and Communication Meet*, pages 14:1—-14:10, New York, NY, USA, 2012. ACM

Chapter 3 As considering each BS individually does not conserve much energy for long power cycles, I describe how all devices of a network can coordinate to reduce the energy consumption even if power cycles are long. I use the example of a wired network to ignore the complexities of wireless transmissions. I quantify the trade-off between the energy consumption and latency considering a network as a whole (instead of each device individually).

- M. Herlich and H. Karl. The Trade-Off between Power Consumption and Latency in Computer Networks. In Vicente Casares-Giner, Pietro Manzoni, and Ana Pont, editors, *Proceedings of the Networking Workshops*, volume 6827 of *Lecture Notes in Computer Science*, pages 273–280. Springer Berlin / Heidelberg, 2011

For the rest of the dissertation I apply the same idea of network-wide coordination to RANs. I use the split of signaling and data traffic (see section 1.3) to analyze the energy consumption of signaling independent from data traffic. The independent analysis makes it possible to analyze these models analytically. An analysis would be far more complex in a combined model.

Chapter 4 I describe how much energy can be conserved in RANs if BSs cooperatively transmit to extend their signaling range. Firstly, I determine the reduction in energy consumption when BSs are placed specifically for energy efficient signaling. Secondly, I determine the reduction when BSs are placed but can be deactivated.

- M. Herlich and H. Karl. Reducing Power Consumption of Mobile Access Networks with Cooperation. In *Proceedings of the International Conference on Energy-Efficient Computing and Networking*, pages 77–86, New York, USA, 2011. ACM

Chapter 5 I analyze the reduction of energy consumption to provide data traffic using cooperation. I quantify the activity probability for a BS depending on the activity of UEs.

- M. Herlich and H. Karl. Energy-Efficient Assignment of User Equipment to Cooperative Base Stations. In *Proceedings of the International Symposium on Wireless Communication Systems*, 2013

Both the reduction of consumed energy in signaling and data traffic depend cooperative transmissions. If cooperative transmissions are not coordinated correctly the radiated power becomes interference at other receivers. This reduced channel quality will increase the time necessary for their transmissions and thus increase energy consumption.

Therefore, it is necessary to keep the radiated power low to prevent disrupting other transmissions.

Chapter 6 I describe how interference from cooperating BSs can be reduced by selecting which BSs actually transmit based on *instantaneous* channel knowledge instead of *average* channel knowledge.

- M. Herlich and H. Karl. Analytic Quantification of Outage Probability and Radiated Power of Cooperative Base Stations. *In preparation*

Chapter 7 As each of the analytical chapters ignore some aspects of a real RAN, I present the results of simulating a RAN which includes all these aspects. This allows me to determine how all aspects interact. I determine the energy consumption of the different approaches with realistic dense urban parameters.

- M. Herlich, T. Hohenberger, and H. Karl. Activation Strategies for Low-Power Radio Access Networks. *In preparation*

Figure 1.9 shows which chapters introduce the concepts used in another chapter. It also provides an overview which chapters are specifically for RANs and which are not. Moreover, it shows which chapters provide analytical results and which are based on simulations. Table 1.2 shows in detail which effects I consider in which chapter.

Table 1.2.: I consider different effects in each chapter. ✓means explicitly considered, (✓) means implicitly considered. - means not applicable

Effect\Chapter	2	3	4	5	6	7
Path loss	-	-	✓	✓	(✓)	✓
Fading	-	-			✓	(✓)
Antenna sectors	-	-				✓
Cooperative transmissions	-	-	✓	✓	✓	✓
Coordinated deactivation	-	✓	✓	✓		✓
Progressing time	✓	(✓)	(✓)		-	✓
Interference	-	-			✓	✓
Overload	✓	✓				✓
Shifting load in time	✓					(✓)
Shifting load in space	-	✓	✓	✓		✓

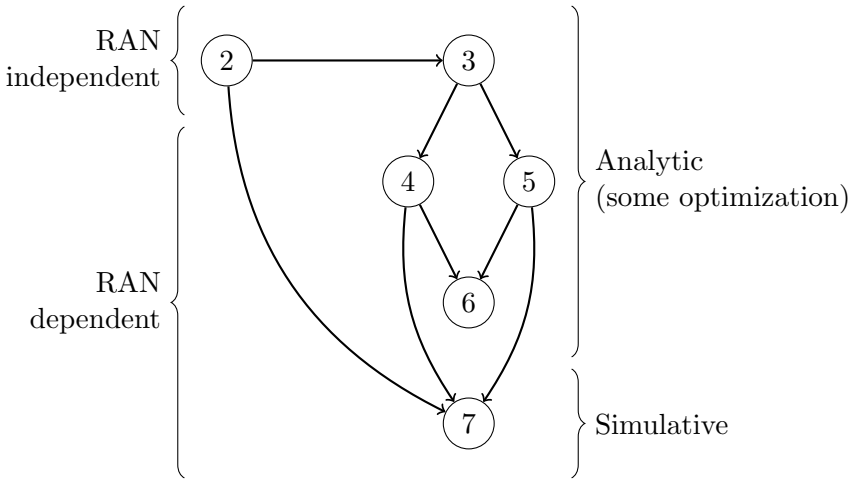

Figure 1.9.: The interdependency guide shows which chapter provides the basis for other chapters.

2. Conserving energy at each BS individually

All methods to reduce the energy consumption at each base station (BS) will directly translate into a reduction of energy consumption of the complete radio access network (RAN). To conserve energy in a RAN it is, thus, natural to try to conserve energy at each BS. In this chapter, I determine the limits of sleep modes for reducing the energy consumption if the sleep modes of the BSs are not coordinated.

I quantify the effect of times on latency and power consumption for a BS with a sleep mode. The calculations in this chapter do not depend on any special properties of BSs and, hence, can be applied to any single server queuing system with a sleep mode.

When a BS is in sleep mode it consumes only negligible power. But changing into sleep mode takes time and during that time power is also consumed. This power-cycle duration can also introduce additional latency into the system because jobs have to wait for the BS to become active again. Deactivating the BS more will often decrease the energy consumption but also increase latency. I first analyze the trade-off between energy consumption and latency for Poisson arrivals. Because the traffic a BS has to serve does not actually form a Poisson process, I also do a competitive analysis for worst-case arrivals. This allows me to describe whether the results are specific for Poisson arrivals or can be generally applied.

To calculate the latency and power consumption for Poisson arrivals I use Markov chains. For the competitive analysis I present arrival patterns which result in high latency and energy consumption and prove that these patterns are the worst case.

2.1. Introduction

To determine the increased latency I use a classical queuing system in which the server has the additional ability change into sleep mode. Being in this sleep mode consumes less power than being active, but the power cycle consumes time and energy.

Different policies can be applied to change between active and sleep mode. For example, a greedy policy deactivates the server as soon as no jobs are in the system and activates the server again when a job arrives.

I want to determine expressions for the worst-case and average latency and power consumption, depending on the power-cycle duration of a BS. Also, I want to quantify the following (intuitively plausible) statements for the different assumptions about the job arrivals: “It is easy to conserve energy, but hard (or impossible) to keep the latency low” and “With higher power-cycle durations, using a sleep mode becomes less and less practical.”

As the assumptions of this chapter are valid for a wide range of queuing systems I formulate them independently of user equipments (UEs) and BSs. Hence, I use the terminology which is usually used in queuing systems (jobs and servers).

2.2. Related work

Chen et al. [CX07, CXSY09] describe a Markov chain to calculate the power consumption of the greedy policy and the accumulate and fire policy. My model is the same as theirs, but I expand upon their work and additionally calculate the mean latency of a job and compare this to the latency obtained when using an oracle. While they focus on the average case for greedy and accumulate and fire policy, I compare this to the optimal values and calculate the competitive ratio.

Bredenbals analyzed the behavior of system with delayed activation and deactivation in his master's thesis [Bre13]. It is based on the same publication [HK12] as this chapter. An interesting result of his work is that it is possible to select activation and deactivation delays, such that the resulting system uses less energy than the greedy policy and at the same time has a lower mean latency.

Irani et al. [ISG02] analyze the competitive ratio of systems with multiple power saving states. In their model the only penalty for a state change is the consumed energy; the model does not include the time necessary for the change. Thus, they make no statements about the effect of conserving energy on the latency. In contrast to this, I describe both latency and power consumption to characterize the trade-off between them.

Ren et al. [RKM05] focus on hierarchical scenarios with non-stationary service requests, but also provide a review of the dynamic power management problem. This includes the power consumption and performance degradation of a system with constant state-change times.

Andrew et al. [AWT09] investigate how to minimize a linear combination of energy and response time. They show that Shortest Remaining Processing Time scheduling is 2-competitive. Stidham [Sti70] also converts waiting time and energy cost into a single metric which then can be minimized. Lam et al. [LLT+09] minimize the sum of energy and latency. Zhang and Chanson [ZC05] limit the average delay and try to minimize the consumed power. In contrast to my model, none of these models considers the time necessary to change states.

Authors providing overview papers of energy-efficient queuing include: Pruhs [Pru07] and Albers [Alb09], who provide overviews of competitive analyses of scheduling problems, Irani and Pruhs [IP05], who provide a description of both open and solved algorithmic problems with respect to power management, and Lu and De Micheli [De 01], who describe an oracle and its application to different traces. Benini et al. [BBPD99] describe how dynamic power management schemes can be modeled using Markov models.

Others [TzJwHj04, Ped99] propose alternative policies to conserve energy. The new policies are usually compared to other polices, while I compare the greedy policy to the optimal values.

Li [Li09] analyses the competitive ratio of a finite-capacity queue for jobs with hard

deadlines. While this model is more detailed than mine it does not consider energy consumption. Baptiste et al. [BCD07] provide a polynomial time algorithm to calculate the minimum-energy schedule for a given set of jobs. While they provide an offline algorithm, I analyze online algorithms.

Instead of limiting the latency by deadlines, Chan et al. [CLMW07] limit the energy consumption and calculate the competitive ratio for throughput. I do not limit any resource but analyze the interaction between latency and energy consumption described by competitive ratios and averages.

Heyman [Hey69] models similar problems as an M/G/1 with (de-)activation costs and converts waiting time into the same costs. He determines that the optimal policy is the accumulate and fire policy used by Chen [CX07] and in this dissertation. Heyman considers the arrivals to be a Poisson process, while I also consider competitive ratios.

Methodically different approaches have been considered to develop and analyze power management policies as well: machine learning [BGN+10], Petri nets [QW00], and model checking [NP02]. Each of these uses different techniques to approach problems related to mine.

Another type of work focuses on realistic values for power consumption and job arrival patterns. Prominent examples are web servers [BEK+02] and hard drives [SBdM00]. They use traces to compare different energy conservation methods, while I focus on the worst case for arbitrary systems.

2.3. Model

2.3.1. Queuing system

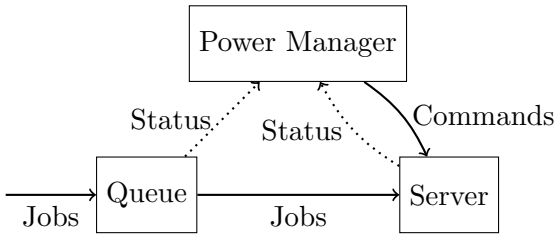

Figure 2.1.: In my model jobs arrive at the queue to be processed by the server. The power manager observes both the queue and the server and initiates state changes of the server.

Figure 2.1 shows the model that I use for power cycles. It consists of jobs, a queue, a server, and a power manager. The jobs arrive at the system and have to be processed by the server. Because the server cannot handle more than a single job at a time, the queue buffers arriving jobs until the server can process them in a first-in-first-out (FIFO) manner. The size of the queue is unlimited. The power manager observes the server and the queue and initiates activation and deactivation of the server.

While the model that only a single job can be processed at a time is widely used in computing it does not hold for BSs which process transmission requests concurrently. They will usually use a scheduling policy to fairly distribute the radio resource to all currently running transmissions. But under the assumption that the durations of jobs are exponentially distributed, processing one job after the other and processing all jobs in parallel is equivalent. The reason for this is that the exponential distribution is memoryless and, thus, the finishing rate of jobs is the same independently of the processing strategy. Therefore, the distribution of number of jobs in the system is the same and, thus, also the mean latency (see Little's law [Nel95]). Note that this is only valid as long as the server does not know the processing time in advance and can base its scheduling on it.

Because the M/M/1 system I describe models that a job is finished faster when it is alone in the system compared to a crowded system, it is a good representation for data transfers. As the duration of voice (and video) calls does not depend on the load of the BS, a M/M/1 is not a good model for it. For voice traffic a M/M/k model (without queue) is more suited, where k is the maximum number of concurrent calls a BS can handle. However, as I consider data transfers a M/M/1 is the more suitable model.

2.3.2. Server states and timing

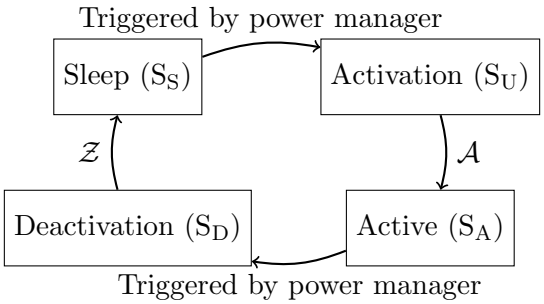

Figure 2.2.: State changes of the server are triggered by the power manager or take a random time to complete – depending on the type of state change.

Figure 2.2 shows all states of the power cycle of the server: Sleeping (S_S), activating (starting up) (S_U), active (S_A), and deactivating (S_D). The changes from deactivating to sleeping and from activating to active are determined by the (random) variables $\mathcal{A}$ and $\mathcal{Z}$, while the other two are triggered by the power manager. I assume the power manager can only issue the state changes stated above. It cannot change the state from deactivating to activating or change the order of jobs in the queue. Other models for example allow the server to abort the deactivation and directly enter the active state without the penalty of going through the activation state again. However, I chose my model as it is the most restrictive and the results can also be applied to the other models, but not the other way around.

The following four values describe the behavior of the system. (1) The inter-arrival time of jobs $\mathcal{L}$ (at rate λ and mean $\mathbb{E}[\mathcal{L}] = \overline{\mathcal{L}}$). (2) The time $\mathcal{M}$ the server needs to process a job (at rate μ and mean $\mathbb{E}[\mathcal{M}] = \overline{\mathcal{M}}$). (3) The time $\mathcal{Z}$ the server needs to deactivate (at rate ω and mean $\mathbb{E}[\mathcal{Z}] = \overline{\mathcal{Z}}$). (4) The time $\mathcal{A}$ the server needs to activate (at rate α and mean $\mathbb{E}[\mathcal{A}] = \overline{\mathcal{A}}$).

I analyze scenarios in which the four variables are determined in different ways: (1) random variables that follow exponential distributions, (2) constant values, and (3) values selected by an adversary for construction of the worst case. For all different cases I only consider scenarios which have infinite length to ignore transient effects. This is usually done in case (1) and easy to construct by repeating a scenario infinitely often in cases (2) and (3).

Using Markov chains the exponential distribution is easiest to analyze. For the competitive analysis constant times are easiest to analyze. But in addition to being easier or more complicated to analyze, different systems will have different distributions of power-cycle durations. For very simple system (e.g., a hand-held calculator) the deactivation is zero because the power can simply be turned off. More complex systems (e.g., a personal computer) will have to perform tasks before it can be go into sleep mode. This can include informing connected devices that it is going into sleep mode and preparing the sleep mode itself. The time this preparation takes might depend on the system state, for example, the time it takes for a computer to suspend-to-disk depends on the amount of RAM that has to be written to the disk.

The activation time can also vary depending on what the system has to prepare to fulfill its functions. A computer might run periodic integrity and virus scans or upgrade the system. Networked systems (such as BSs) need time to activate their networking components and synchronize their state with their neighbors or the rest of the network. Because many of these effects are hard to predict and are usually outside the control of the power manager, using a randomly distributed power-cycle duration is reasonable. The reason to pick exponentially distributed power-cycle durations is that they can be modeled easily in Markov chains.

2.3.3. Policies and metrics

I compare four different policies of the power manager: (1) a greedy policy GP, which deactivates the server as soon as no jobs are in the system and activates it as soon as a new job arrives. (2) The "accumulate and fire" policy AP(k) [CX07], which behaves like greedy but waits for k jobs to activate the server again. (3) An energy-minimizing policy EP with an oracle, which I use as a reference for energy consumption. (4) A latency-minimizing policy LP with an oracle, which I define to use the minimal amount of energy to achieve the same latency as an always-on policy OnP. I allow the energy- and latency-minimizing policies to use an oracle, which has knowledge of all future values of all random variables, that is inter-arrival time, processing time, activation time, and deactivation time. I show in section 2.4.3 that the energy-minimizing policy does not need the oracle. This makes the latency-minimizing policy the only policy that uses the oracle.

I consider different metrics to compare a policy X to others: (1) power consumption $\mathrm{P}(X)$ of the server, averaged over time, (2) latency $\mathrm{L}(X)$ as the mean time a job is in the system (waiting time plus processing time), (3) the ratio of latency compared to the always-on policy and, (4) the ratio of power consumption compared to the energy-minimizing policy. For both ratios I calculate the worst case compared to an offline algorithm (competitive ratio [BEY05] $\mathrm{CR}(X)$) and the mean under Poisson arrivals (Poisson ratio $\mathrm{PR}(X)$). I denote the ratios for latency with an index L ($\mathrm{CR}_\mathrm{L}(X)$ and $\mathrm{PR}_\mathrm{L}(X)$) and the ratios for power consumption with an index P ($\mathrm{CR}_\mathrm{P}(X)$ and $\mathrm{PR}_\mathrm{P}(X)$).

To calculate the power consumption of the system, I assume the server is the main consumer of power and ignore all other components. To further simplify the analysis I assume the server consumes 1 unit of power while active, activating, and deactivating, and zero power while sleeping. If these assumptions do not apply, the results can be scaled accordingly.

2.4. Poisson arrivals

In this section, I analyze the policies under the assumption that all variables (namely, $\mathcal{A}, \mathcal{L}, \mathcal{M}$, and $\mathcal{Z}$) are exponentially distributed. This allows the construction of a continuous-time Markov chain [Bol98] to calculate the mean power consumption and mean latency. Chen et al. [CX07] provide a similar analysis for the power consumption. I summarize their results and explain how I derive the mean latency. I used Maxima 5.25.1 to do the calculations I present in this chapter.

2.4.1. Greedy policy

Chen et al. [CX07] model the greedy policy as a Markov chain as shown in figure 2.3. The states in the chain are labeled as (State of server, Number of jobs in the system). I denote the probability to be in state X in the stationary distribution as $\mathbb{P}_\mathrm{st}[X]$.

Using the same steps as Chen et al. [CX07] I express the probability that the server is consuming power P using the greedy policy GP as

$$\mathrm{P}(\mathrm{GP}) = 1 - \mathbb{P}_\mathrm{st}[\mathrm{S_S}, 0] = \frac{\lambda\left(\mu\left(\omega+\alpha\right)\lambda + \omega\left(\mu\omega + \alpha\omega + \alpha\mu\right)\right)}{\mu\left(\left(\omega+\alpha\right)\lambda\left(\lambda+\omega\right) + \alpha\omega^2\right)}. \tag{2.1}$$

As I assume that the server consumes no power when in sleep mode and 1 unit of power else, this is also the mean power consumption of the greedy policy over time.

I calculate the mean number of users in the system by

$$\mathbb{E}[\text{Users in System}] = \sum_{i=1}^{\infty} i\left(\mathbb{P}_\mathrm{st}[\mathrm{S_A}, i] + \mathbb{P}_\mathrm{st}[\mathrm{S_D}, i] + \mathbb{P}_\mathrm{st}[\mathrm{S_U}, i]\right). \tag{2.2}$$

With this result and Little's law [Nel95] I calculate the average latency of a job to be

$$\mathrm{L}(\mathrm{G}) = \frac{\mathrm{U}_1\frac{\lambda^3}{\omega} + \left(\frac{\omega}{\alpha} - \frac{\mu}{\omega}\mathrm{U}_1\right)\lambda^2 - \left(\mu\mathrm{U}_1 + \alpha\right)\lambda - \left(\mu+\alpha\right)\omega}{\left(\omega+\alpha\right)\frac{\lambda^3}{\omega} + \left(\omega - \frac{\alpha\mu}{\omega} - \mu + \alpha\right)\lambda^2 + \mathrm{U}_2\lambda - \alpha\mu\omega}, \tag{2.3}$$

where $\mathrm{U}_1 = \frac{\omega}{\alpha} + \frac{\alpha}{\omega} + 1$ and $\mathrm{U}_2 = \alpha\omega - \mu\omega - \alpha\mu$.

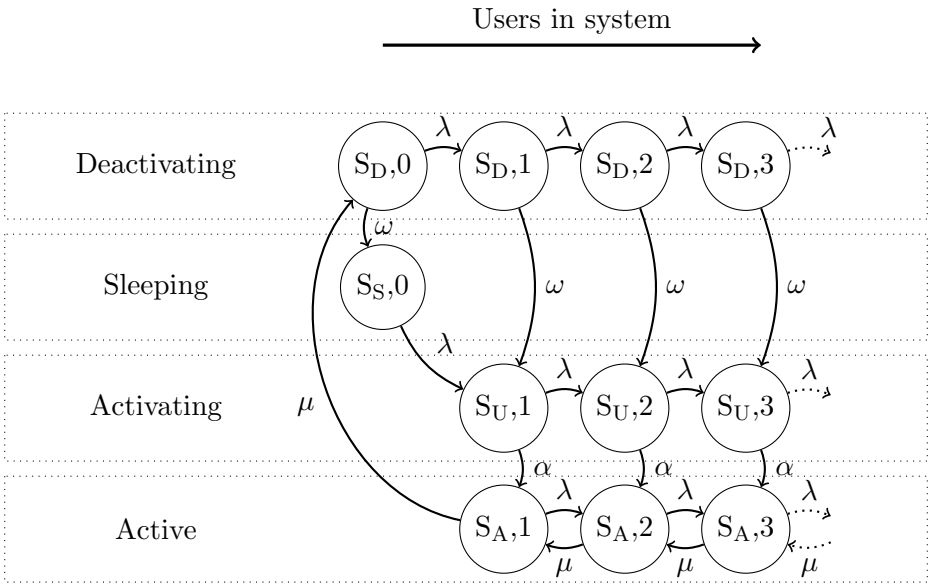

Figure 2.3.: A Markov model for the behavior of the greedy policy [CX07] allows both latency and power consumption to be determined.

2.4.2. Accumulate & fire policy

A generalization of the greedy policy is the accumulate and fire policy. It deactivates the server as soon as no jobs are left in the system and activates it again when the queue contains k jobs [CX07]. Figure 2.4 shows the Markov chain of the accumulate and fire policy. The probability to consume power and thus also the mean power consumption using accumulate and fire using my model is [CX07]

$$\mathrm{P}(\mathrm{AP}(k)) = \frac{\alpha\lambda^{k+2} - \alpha\lambda(\lambda+\omega)^k\left(\lambda - \left(\frac{\mu}{\alpha}+k\right)\omega - \mu\right)}{\mu\omega\left(\lambda+\alpha k\right)(\lambda+\omega)^k + \alpha\mu\lambda^{k+1}}. \tag{2.4}$$

Using the same method to calculate the mean number of jobs in the system as used for the greedy policy and Little's law [Nel95] I calculate the mean latency to be

$$\mathrm{L}(\mathrm{AP}(k)) = \frac{(\lambda+\omega)^k\omega^2\mathrm{V}_1 + 2\lambda^k\alpha\lambda\mathrm{V}_2}{2\alpha\omega\lambda\left(\lambda-\mu\right)\left(\left(\lambda+\alpha k\right)\omega(\lambda+\omega)^k + \alpha\lambda^{k+1}\right)}, \tag{2.5}$$

where $\mathrm{V}_1 = 2\lambda^3 + 2\left(\alpha k - \mu - \alpha\right)\lambda^2 + \alpha k\left(\alpha k - 2\mu - 3\alpha\right)\lambda + \alpha^2\left(1-k\right)k\mu$ and $\mathrm{V}_2 = \left(\omega+\alpha\right)\lambda^2 + \left(\left(-\mu+\alpha k-\alpha\right)\omega - \alpha\mu\right)\lambda - \alpha k\mu\omega$.

2.4.3. Energy-minimizing policy

To minimize the energy consumption the server needs to spend as little time doing state changes and running idle as possible. This can be achieved by letting the parameter

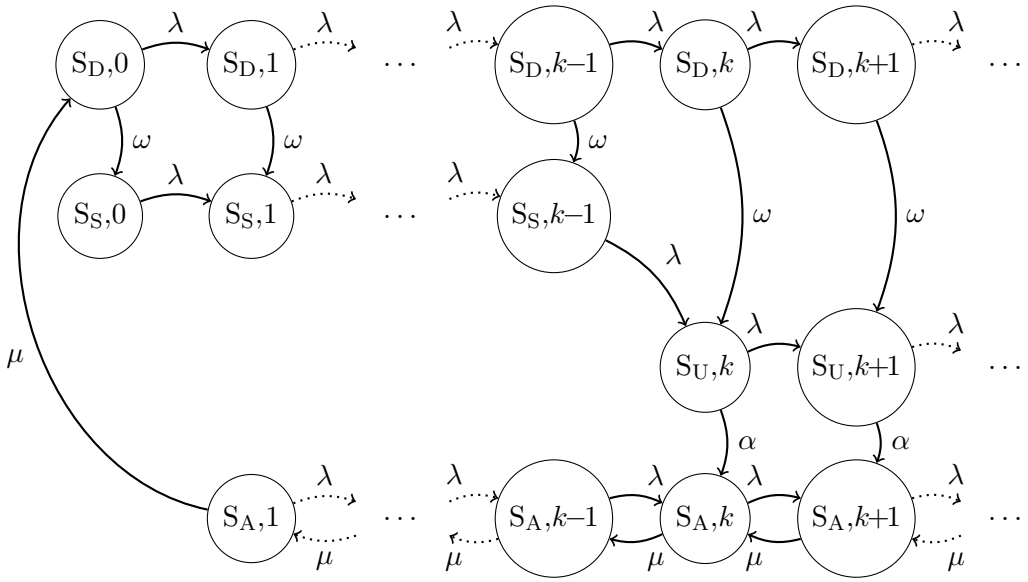

Figure 2.4.: A Markov model for the behavior of accumulate and fire policy [CX07] allows both latency and power consumption to be determined.

k of the accumulate and fire policy go towards infinity. Intuitively the power manager waits for a lot of jobs to be scheduled before activating the server up to process them, amortizing the power-cycle cost over the jobs that are processed at once. Note that this policy does not need an oracle because all the information to implement it is already available without an oracle.

As every policy needs to spend at least $\lambda\overline{\mathcal{M}}$ energy per time unit on average just to process the jobs, no policy can consume less than $\lambda\overline{\mathcal{M}} = \lambda/\mu$ power on average and still process all jobs. Because the mean power consumption using the accumulate and fire policy when k tends towards infinity becomes exactly λ/μ, this limiting case is the energy-minimizing policy. With increasing k, however, the mean latency also tends towards infinity.

As the latency is unbounded, the factor by which the latency is higher than the minimum possible latency is also unbounded. In contrast, to this the power consumption of any policy can be at most 1 (the maximum power consumed by the server). Because the most energy-efficient policy consumes λ/μ power, the ratio for power consumption for any policy cannot be higher than μ/λ.

2.4.4. Latency-minimizing policy

I analyze the latency-minimizing policy under the assumption that it can access an oracle that knows all future values of all random variables and not only their distributions. First, I assume all random variables are exponentially distributed and their realizations

are known to the oracle. Then, I apply the same calculation to constant power-cycle durations. In the second case, the oracle is still necessary to predict the inter-arrivals times.

The minimum latency is achieved by always keeping the server active. Hence, the latency of the latency-minimizing policy LP is the same as of the always-on policy OnP. It is [Bol98]

$$\mathrm{L(LP)} = \mathrm{L(OnP)} = \frac{1/\mu}{1-(\lambda/\mu)} = \frac{1}{\mu-\lambda}. \tag{2.6}$$

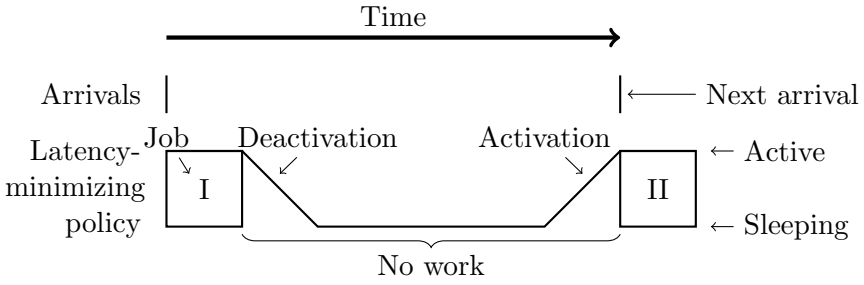

Figure 2.5.: The latency-minimizing policy can only deactivate the server during idle periods that are long enough. It can only do this reliably because it has access to an oracle for the knowledge of inter-arrival and power-cycle durations.

To reach the same latency, the latency-minimizing policy must finish each job at the same time and, thus, also start processing each job at the same time as the always-on policy would. The only time when the latency-minimizing policy can save energy is during idle periods that are long enough: when the server is idle and the time for a power cycle is smaller than the time until the next arrival, it can be deactivated. Figure 2.5 illustrates this.

I calculate the probability that the latency-minimizing policy can deactivate the server between two arrivals from the probability that the server is idle after processing the job and the probability that there is enough time for a power cycle. The probability that a job leaves the system empty is the same as the probability that a job finds an empty system because the M/M/1 is reversible [Nel95]. And because "Poisson arrivals see time averages" (PASTA) [Nel95], this is the same as the time average $\mathbb{P}_{\mathrm{st}}[0] = 1-\lambda/\mu$. Hence, the probability that a deactivation is possible after the server has processed a job is

$$\begin{aligned}\mathbb{P}[\text{Deactivation possible}] &= \overbrace{\mathbb{P}_{\mathrm{st}}[0]}^{\text{Job leaves system empty}}\mathbb{P}[\mathcal{A}+\mathcal{Z}<\mathcal{L}] \\ &= \mathbb{P}_{\mathrm{st}}[0]\underbrace{\mathbb{P}[\mathcal{Z}<\mathcal{L}]}_{\text{Enough time to deactivate}}\underbrace{\mathbb{P}[\mathcal{A}+\mathcal{Z}<\mathcal{L}|\mathcal{Z}<\mathcal{L}]}_{\text{Enough time to activate again}}.\end{aligned} \tag{2.7}$$

I derived the third part of the equation with law of the total probability and the knowledge that $\mathbb{P}[\mathcal{A}+\mathcal{Z}<\mathcal{L}|\mathcal{Z}>\mathcal{L}] = \mathbb{P}[\mathcal{A}<0] = 0$. When $\mathcal{L}$ is exponentially

distributed and thus memoryless, I can simplify this to

$$\mathbb{P}[\text{Deactivation possible}] = \mathbb{P}_{\text{st}}[0]\mathbb{P}[\mathcal{Z} < \mathcal{L}]\mathbb{P}[\mathcal{A} < \mathcal{L}]. \tag{2.8}$$

With probability $\mathbb{P}[\text{Deactivation possible}]$ the server can be deactivated for a time of $(\mathcal{L} - \mathcal{A} - \mathcal{Z} | \mathcal{A} + \mathcal{Z} < \mathcal{L}) = \mathcal{L}$. Because the average time the server spends in sleep mode between two arrivals is $\mathbb{P}[\text{Deactivation possible}]\overline{\mathcal{L}}$ and the mean time between two arrivals is $\overline{\mathcal{L}}$, $\mathbb{P}[\text{Deactivation possible}]$ is also the fraction of time spent in sleep mode in the steady state distribution with the latency-minimizing policy. Given the binary energy consumption model, this results in a mean power consumption of

$$\mathrm{P(LP)} = \frac{\lambda + \omega + \alpha + \frac{\alpha\omega}{\mu}}{\lambda + \omega + \alpha + \frac{\alpha\omega}{\lambda}}, \tag{2.9}$$

when all random variables are exponentially distributed.

To calculate $\mathbb{P}[\text{Deactivation possible}]$ I only needed to assume that $\mathcal{A}$ or $\mathcal{Z}$ are exponentially distributed for the last step. Using, as example, constant values for $\mathcal{A}$ and $\mathcal{Z}$, the mean power consumption simplifies to

$$\mathrm{P_C(LP)} = 1 - \overbrace{\left(1 - \frac{\lambda}{\mu}\right)}^{\text{Time not processing jobs}} \underbrace{\mathrm{e}^{-(\mathcal{Z}+\mathcal{A})\lambda}}_{\text{Fraction of that time spent in sleep mode}} \tag{2.10}$$

A similar analysis can be made for constant service times: The probability that a job leaves the system behind empty is given by the queue-length distribution of an M/D/1 [Nel95].

2.4.5. Comparison

In this section, I first show that both the greedy policy and the latency-minimizing policy can consume less energy than the other depending on the arrival pattern. Then I show that the latency-minimizing policy consumes less energy than the greedy policy, in the long run, when all random variables (inter-arrival time, processing time, power-cycle durations) are exponentially distributed.

Figure 2.6 shows an arrival pattern that causes the latency-minimizing policy to consume more energy than the greedy policy as it does not deactivate the server between the arrivals. The greedy policy will process the first job later and the second right after the first one. In contrast to this, Figure 2.7 shows an arrival pattern in which the future knowledge of the latency-minimizing policy prevents it from deactivating the server because this would delay the second job. The idle time of length ϵ is necessary to give the greedy policy time to initiate the deactivation. The latency-minimizing policy not only reduces latency, but also conserves energy in this case because it deactivates the server for a longer period later.

Given these two examples, it is clear that both policies can be more energy-efficient than the other for given examples. Next I determine which of these effects outweighs the other when arrivals are given by a Poisson process.

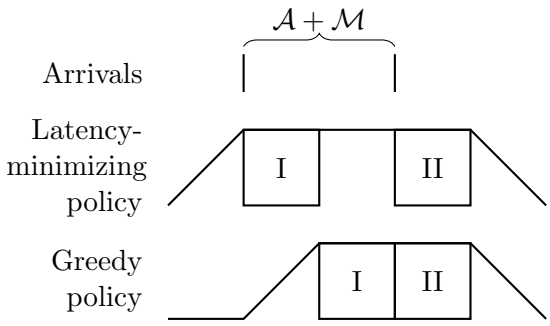

Figure 2.6.: This is an arrival pattern in which the greedy policy consumes less energy than the latency-minimizing policy.

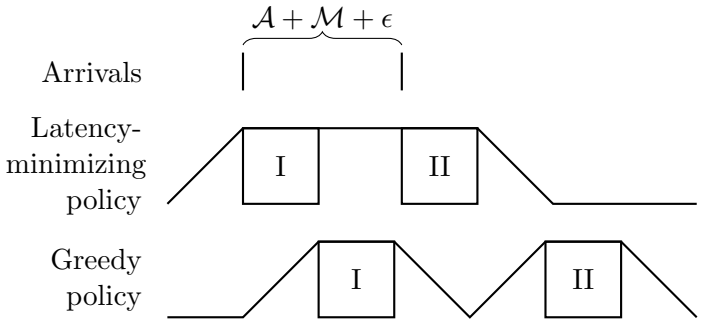

Figure 2.7.: This is an arrival pattern in which the latency-minimizing policy consumes less energy than the greedy policy.

I already calculated the power consumption of the greedy policy P(GP) and the power consumption of the latency-minimizing policy P(LP). Subtracting the power consumption of the latency-minimizing policy from the greedy policy and simplifying yields a strictly positive result (with the reasonable assumptions $\alpha, \omega > 0$ and $\mu > \lambda > 0$). Hence, the greedy policy consumes more power on average than the latency-minimizing policy when the arrivals are a Poisson process.

Given the mean values for the latency of the greedy policy and the always-on policy, I calculate the factor $\mathrm{PR_L(GP)} = \mathrm{L(GP)/L(OnP)}$ by which greedy policy's latency is higher than that of the always-on policy

$$\mathrm{PR_L(GP)} = 1 + (\overline{\mathcal{A}} + \overline{\mathcal{Z}})(\mu - \lambda) + R, \tag{2.11}$$

where

$$R = \frac{\overline{\mathcal{A}\mathcal{Z}}^2\lambda^3 - \mu\overline{\mathcal{A}\mathcal{Z}}^2\lambda^2 + \overline{\mathcal{A}\mathcal{Z}}\lambda^2 - \mu\overline{\mathcal{A}\mathcal{Z}}\lambda + \overline{\mathcal{Z}}\lambda - \mu\overline{\mathcal{Z}}}{\overline{\mathcal{Z}}^2\lambda^2 + \overline{\mathcal{A}\mathcal{Z}}\lambda^2 + \overline{\mathcal{Z}}\lambda + \overline{\mathcal{A}}\lambda + 1} < 0 \tag{2.12}$$

is always negative, and thus $\mathrm{PR_L(GP)}$ grows at most linear in $\overline{\mathcal{A}}$ and $\overline{\mathcal{Z}}$. Analogously the factor for mean power consumption of the greedy policy $\mathrm{PR_P(GP)} = \mathrm{P(GP)/P(EP)}$ is

$$\mathrm{PR_P(GP)} = 1 + \frac{\mu - \lambda}{\lambda} W, \tag{2.13}$$

where

$$W = 1 - \frac{1}{\left(\overline{\mathcal{A}} + \overline{\mathcal{Z}}\right)\lambda\left(\overline{\mathcal{Z}}\lambda + 1\right) + 1} \tag{2.14}$$

takes values between 0 and 1 and describes how much of the power that the energy-minimizing policy saves is conserved by the greedy policy.

2.5. Adversary-controlled arrivals

In this section, I compare the *mean* latency and the *mean* power consumption of the greedy policy to the respective optima. I do this for an adversary who has control over the inter-arrival times of new jobs. Letting an adversary try to maximize latency or power consumption by selecting values for the variables is the same as calculating the worst case for a random variable, because the adversary can select the worst case.

I begin by considering both the job processing time and the (de-)activation durations to be constant values, because this is easiest to analyze. Later, I generalize the result to arbitrary distributions.

2.5.1. Latency ratio: greedy policy

I show that the competitive ratio of the greedy policy for latency is $\mathrm{CR_L(GP)} = 1 + \frac{\mathcal{A}+\mathcal{Z}}{\mathcal{M}}$. To do this I provide an example of job inter-arrival times in which the mean latency of the greedy policy is a factor of $1 + \frac{\mathcal{A}+\mathcal{Z}}{\mathcal{M}}$ higher than the latency of the always-on policy. Later I show that this is the maximum for all inter-arrival times.

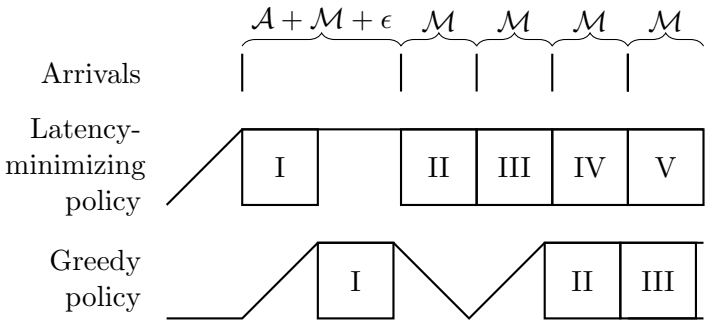

Figure 2.8.: This is the arrival pattern with the worst competitive ratio for latency $\mathrm{CR_L}$ for the greedy policy.

The basic idea for the example is to create the highest possible latency for the greedy policy and never let it catch up with the always-on policy again. Figure 2.8 illustrates the idea. It is achieved by creating an idle time gap of size ϵ between the end of the first job and the arrival of the second job. While the server of the always-on policy will stay active, the greedy policy will undergo a full power cycle and finish processing the first job $\mathcal{A}+\mathcal{Z}-\epsilon$ later than the always-on policy. For the rest of the (infinitely long) example, the arrival times of new jobs equal the finish times of the last job of the always-on policy. Hence, every following job will be finished $\mathcal{A}+\mathcal{Z}-\epsilon$ later using the greedy policy than the always-on policy.

Using this allocation of job inter-arrival times, the always-on policy will finish processing each job $\mathcal{M}$ time units after it arrived, while the greedy policy will finish all (infinitely many) but the first job after $\mathcal{A}+\mathcal{Z}-\epsilon+\mathcal{M}$ time. Or, equivalently, a factor of $1+\frac{\mathcal{A}+\mathcal{Z}}{\mathcal{M}}$ later than the always-on policy for ϵ going to zero.

To show that this is the worst the greedy policy can behave, I show that the greedy policy can never be more than $\mathcal{A}+\mathcal{Z}$ late when finishing a job. To do this I show that the invariant "the arriving job is finished at most $\mathcal{A}+\mathcal{Z}$ later than using the always-on policy" holds for all job arrivals during any sequence of inter-arrival times using mathematical induction.

First I show that the invariant holds at the first arrival time t_0: The always-on policy will immediately start processing the job if it arrives at t_0, while the greedy policy will initiate a state change at t_0 and finish at $t_0+\mathcal{A}$. Hence, the job is finished at time $t_0+\mathcal{M}$ using the always-on policy and $t_0+\mathcal{A}+\mathcal{M}$ using the greedy policy.

Now I show that the invariant also holds for all subsequent arrivals. I make a case distinction over how many jobs are in the system of the always-on policy when the new job arrives, called n. For $n>0$, both policies will start processing the new job right after the last job finishes. Hence, the new job will have the same delay as the previous, which by the invariant is at most $\mathcal{A}+\mathcal{Z}$.

When a new job arrives and no other jobs are in the always-on system, I define the

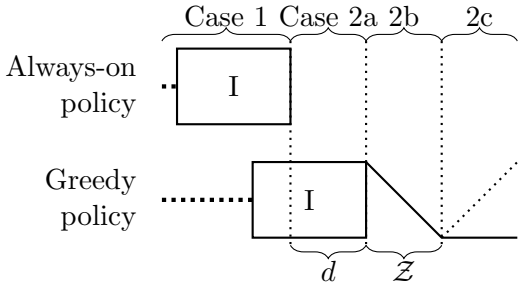

Figure 2.9.: The different possibilities for arriving jobs need to be considered individually to analyze the latency of the greedy policy GP.

time between the previous job finished in the always-on system and in the greedy system as d (note that time d may or may not have passed when the new job arrives). Figure 2.9 illustrates the different possible time windows in which a new job can arrive. If d has not passed, the new job will start processing immediately using the always-on policy and after d has passed using the greedy policy. Hence, this new job will be finished at most d later than using the always-on policy, because $d < \mathcal{A} + \mathcal{Z}$ holds by the invariant.

If the new job arrives after d has passed, the greedy policy will power cycle the server and the delay will not be larger than $\mathcal{A} + \mathcal{Z}$; using the always-on policy processing starts as soon as the new job arrives.

This shows that no job will be finished more that $\mathcal{A} + \mathcal{Z}$ later using the greedy policy than the same job using the always-on policy. Because each job has a running time of $\mathcal{M}$, each job is in the system $\mathcal{M}$ time units compared to $\mathcal{A} + \mathcal{Z} + \mathcal{M}$, which is equivalent to a maximum increase in latency by a factor of $1 + \frac{\mathcal{A}+\mathcal{Z}}{\mathcal{M}}$. Hence, the competitive ratio for latency of the greedy policy is $\mathrm{CR_L(GP)} = 1 + \frac{\mathcal{A}+\mathcal{Z}}{\mathcal{M}}$.

2.5.2. Latency ratio: accumulate & fire policy

As the greedy policy is equal to the accumulate and fire policy with $k = 1$, the competitive ratio for latency $\mathrm{CR_L(AP}(k))$ is $1 + \frac{\mathcal{A}+\mathcal{Z}}{\mathcal{M}}$ for $k = 1$. I now show that for all larger k, the competitive ratio for latency of the accumulate and fire policy is infinite by providing an arrival pattern that can create arbitrarily large latency ratios. The idea is to let the always-on policy handle the first job without delay while the accumulate and fire policy waits for the second job to activate the server. The time n between arrivals can be selected arbitrarily large.

The i-th job arrives at time $i \cdot n$. The accumulate and fire policy will only activate the server when k jobs have arrived, to process them in a batch. Hence, when n goes towards infinity the mean latency of a job goes towards infinity.

2.5.3. Energy ratio: greedy policy

Again, first I give a series of job inter-arrival times that result in the worst case energy consumption for the greedy policy; then I prove that this is the worst case.

The basic idea for the scenario is to let the greedy policy power cycle for each job while the energy-minimizing policy waits until a large batch of jobs has arrived and processes them in one cycle.

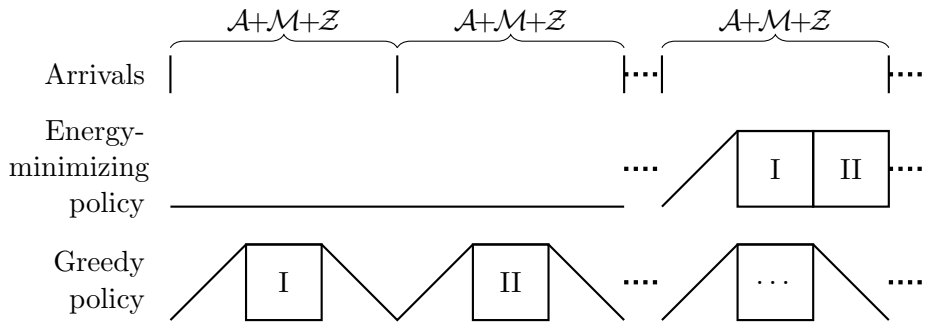

Figure 2.10.: This is the arrival with the worst competitive ratio for power $\mathrm{CR_P}$ for the greedy policy.

The job arrivals are at $i(\mathcal{A}+\mathcal{M}+\mathcal{Z})$, with $i \in \{0,\ldots,n\}$. The greedy policy will power cycle between each arrival and thus consume $n(\mathcal{A}+\mathcal{M}+\mathcal{Z})$ units of energy. This is illustrated in figure 2.10. The energy-minimizing policy will activate once when all jobs have arrived, process all jobs, and deactivated after that. Hence, it will consume $\mathcal{A}+n\mathcal{M}+\mathcal{Z}$ units of energy. When n approaches infinity, the ratio between both energy consumptions approaches $1+\frac{\mathcal{A}+\mathcal{Z}}{\mathcal{M}}$.

In the worst case, the greedy policy will do a full power cycle for each job. Hence, consuming $\mathcal{A}+\mathcal{M}+\mathcal{Z}$ units of energy for each job. To process a job, the server must be active for at least $\mathcal{M}$ units of time, and thus consumes at least $\mathcal{M}$ units of energy. The ratio between the two energy consumptions shows that the competitive ratio for power $\mathrm{CR_P(GP)}$ consumption of the greedy policy is $1+\frac{\mathcal{A}+\mathcal{Z}}{\mathcal{M}}$. Note that the competitive ratio for power is the same for the latency-minimizing policy because it also activates the server for every single arriving job.

2.5.4. Energy ratio: accumulate & fire policy

First, I describe an arrival pattern that results in this ratio and, second, I show that no pattern can be worse. Both arguments closely follow the same idea as for the greedy policy, which figure 2.8 shows; the only difference is that the accumulate and fire policy processes k jobs in each cycle.

The job arrivals are at $i(\mathcal{A}+k\mathcal{M}+\mathcal{Z})$, with $i \in \{0,\ldots,kn\}$. The accumulate and fire policy will power cycle between every k arrivals and thus consume $n(\mathcal{A}+k\mathcal{M}+\mathcal{Z})$ units

of energy. The energy-minimizing policy will activate once when all jobs have arrived, process all jobs, and deactivated after. Hence, it will consume $\mathcal{A} + nk\mathcal{M} + \mathcal{Z}$ units of energy. When n approaches infinity, the ratio between both energy consumptions approaches $1 + \frac{\mathcal{A}+\mathcal{Z}}{k\mathcal{M}}$.

As the accumulate and fire policy only initiates state changes when jobs arrive or finish, it will do a full power cycle every k jobs. Hence, it consumes $\mathcal{A} + k\mathcal{M} + \mathcal{Z}$ units of energy for k jobs. To process k jobs, the server must be active for at least $k\mathcal{M}$ units of time, and thus consumes at least $k\mathcal{M}$ units of energy. The ratio between the two energy consumptions shows that the competitive ratio for power consumption of the Accumulate and Fire policy $\mathrm{CR_P}(\mathrm{AP}(k))$ is $1 + \frac{\mathcal{A}+\mathcal{Z}}{k\mathcal{M}}$.

Note that the competitive ratio for power $\mathrm{CR_L}(\mathrm{AP}(k))$ for the accumulate and fire policy approaches 1 when k approaches infinity. The reason for this is that the accumulate and fire policy becomes the energy-minimizing policy when k approaches infinity.

2.5.5. Impossible trade-offs between energy and latency

In this section, I prove two theorems that give lower bounds on the competitive ratios for algorithms without an oracle.

Theorem 2.1. *No single policy (without an oracle) can have a competitive ratio for latency CR_L strictly lower than $1 + \frac{\mathcal{A}+\mathcal{Z}}{\mathcal{M}}$ and a finite competitive ratio for power CR_P.*

Proof. To reach a finite competitive ratio for power, every policy has to deactivate the server at some point in time. The same example from the calculation of the competitive ratio $\mathrm{CR_L}$ of the greedy policy (figure 2.8) can now be used again: after the deactivation has been initiated the adversary can then make jobs arrive at the system. When the server is active again it has a delay of $\mathcal{A} + \mathcal{Z} - \epsilon$ compared to the always-on policy. From then on the adversary can keep the server utilized all the time, thus never letting any policy catch up with the always-on policy. Hence, all jobs will finish after $\mathcal{A} + \mathcal{Z} - \epsilon + \mathcal{M}$, while the always-on policy finishes them after $\mathcal{M}$ time units. For $\epsilon \to 0$, this results in a competitive ratio for latency of $\mathrm{CR_L} = 1 + \frac{\mathcal{A}+\mathcal{Z}}{\mathcal{M}}$. □

Theorem 2.2. *No single policy (without an oracle) can have a finite competitive ratio for latency CR_L and a competitive ratio for power CR_P strictly lower than $1 + \frac{\mathcal{A}+\mathcal{Z}}{\mathcal{M}}$.*

The idea of the proof is that in order to have a competitive ratio $\mathrm{CR_P}$ for energy lower than $1 + \frac{\mathcal{A}+\mathcal{Z}}{\mathcal{M}}$, the policy must process more than one job during one power cycle on average. However, because jobs can be arbitrarily far apart, this will result in an arbitrarily large mean latency.

Proof. An arrival pattern for which no policy can perform under both boundaries is the example already provided to show that any accumulate and fire policy with $k > 1$ has infinite competitive ratio for latency: The job i arrives at time in, where n approaches infinity. A policy that executes each job in a single power cycle will consume $\mathcal{A} + \mathcal{M} + \mathcal{Z}$ energy for each job and thus have a $\mathrm{CR_P}$ of $1 + \frac{\mathcal{A}+\mathcal{Z}}{\mathcal{M}}$. Thus, it violates the first boundary. Define the mean number of jobs executed in a power cycle as $c > 1$. The competitive

ratio for power is $1+\frac{\mathcal{A}+\mathcal{Z}}{c\mathcal{M}}$, which fulfills the first boundary. To process c jobs in a single power cycle, $c-1$ jobs must be delayed by at least n units of time. Hence, the mean latency of all jobs must be at least

$$\mathrm{L} \geq \overbrace{(0+\mathcal{M})\frac{1}{c}}^{\text{Non-delayed jobs}} + \overbrace{(n+\mathcal{M})\frac{c-1}{c}}^{\text{Delayed jobs}}. \tag{2.15}$$

For every fixed $c > 1$, this sum tends towards infinity, when n tends toward infinity. □

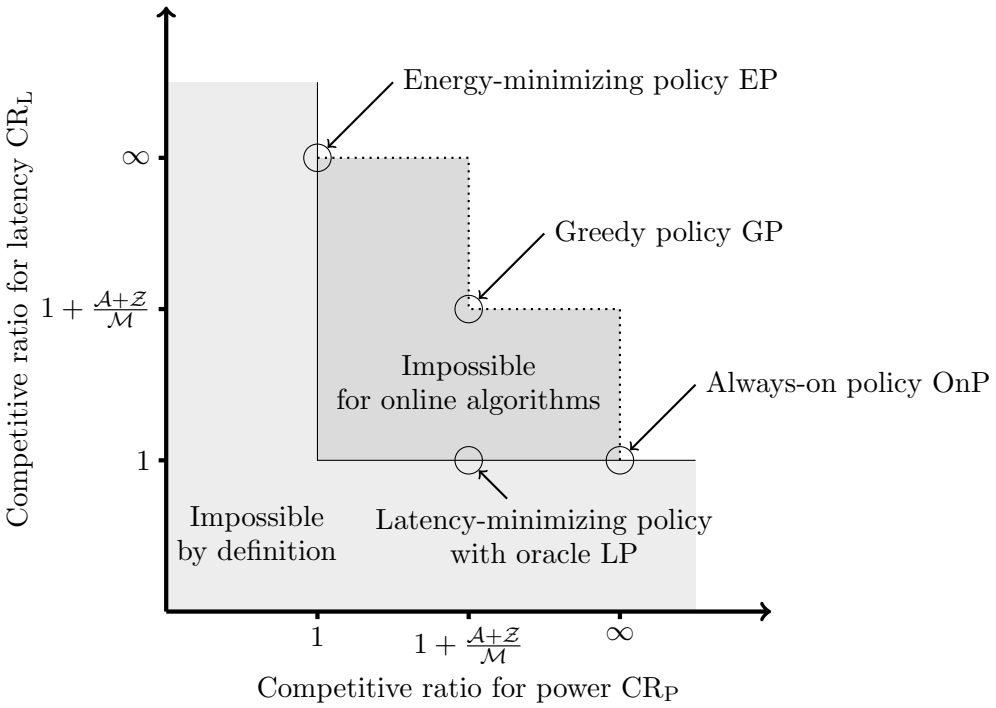

Figure 2.11.: This is an illustration of the competitive ratios of the different policies. Note that this is not to any scale.

Figure 2.11 gives an overview of the different competitive ratios and which combinations of ratios are not possible for online algorithms. The figure is only meant as a visual reminder of the possible trade-offs of the different policies; it is not to any scale. The results only hold for deterministic algorithms. An analysis for randomized algorithms that is based on my work can be found in Nico Bredenbals' masters's thesis [Bre13].

2.5.6. Arbitrarily distributed random variables

If the adversary has control over the inter-arrival times of new jobs and all other values $(\mathcal{A}, \mathcal{M}, \mathcal{Z})$ are general random variables or also under adversarial control, the competitive

ratios become

$$\mathrm{CR_P(GP)} = \frac{\mathbb{E}[\mathcal{A} + \mathcal{Z})]}{\mathbb{E}[\mathcal{M}]} \tag{2.16}$$

and

$$\mathrm{CR_L(GP)} = \frac{\max(\mathcal{A} + \mathcal{Z})}{\mathbb{E}[\mathcal{M}]}. \tag{2.17}$$

This holds for all distributions of $\mathcal{A}, \mathcal{M}$ and $\mathcal{Z}$. I assume the adversary knows all realizations of all random variables not under its control. If the adversary has control over $\mathcal{A}, \mathcal{M}$, or $\mathcal{Z}$, both the expected value $\mathbb{E}$ and the maximum value max can be calculated from the choices I allow the adversary. If the numerator is infinite or the denominator approaches zero, the corresponding competitive ratio is infinite.

As the proofs of these statements are similar to the proofs done in the earlier sections, I only outline the differences here: The maximum power cycle time determines the competitive ratio for latency and the mean power cycle time determines the competitive ratio for power. The reason for this is that a single long power cycle can result in high latency for all following jobs, but only consumes additional energy once.

To create the worst ratio for power consumption, the adversary provides the jobs so sparsely that the greedy policy will power cycle for each job, thus consuming $\overline{\mathcal{A}}+\overline{\mathcal{M}}+\overline{\mathcal{Z}}$ energy while the energy-minimizing policy consumes $\overline{\mathcal{M}}$ energy asymptotically.

The worst ratio for latency can be achieved by letting the greedy policy power cycle repeatedly. Once a power cycle of maximum (or any arbitrarily large) length starts, the rest of the arrivals can be scheduled just when the always-on policy finished the previous job. This results in a mean latency of the always-on policy of $\overline{\mathcal{M}}$, while the mean latency for the greedy policy is $\max(\mathcal{A} + \mathcal{Z}) + \overline{\mathcal{M}}$. It is possible to show this using the same distinction as in section 2.5.1.

The statements made in this section hold for all arrival patterns of infinite length. They are not true for arrival patterns of finite length, but the longer the arrival pattern, the more it approaches the values calculated in this section. The reason for this is that the adversary needs time and in some cases repeated tries to generate the worst case. When the time is limited the amount of tries to for example find a long power cycle is limited and, hence, he cannot create the worst case reliably.

2.6. Results

Unless noted otherwise, I arbitrarily selected the arrival rate $\lambda = 1/4$, the processing rate $\mu = 1$, the activation rate $\alpha = 2$, and the deactivation rate $\omega = 2$ for the plots shown in this section.

To gain confidence in the correctness of the analytic derivation of the energy consumption and the latency, I implemented a simple event-based simulation from the model. Figures 2.12 and 2.13 show that the results derived in the previous section closely match the results from the simulation. Hence, I conclude that the analytical results are correct. Additonally, both figures show that a higher activation threshold increaes latency, but the effect on power consumption is negligible.

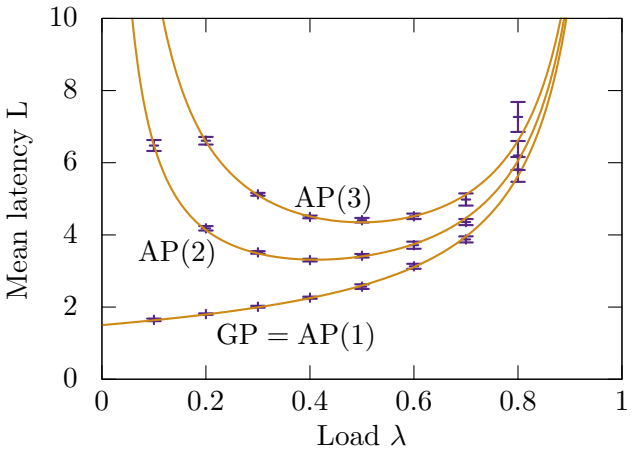

Figure 2.12.: The analytic results for the latency match my simulation. Note that the latency increases when the load approaches 0 for the accumulate and fire policy for $k > 1$ because often single jobs wait for others to arrive and activate the server.

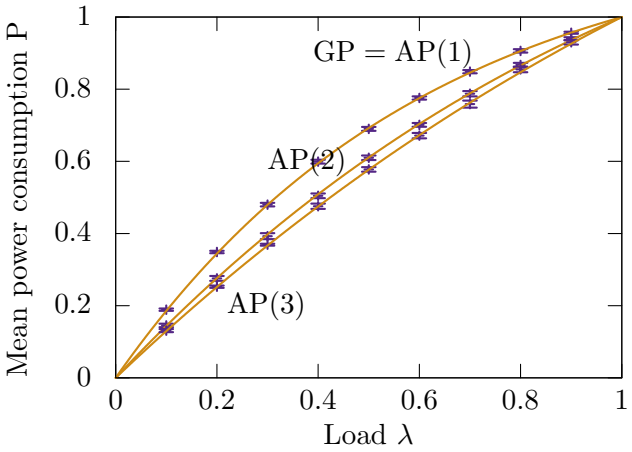

Figure 2.13.: The analytic results for the power consumption match my simulation.

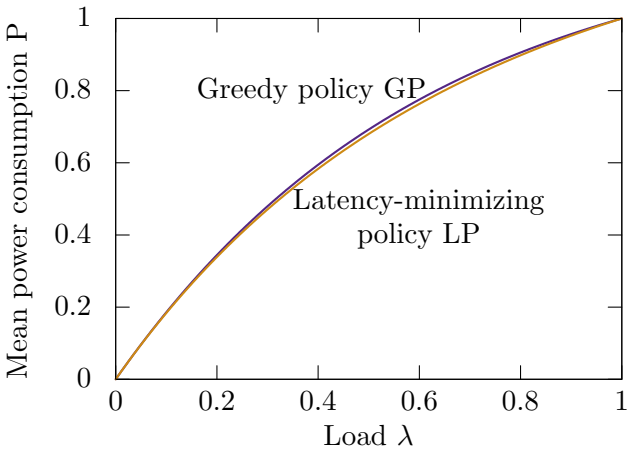

Figure 2.14.: The greedy policy conserves nearly as much power as the latency-minimizing policy.

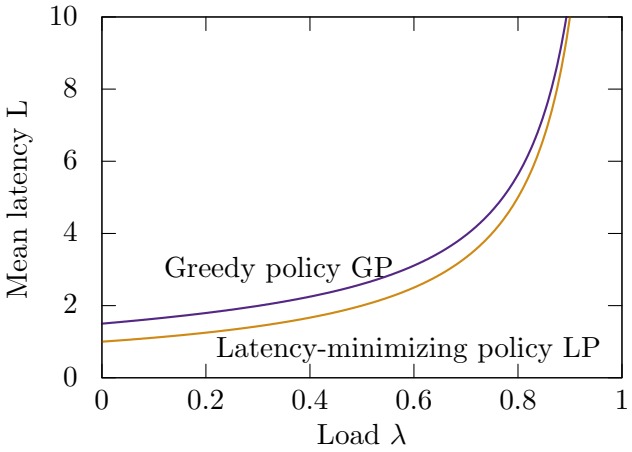

Figure 2.15.: The greedy policy has a higher latency than the latency-minimizing policy.

Figure 2.14 and 2.15 compare the latency and the power consumption of the greedy policy and the latency-minimizing policy. They show that the greedy policy conserves nearly as much power as the latency-minimizing policy, but pays for this with increased latency.

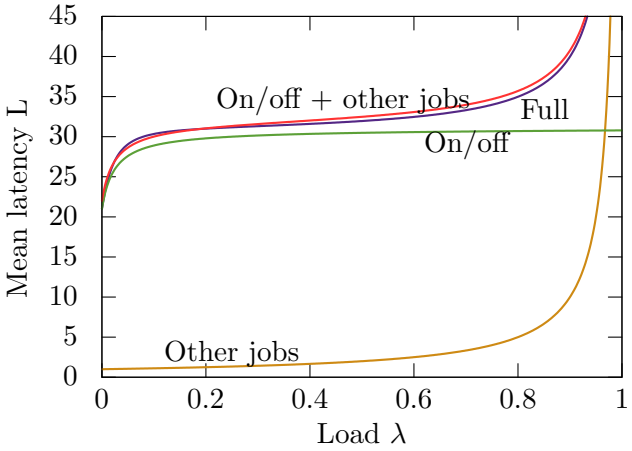

Figure 2.16.: The S-Curve of the full model is the sum of two different effects: Waiting for the server to become active, which dominates under low load, and waiting for the server to finish other jobs, which dominates at high load. I use $\overline{\mathcal{A}} = \overline{\mathcal{Z}} = 20$ to show both effects in this plot.

Figure 2.16 shows an S-curve that appears when the activation and deactivation times are higher ($\overline{\mathcal{A}} = \overline{\mathcal{Z}} = 20$ in this plot). The S-curve is the result of the overlap of two different reasons for latency: Waiting for the server to reach the active state and waiting for other jobs to finish processing. For this comparison I calculated the time a job waits on other jobs to finish from the always-on policy; and the time waiting for the server to become active from a Markov chain described in section 2.4 that only has one customer. Figure 2.16 shows that the sum of these two effects closely describes the effect seen in the full model.

Figure 2.17 illustrates the possible trade-offs that the greedy policy and the accumulate and fire policies allow. It shows how latency and power can be traded against each other and where the theoretical boundaries are.

Figures 2.18 and 2.19 show how higher activation and deactivation durations influence the power consumption and the latency of the greedy policy and the latency-minimizing policy. The power consumption of the greedy policy is influenced more by power cyclepower-cycle durations than the latency-minimizing policy. While the latency of the latency-minimizing policy is independent of activation and deactivation durations, the latency of the greedy policy increases approximately linearly (see section 2.4).

Figure 2.20 shows what happens if the time required for a power cycle is not equally distributed between activation and deactivation phases. The left side shows small ac-

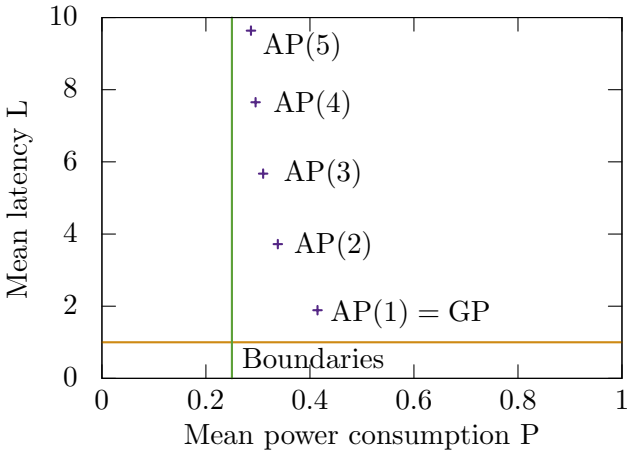

Figure 2.17.: The greedy policy and the accumulate and fire policies allow different trade-offs between latency and power consumption. The bounds are the theoretical lower bounds.

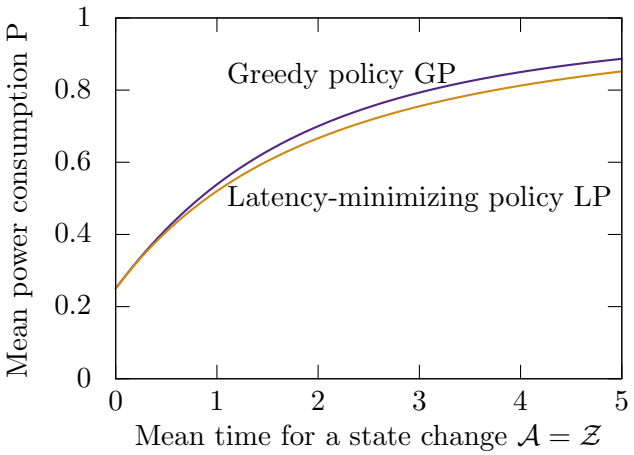

Figure 2.18.: The power consumption increases with increasing power state change times. The effect using the greedy policy is stronger than using the latency-minimizing policy.

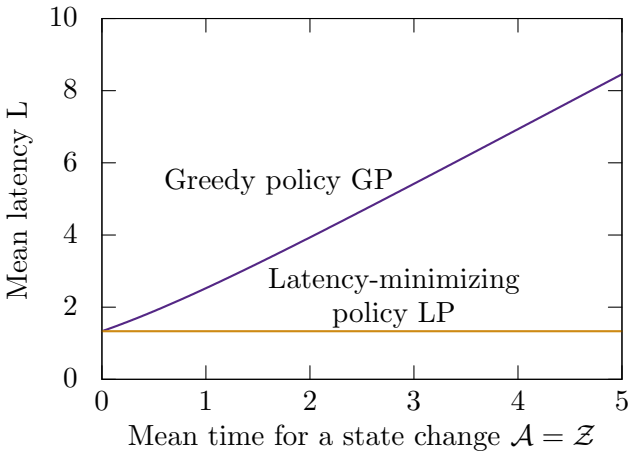

Figure 2.19.: The latency increases with increasing power state change times using the greedy policy, but not using the latency-minimizing policy.

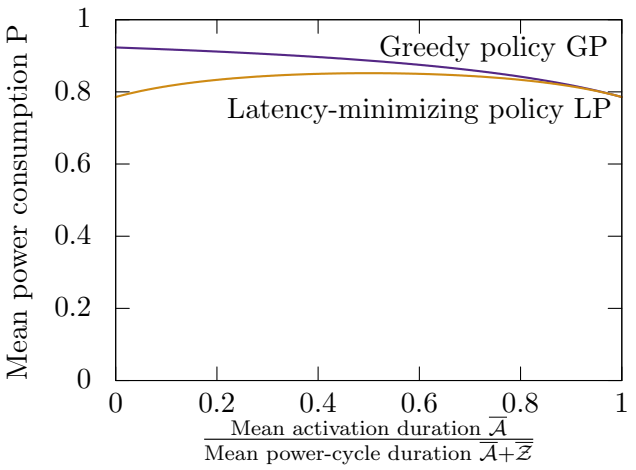

Figure 2.20.: While the power consumption with the latency-minimizing policy with an oracle is symmetric, it is asymmetric with the greedy policy. I fixed $\overline{\mathcal{A}}+\overline{\mathcal{Z}} = 10$ here to highlight the asymmetry.

tivation durations and high deactivation durations and the opposite is on the right. I normalized the state change durations so that the expected value of their sum is the same.

The power consumption of the latency-minimizing policy is symmetric with respect to the activation and deactivation durations: The latency-minimizing policy will produce the same result if the expected values for $\mathcal{A}$ and $\mathcal{Z}$ are exchanged. More generally, the power consumption depends only on the distribution of $\mathcal{A} + \mathcal{Z}$, but on neither one individually. For exponentially distributed $\mathcal{A}$ and $\mathcal{Z}$ this only holds if their expected values are switched; if the sum of their expected values is the same, the distribution will generally not be the same (formally: $X_i \sim \exp(\lambda_i) \Rightarrow (X_1 + X_2) \sim (\widetilde{X}_1 + \widetilde{X}_2)$ if and only if $\lambda_1 = \widetilde{\lambda}_1 \wedge \lambda_2 = \widetilde{\lambda}_2$ or $\lambda_1 = \widetilde{\lambda}_2 \wedge \lambda_2 = \widetilde{\lambda}_1$).

In contrast to the latency-minimizing policy the power consumption of the greedy policy is not symmetric: If the time of the power cycle is mostly spent in activation phases the power consumption is lowest. The reason for this is that long activation durations will create batches of jobs that will be processed at once. At the extreme of zero deactivation duration, the power consumption of the greedy policy is the same as of the latency-minimizing policy. However, creating batches of jobs introduces additional waiting time. This effect is shown in figure 2.21. If the activation is the dominant part of the power cycle, the example presented in figure 2.7 results in a higher power consumption of the greedy policy.

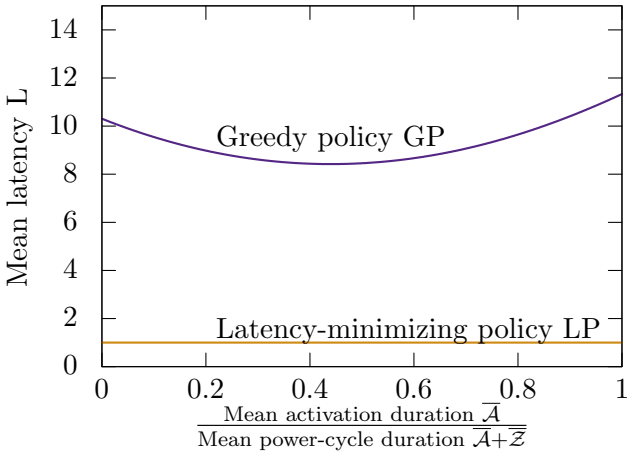

Figure 2.21.: While the mean latency of the latency-minimizing policy with an oracle power is independent of the power-cycle durations, the mean latency of the greedy policy changes. I fixed $\overline{\mathcal{A}}+\overline{\mathcal{Z}} = 10$ here to highlight the asymmetry.

Figure 2.21 shows how the mean latency using the greedy policy increases if the power-cycle durations are dominated by activation durations. It additionally shows that the mean latency does not have its minimum when the activation time is lowest: When the mean activation time is low, the mean deactivation time is high and the probability for a job to arrive during the deactivation of the previous job increases more than what the

lower activation time compensates for. The minimum mean latency is achieved when the mean activation time is

$$\overline{\mathcal{A}} = \frac{3S^2\lambda^2 + 2S\lambda + 3 - (S\lambda + 1)\sqrt{S^2\lambda^2 + 2S\lambda + 9}}{4S^2\lambda^2} S, \tag{2.18}$$

where $S = \overline{\mathcal{A}} + \overline{\mathcal{Z}}$ is the fixed expected time for a power cycle. Note that the mean latency of the latency-minimizing policy does not depend on the distribution between activation and deactivation times (not even their actual values) because it guarantees the same mean latency as the always-on policy.

2.7. Conclusion

In this chapter, I analyzed how the power-cycle durations of a single BS influence the mean power consumption and mean latency. I calculated the mean latency and power consumption both under Poisson and worst case arrivals. In the worst case, both the latency and the power consumption are linear in power-cycle durations. For Poisson arrivals, the power-cycle duration determines both to a large extent. However, the result for Poisson arrivals is asymmetric for other distributions of power-cycle duration to activation and deactivation durations. In summary, I quantified how a high ratio of power cycle to job processing time makes the sleep mode impractical.

Disabling a BS without coordinating with its neighbors is only useful, if (1) the jobs can tolerate a high amount of delay or (2) the power cycle of the BS is fast compared to processing a job. But neither of these are guaranteed to hold in future RANs. Hence, I need to look beyond a single BS.

The results of this chapter apply only when considering a single BS, but not when considering a complete RAN in its entirety. To show how a RAN can conserve more energy, I first show how a system composed of multiple smaller components can conserve energy without depending on fast power-cycle durations in the next chapter.

3. Conserving energy by coordinating sleep modes network-wide

The propagation of radio waves is complex due to effects such as path loss, fading, shadowing, and interference. Hence, it is hard to analyze the effects of energy-conservation methods in a radio access network (RAN). To show these effects in a simpler example, I consider a wired network in this chapter. I selected the example to show that the effects of energy-conservation methods on latency can be measured with different latency metrics. Also, this chapter shows how a system can adapt to traffic with sleep modes even if the power cycles are long. Long power cycles in this context means that the power cycles are so long that adapting to single demands of the users (as in the previous chapter) is impractical. In this chapter, I assume that power-cycle durations are short compared to changes of *average* load.

With this example I also show how different methods to conserve energy interact. I analyze the influence of deactivating network components on power consumption and latency of data transfers. Then, I compare the power consumption necessary to bound latency without depending on the power-cycle durations of the devices. This results in description of the trade-off between power consumption and latency. It also illustrate how a complex system can approach a linear power profile even though all sub-components have a binary power profile.

The previous chapter showed an example how demands can be moved in time. This chapter is an example how demands can be moved in space by rerouting. In both cases I measure the effect for the users as increased latency. In the previous chapter the increase in latency was due to queuing effects and in this chapter they are due to the increased distance the signals have to travel.

3.1. Introduction

In this chapter, I describe how energy can be conserved in a wired network not operating under full load. To reduce the energy consumption I allow traffic to be rerouted away from some connections and then deactivate the line cards of these connections. Rerouting traffic increases the latency of the rerouted traffic and, thus, reduces the users experience.

I formalize the problem of deactivating lines to conserve power while fulfilling all demands. Then I define network configurations that minimize power consumption and configurations that minimize latency as well as trade-offs in between. Unfortunately, the corresponding decision problem (capacitated fixed-charge multi-commodity minimum-cost flow network) is nondeterministic polynomial time (NP)-complete [JLR78]. Because it is not feasible to find the optimal solution for NP-complete problems, I use mixed

integer linear programming (MILP) to approximate the solutions for a given realistic network. The combinatorial explosion of possible states that makes this NP-complete is typical for energy conservation methods in networks and will be important in the later analysis of the energy consumption of RANs.

To analyze the trade-off between power consumption and latency, I begin by formalizing the problem as a graph. Then, I prove a bound between different possible metrics of latency and explain the relationship between latency and bandwidth-delay product (BDP). After that I create an optimization model from the formalization of the problem and show the results I obtain with it. My results show that it is possible to conserve considerable power by deactivating line cards in wired networks while increasing the latency only slightly.

3.2. Related work

Many papers present work on minimizing the cost in fixed-charge networks [HS89, KF91], but not many of them consider the effects of deactivation on latency. One example which does is the work of Lin et al. [LS12]. While they focus on developing efficient heuristics I use an optimization model and compare the different ways to measure the increased latency of the demands.

Chiaraviglio et al. [CMN09a, CMN09b] describe several centralized heuristics to approximate the most power-efficient topology for the binary power-consumption model. They assume that line cards as well as routers can be deactivated and demands of end-to-end flows are known. The heuristics provide results that are close to minimal power consumption, but they do not analyze latency.

Vasić and Kostić [VK10] describe a distributed algorithm that uses adaptive link rates to reduce power consumption by modifying both topology and multi-path routing. Their idea is to distribute the load so that lines can be set to low-speed low-power operating modes. Their work provides a possible approach to actually implement power-saving measures, but does not focus on its effects on latency.

Revirigo et al. [RMHL10] analyze how adaptive link rates and burst transmissions can reduce the idle power consumption of single links. This method considers each link individually while I consider the whole network.

Other heuristics for finding solutions to the fixed-charge network flow problem exist [HNS09] which provably find high-quality solutions quickly, but these do not address the latency as I consider it.

Magnanti and Wong [MW84] provide an overview under which assumptions the network design problem is equivalent to other known problems and their complexity. This overview allows a quick estimation of how complex it would be to find energy-efficient configurations of models which consider different effects. An alternative formulation by Lin et al. [LS12], for example considers to deactivate unnecessary cables in bundled links.

The mathematical concept of k-spanners [CC95, PS89] formalizes the maximal stretch that the shortest path between two nodes may have in a subgraph. Analysis of k-Spanners for weighted graphs exist [ADD+93] as well as k-Spanners with fault toler-

ance [CLP09], but because k-Spanners have no notion of capacity they cannot fulfill all constraints that are needed for a feasible configuration. The results for k-Spanners can thus only be used if no edge reaches is capacity limit. I focus on the general case in which not all capacities can be assumed to be sufficient for all demands. In addition, k-Spanners only consider the maximum stretch, while I also consider different metrics, such as the average stretch.

Mean distances in graphs have been analyzed in general [DG77] and together with the minimum degree [KW97]. The difference between these two publications and the analysis of this chapter is that they focus on proving bounds for the mean distance and ignore capacities. Dankelmann and Entringer [DE00] show that spanning trees with certain mean distances exist, but do not take the capacity restrictions I consider into account either. Plesnik [Ple84] gives general results for the sum of all distances.

Another research area that uses topology control is ad hoc and sensor networks [FS07]. Topology control in this chapter as well as in ad hoc and sensor networks deactivate edges in a network graph to conserve power while maintaining desired properties such as low latency. So both have similar goals but the assumptions for the underlying network are different: While quality of a link to a neighbor can change rapidly in a wireless network, this will only rarely happen in wired networks. Additionally, reducing the node degree is not important in a wired network because I assume a network to be built only with node degrees it is able to handle.

3.3. Model

In this section, I describe the modeling assumptions I make for this chapter. The modeling in this chapter is different from the following chapters because it considers a wired network instead of wireless network.

3.3.1. Network graph

I formalize the network as a graph G with the vertexes $\mathrm{V_G}$ representing routers and the edges $\mathrm{E_G}$ representing line cards and wirings that connect routers. For each edge e the capacity $\mathrm{cap}(e)$ is the maximal data rate, the latency $\mathrm{L}(e)$ as the propagation delay, and $\mathrm{P}(e)$ is the power consumption under full load. The network has to serve unicast, static demands $\mathrm{D} \subseteq \mathrm{V_G} \times \mathrm{V_G}$; the function $\phi : \mathrm{D} \to \mathbb{R}^+$ maps the demands to the required data rates.

As I analyze propagation delay caused by *edges*, I consider only the power consumption of *edges*. Thus, to conserve power I deactivate edges and call the resulting states *active* and *sleep*. I represent the status of edges using configurations: a *configuration* C is a set of active edges together with routes for the demands using only active edges. A special configurations is a (not necessarily unique) configurations $\mathrm{C_L}$ which activates every edge and routes every demand on its shortest path to minimize latency. Additionally, I define the amount of data of a flow of demand d on an edge e as $\sigma^d_{\mathrm{C}}(e)$ and the utilization of

edge e as

$$\mathrm{U_C}(e) = \frac{\sum_{d \in \mathrm{D}} \sigma^d_{\mathrm{C}}(e)}{\mathrm{cap}(e)}. \tag{3.1}$$

This is equivalent to the always-on policy OnP and routing on shortest paths.

3.3.2. Power consumption

The power models for base stations (BSs) (see section 1.2) also apply well to wired network hardware. The binary power model models today's non-load-adaptive hardware well [CSB+08]. The linear model is motivated by techniques like adaptive link rates and burst transmissions that may reduce idle power consumption [RMHL10].

An algorithm which controls the topology to conserve power must consider idle power consumption, lest it might consume more power [PVC+09]. The reason for this is: Deactivating edges conserves power, but rerouting increases power consumption. If the later is greater than the former the total power consumption increases.

I apply the same power model I describe for BSs for lines in this chapter. The notation is as follows. I assume an active edge e consumes $\mathcal{P}\mathrm{P}(e)$ power when idle, where $\mathcal{P} \in [0, 1]$ is the fraction of maximum power consumed when idle. Power consumption scales linearly in utilization $\mathrm{U_C}(e)$ from $\mathcal{P}\mathrm{P}(e)$ when idle to $\mathrm{P}(e)$ under full load.

3.3.3. Latency

As I restrict my model of latency to propagation delays, the latency of a demand d is the sum of the latencies of its constituting edges; I call it $\mathrm{L_C}(d)$. To compare the latencies in two different configurations, I define the *stretch* of a configurations as the increase factor in all latencies caused by rerouting.

There are at least five different, intuitively reasonable ways to define the stretch of a configurations C compared to the latency-minimizing configurations $\mathrm{C_L}$. I introduce each metric and give its value for a simple example: a ring-shaped network with 8 nodes where I deactivate one edge to conserve energy. Figure 3.1 shows this example network. Every edge $e \in \mathrm{E_G}$ has latency $\mathrm{L}(e) = 1$ and there is a demand of $\phi(d) = 1$ from every node to every other node. I assume the capacities are large enough that an edge can be deactivated. In the following paragraphs, I give reasons why each metric is useful and give their values for the ring example with $n = 8$ (and the value it approaches when n approaches infinity).

The first two metrics express that *no* demand should suffer from a high latency due to power conservation. I define S^{MS} as the maximum stretch a *single* demand suffers in configurations C:

$$S^{\mathrm{MS}}{}_{\mathrm{C}}(\mathrm{D}) = \max_{d \in \mathrm{D}} \left(\frac{\mathrm{L_C}(d)}{\mathrm{L_{C_L}}(d)} \right). \tag{3.2}$$

In contrast, S^{SM} describes the stretch the maximum of *all* demands suffers in configu-

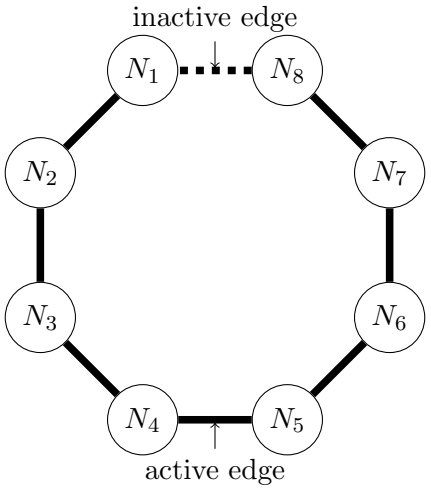

Figure 3.1.: An 8-Ring with a single inactive edge is an 8-Path. I use this example to demonstrate the different ways to compare latencies.

rations C:

$$S^{\mathrm{SM}}{}_{\mathrm{C}}(\mathrm{D}) = \frac{\max\limits_{d\in\mathrm{D}} \mathrm{L}_{\mathrm{C}}(d)}{\max\limits_{d\in\mathrm{D}} \mathrm{L}_{\mathrm{C_L}}(d)}. \tag{3.3}$$

While S^{MS} compares each latency to its *own* latency in the latency-minimizing configuration, S^{SM} compares the maxima of *all* latencies. In the 8-Ring, S^{MS} is 7 ($\lim_{n\to\infty} S^{\mathrm{MS}} = \infty$), because the highest increase in latency is from 1 (1) to 7 ($n-1$) and S^{SM} is $7/4 = 1.75$ ($\lim_{n\to\infty} S^{\mathrm{SM}} = 2$) the highest latency increases from 4 ($\lfloor n \rfloor$) to 7 ($n-1$).

To describe the tendency of the latencies, the next two metrics use the weighted arithmetic mean (weighted by the amount of transferred data $\phi(d)$) instead of the maximum. This is reasonable if it is acceptable for some demands to suffer from a high latency as long as the average stays low. Analogous to S^{MS} and S^{SM}, I define (a) S^{AS} as the weighted arithmetic mean of the stretch and (b) S^{SA} as the stretch of the weighted arithmetic mean latency:

$$S^{\mathrm{AS}}{}_{\mathrm{C}}(\mathrm{D}) = \operatorname*{avg}_{d\in D} \frac{\mathrm{L}_{\mathrm{C}}(d)}{\mathrm{L}_{\mathrm{C_L}}(d)} = \frac{\sum\limits_{d\in\mathrm{D}} \frac{\mathrm{L}_{\mathrm{C}}(d)}{\mathrm{L}_{\mathrm{C_L}}(d)}\phi(d)}{\sum\limits_{d\in\mathrm{D}} \phi(d)} \tag{3.4}$$

and

$$S^{\mathrm{SA}}{}_{\mathrm{C}}(\mathrm{D}) = \frac{\operatorname*{avg}\limits_{d\in\mathrm{D}} \mathrm{L}_{\mathrm{C}}(d)}{\operatorname*{avg}\limits_{d\in\mathrm{D}} \mathrm{L}_{\mathrm{C_L}}(d)} = \frac{\sum\limits_{d\in\mathrm{D}} \mathrm{L}_{\mathrm{C}}(d)\phi(d)}{\sum\limits_{d\in\mathrm{D}} \mathrm{L}_{\mathrm{C_L}}(d)\phi(d)}. \tag{3.5}$$

For the 8-ring example S^{AS} is $10/7 \approx 1.43$ ($\lim_{n\to\infty} S^{\mathrm{AS}} = 1.5$) and S^{SA} is $21/16 = 1.3125$ ($\lim_{n\to\infty} S^{\mathrm{SA}} = 4/3 \approx 1.33$).

One of the problems with the first four metrics is that they give different values when the order of combining (avg and max) and stretch calculation is reversed. The idea to use the weighted geometric mean S^{GS} is that this order is irrelevant and thus $S^{\mathrm{GS}} = S^{\mathrm{SG}}$ (see section A.1). They are defined by

$$S^{\mathrm{GS}}{}_{\mathrm{C}}(\mathrm{D}) = \operatorname*{geo}_{d\in\mathrm{D}} \frac{\mathrm{L_C}(d)}{\mathrm{L_{C_L}}(d)} = \left(\prod_{d\in\mathrm{D}} \left(\frac{\mathrm{L_C}(d)}{\mathrm{L_{C_L}}(d)}\right)^{\phi(d)}\right)^{\frac{1}{\sum_{d\in\mathrm{D}}\phi(d)}} \tag{3.6}$$

and

$$S^{\mathrm{SG}}{}_{\mathrm{C}}(\mathrm{D}) = \frac{\operatorname*{geo}_{d\in\mathrm{D}} \mathrm{L_C}(d)}{\operatorname*{geo}_{d\in\mathrm{D}} \mathrm{L_{C_L}}(d)} = \frac{\left(\prod_{d\in\mathrm{D}} \mathrm{L_C}(d)^{\phi(d)}\right)^{\frac{1}{\sum_{d\in\mathrm{D}}\phi(d)}}}{\left(\prod_{d\in\mathrm{D}} \mathrm{L_{C_L}}(d)^{\phi(d)}\right)^{\frac{1}{\sum_{d\in\mathrm{D}}\phi(d)}}} \tag{3.7}$$

and are approximately 1.22 ($\lim_{n\to\infty} S^{\mathrm{GS}} = 2/\sqrt{\mathrm{e}} \approx 1.21$, see section A.5) in the 8-Ring.

The values for the different metrics show that most metrics do not differ by much in this example. In the next section I analyze the relationships between them in general and give a bound for their ratio.

Table 3.1.: My modeling decisions for the wired example are highlighted in **bold**).

Option	Possible choices	
Topology control	**deactivated edges**	deactivated routers
Change of demands	**static**	dynamic
Power consumption model	**binary**	linear
Considered delay	**propagation**	queuing
Topology control and routing	distinct	**combined**
Location of calculations	distributed	**centralized**
Knowledge of demands	edge utilization	**end-to-end demands**

Table 3.1 summarizes the modeling choices I have mode for this chapter.

3.4. Analysis of metrics for latency aggregation

In this section, I first analyze the relationships between the different latency metrics in the same situation. Then, I describe the relation of BDP and latency.

3.4.1. Relationships between the latency metrics

First note that S^{MS} is always larger than or equal to any of the other four metrics. Between any two of the other metrics, only one comparative statement holds in general:

$S^{\mathrm{GS}} \leq S^{\mathrm{AS}}$. This is a direct implication of generalizing the inequality of arithmetic and geometric means to weighted means [Ste04]. While all possible orders of the metrics S^{SM}, S^{AS}, and S^{SA} exist (see section A.2) not all are possible with two unweighted demands (see section A.3).

I define the *skew* of configurations C skew(C) as the ratio between maximal and minimal latency:

$$\mathrm{skew}(\mathrm{C}) = \frac{\max\limits_{d \in \mathrm{D}} \mathrm{L_C}(d)}{\min\limits_{d \in \mathrm{D}} \mathrm{L_C}(d)}. \tag{3.8}$$

The bound

$$A \leq \mathrm{skew}(\mathrm{C}) \cdot \mathrm{skew}(\mathrm{C_L}) \cdot B \tag{3.9}$$

holds for all combinations of A and B from the five metrics (see section A.4). It implies that all five metrics yield similar values when the ratio between maximum and minimum latency in both the latency-minimizing configurations $\mathrm{C_L}$ and in configurations C is small. For some pairs of metrics tighter bounds are possible, but because examples exist that are arbitrarily close to this bound it is the tightest possible general bound. Note that this bound depends on neither number nor size of demands and thus the number of demands can be arbitrarily large and all metrics still fulfill the inequality above.

I conclude that the metrics are similar for low skew and are bounded independently of the number and size of demands. I analyze how much the metrics differ for a practical scenario in section 3.6. Next, I show a connection between BDP and latency.

3.4.2. Bandwidth-delay product and latency

The bandwidth-delay product (BDP) is the product of the data rate of an edge with its latency:

$$\mathrm{BDP}(e) = \mathrm{cap}(e)\mathrm{L}(e). \tag{3.10}$$

I define the *used* BDP of an edge e as the latency times the actually used data rate. In contrast, the *unused* BDP is the BDP which is *not* used to transfer data and *sleeping* BDP is the BDP of edges in sleep mode. Theorem 3.1 shows that if the used BDP is low, the average latency will be low and so will S^{SA}.

Theorem 3.1. *The mean used bandwidth-delay product (BDP) of all edges is proportional to the mean latency of all demands.*

Proof. The idea of the proof is that the order in which the total sum latency in the system of determined (sum over flows and edges or over edges and flows) does not matter.

Recall the definition of the weighted arithmetic mean latency

$$\operatorname*{avg}_{d \in \mathrm{D}} \mathrm{L_C}(d) = \sum_{d \in \mathrm{D}} \mathrm{L_C}(d)\phi(d) / \sum_{d \in \mathrm{D}} \phi(d). \tag{3.11}$$

I consider the used BDP averaged over all edges, captured per edge as the product of its utilization $\mathrm{U_C}$ in configuration C it capacity cap and its latency L. It is proportional

to mean latency, because

$$\begin{aligned}
\operatorname*{avg}_{e\in \mathrm{E_G}} \overbrace{\mathrm{U_C}(e)\mathrm{cap}(e)\mathrm{L}(e)}^{\text{used BDP on edge } e} &= \frac{1}{|\mathrm{E_G}|}\sum_{e\in \mathrm{E_G}} \sigma(e)\mathrm{L}(e) \\
&= \frac{1}{|\mathrm{E_G}|}\sum_{e\in \mathrm{E_G}}\sum_{d\in \mathrm{D}} \sigma^d_{\mathrm{C}}(e)\mathrm{L}(e) \\
&= \frac{1}{|\mathrm{E_G}|}\sum_{d\in \mathrm{D}}\sum_{e\in \mathrm{E_G}} \sigma^d(e)\mathrm{L}(e) \\
&= \frac{1}{|\mathrm{E_G}|}\sum_{d\in \mathrm{D}} \mathrm{L_C}(d)\phi(d) \\
&= \frac{\sum_{d\in \mathrm{D}}\phi(d)}{|\mathrm{E_G}|}\operatorname*{avg}_{d\in \mathrm{D}} \mathrm{L_C}(d) \\
&\sim \operatorname*{avg}_{d\in \mathrm{D}} \underbrace{\mathrm{L_C}(d)}_{\text{latency of demand } d} \qquad (3.12)
\end{aligned}$$

□

The following related theorem follows directly from the definition the binary power profile.

Theorem 3.2. *When the power consumption of each edge is proportional to its total BDP and the binary model for power consumption is assumed, the BDP of sleeping edges is proportional to the conserved power.*

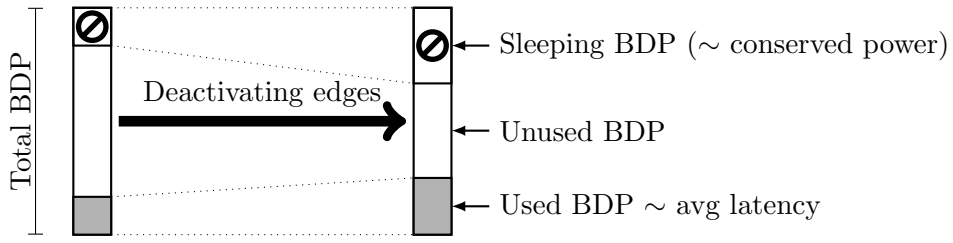

Figure 3.2.: Deactivating edges and rerouting demands conserves power, but increases latency. Both of which can be quantified in terms of BDP.

Figure 3.2 illustrates how BDP, power consumption, and latency are related as described by the theorems 3.1 and 3.2. Each of the two vertical bars represents the total BDP that is available in a network and shows how it is used. Rerouting traffic and deactivating edges will increase the used BDP, but also allows edges to be deactivated.

3.5. Optimization model

To get a general understanding of the power consumption of practical networks I formulate the problem as an MILP. It is based on the following assumptions: A single algorithm controls routing and topology, the algorithm has global knowledge of end-to-end flows, the demands are static, and power consumption follows the linear model (with the binary power model a special case).

$$\forall e \in \mathrm{E_G} \quad : \quad \sum_{d \in \mathrm{D}} \max(0, \sigma_{\mathrm{C}}^d(e)) \leq \mathrm{cap}(e)\mathrm{U_C}(e) \tag{3.13}$$

$$\forall e \in \mathrm{E_G} \quad : \quad \mathrm{U_C}(e) \leq x_{\mathrm{C}}(e) \tag{3.14}$$

$$\forall d \in \mathrm{D}; u, v \in \mathrm{V_G} \quad : \quad \sigma_{\mathrm{C}}^d((u,v)) = -\sigma_{\mathrm{C}}^d((v,u)) \tag{3.15}$$

$$\forall d \in \mathrm{D}, u \in \mathrm{V_G} \quad : \quad \sum_{v \in N(u)} \sigma_{\mathrm{C}}^d((v,u)) - \sum_{v \in N(u)} \sigma_{\mathrm{C}}^d((u,v)) = \begin{cases} -\phi(d), & \text{if } u = \mathrm{src}(d) \\ \phi(d), & \text{if } u = \mathrm{dest}(d) \\ 0, & \text{else} \end{cases} \tag{3.16}$$

$$\sum_{e \in \mathrm{E_G}} \mathrm{U_C}(e)\mathrm{cap}(e)\mathrm{L}(e) / \operatorname*{avg}_{d \in \mathrm{D}} \mathrm{minLat}(d) \leq B^{\mathrm{SA}} \cdot \sum_{d \in \mathrm{D}} \phi(d) \tag{3.17}$$

$$\sum_{d \in \mathrm{D}} \sum_{e \in \mathrm{E_G}} \mathrm{L}(e) \max(0, \sigma_{\mathrm{C}}^d(e)) / \mathrm{minLat}(d) \leq B^{\mathrm{AS}} \cdot \sum_{d \in \mathrm{D}} \phi(d) \tag{3.18}$$

$$\min \sum_{e \in \mathrm{E_G}} \mathrm{P}(e)((1 - \mathcal{P})\mathrm{U_C}(e) + \mathcal{P} x_{\mathrm{C}}(e)) \tag{3.19}$$

I use a network flow model to calculate allocation of demands to edges. To express whether an edge is active or sleeping I use the binary variable $x_C(e)$. Using the notation introduced in section 3.3, I define the capacity constraint in equation 3.13. I use $\max(0, f)$ to model full duplex links. Equation 3.14 guarantees that only active edges can transfer data. Equation 3.15 is the skew symmetry. I define the flow conservation in equation 3.16 so that the allocation of flows meets all demands. $N(v)$ denotes the set of vertices incident to the vertex v. Note that these definitions allow multi-path routing. To specify an upper bound B^{SA} (B^{AS}) on the S^{SA} (S^{AS}) metric I use equation 3.17 (3.18). I only use one of these constraints at a time. Here $\mathrm{minLat}(d)$ is the minimal latency necessary to route demand d. I minimize the power consumption specified in Term 3.19.

Using the MILP I am able to specify upper bounds for the average stretch metrics S^{AS} and S^{SA}. Because the geometric upper bound cannot be written as a linear constraint, I am not able to use it in the optimization model. Because my network model is based on the idea of flows, which allows multi-path routing, I cannot calculate the maximum latency in the linear program either. Hence, I cannot bound the maximum stretch metrics S^{SM} and S^{MS}.

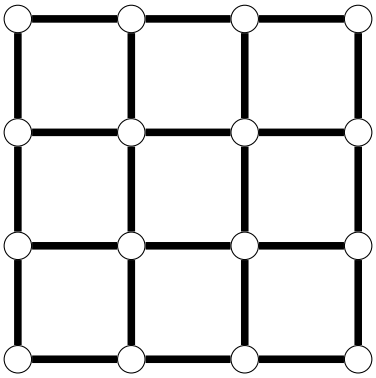

Figure 3.3.: The (4×4)-grid I use as an example network. The capacity, latency and energy consumption of all edges is 1. The demands are created from every node to every other node with equal sizes.

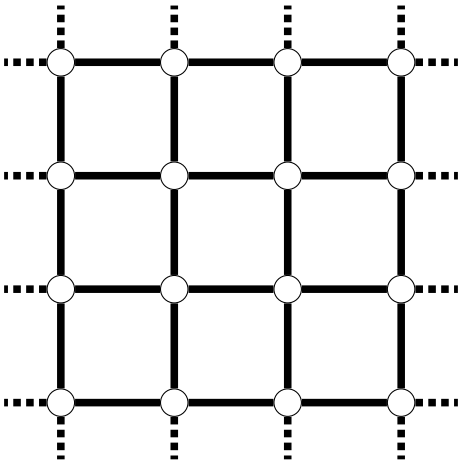

Figure 3.4.: The 4-dimensional hypercube is topologically equivalent to (and easiest to draw in 2 dimensions as) a (4×4)-grid with wrap-around edges. The capacity, latency and energy consumption of all edges is 1. The demands are created from every node to every other node with equal sizes.

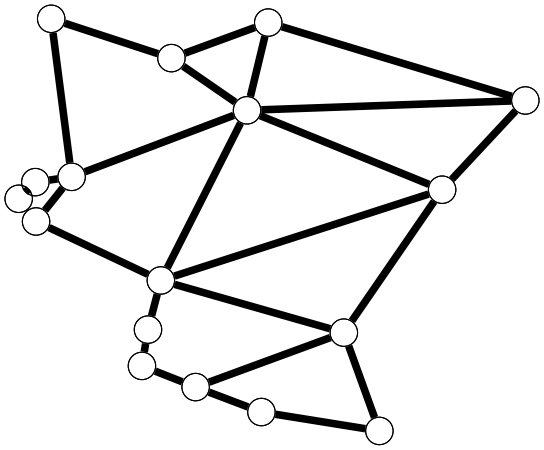

Figure 3.5.: The `nodel-germany` network is from SNDlib [OPTW07]. The demands, capacity and energy consumption of all edges is provided in the model. I determined the latency from the geometric distances.

3.6. Results

I use my model to calculate the power consumption in different networks: a two-dimensional Grid (see figure 3.3), a hypercube (see figure 3.4), and as a practical network I use the `nobel-germany` (see figure 3.5) network from the Survivable fixed telecommunication Network Design library (SNDlib) [OPTW07]. The SNDlib includes power consumption (cost), capacity and a demand pattern, while I estimate the latencies from the physical distances of the routers (locations are present in the model). For both theoretical networks I set all capacities, latencies, and power consumption values to 1 and assume equal demands from every nodes to every other node. I scale the network load with a single scalar factor and solve the optimization problems for minimal power consumption with the GNU Linear Programming Kit (v4.44) and the Gurobi Optimizer 4.0. The error bars show the upper and lower bounds provided by the solver.

Figure 3.6 compares both the most power-efficient configurations and the latency-minimizing configurations as well as the power models. Because the difference between the two power models is small, I use the simpler binary model for further analysis. Additionally, Figure 3.6 shows that it is possible to conserve up to 45% of power. The amount of conserved power depends on the type of network considered, because a minimum-power spanning tree is needed for connectivity between all nodes. Because the load on the edges is inhomogeneous and I use a single scale factor, it is possible to deactivate edges even at the highest load the network can transfer.

Figure 3.7 illustrates the power necessary to keep the stretch below a given bound. Allowing the latency to increase by 20% saves up to 39% of power. Although both average metrics, S^{AS} and S^{SA}, are generally different, they produce similar results in this scenario. Figure 3.8 shows the difference when bounding the metrics by 1.2 for

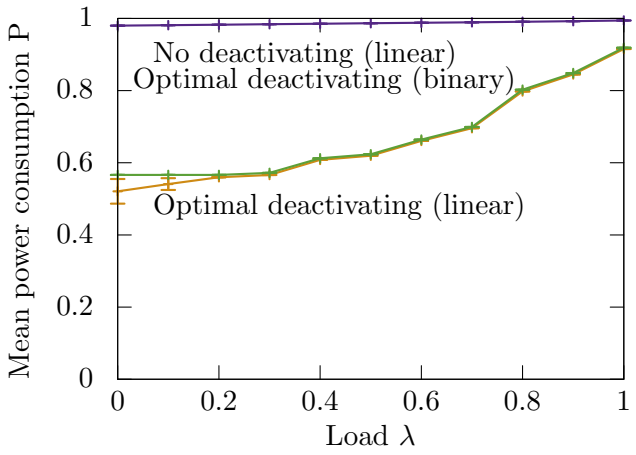

Figure 3.6.: Power consumption with different models using the `nobel-germany` network; The idle power consumption $\mathcal{P}$ is 0.98 in the linear model.

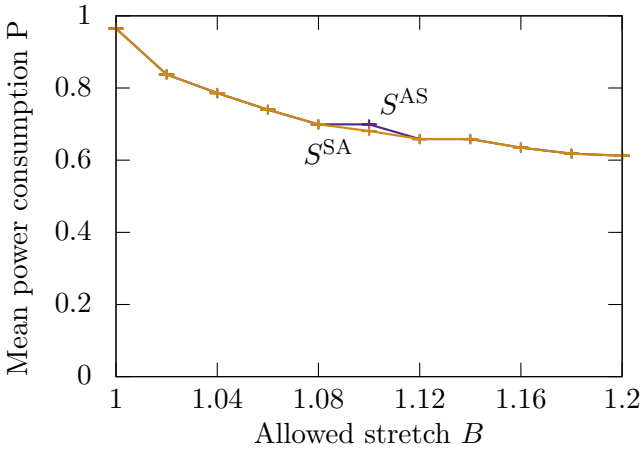

Figure 3.7.: Power consumption for different upper bounds for the stretch metrics S^{AS} and S^{SA} in the `nobel-germany` network

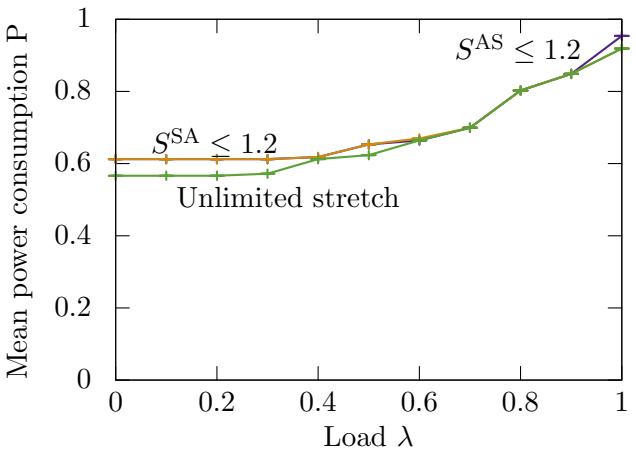

Figure 3.8.: Power consumption increases only slightly for S^{AS} and S^{SA} bounds of 20% increased latency in the `nobel-germany` network.

different load levels. Again both metrics do not differ much and allowing a stretch of 1.2 consumes less than 5% more power than not limiting the stretch at all.

Figure 3.9 (3.10) shows the power consumption of different methods to conserve power in a 4-dimensional hypercube ((4×4)-grid). I compare the optimal deactivating of edges to a method of reducing the idle power consumption $\mathcal{P}$ of all edges to 50% and the combination of both. The results show that both approaches reduce the consumed power in both networks and the power consumption is further reduced by combining both approaches. This exemplifies how methods to reduces the energy consumption on different levels interact and are necessary to conserve most energy.

3.7. Conclusion

I formalized the problem of conserving power in wired networks, applied my model to different networks, and analyzed power consumption and latency. Assuming the binary power-consumption model, this allows the consumed power to be reduced by 39% while increasing the latency by only 20% in the `nobel-germany` network. I conjecture that for most networks and demand patterns configurations exist that conserve large amounts of power and increase latency only slightly.

There are many ways to describe the cost in latency that the users have to accept for reducing the energy consumption. I analytically showed that they cannot be arbitrarily different and their difference depends on the skewness of the latency of the demands. I also showed that for a practical example the difference between the different metrics is negligible.

In future work more complex models can be analyzed, for example arbitrary power profiles. Another approach is to consider not only propagation delays, but to include

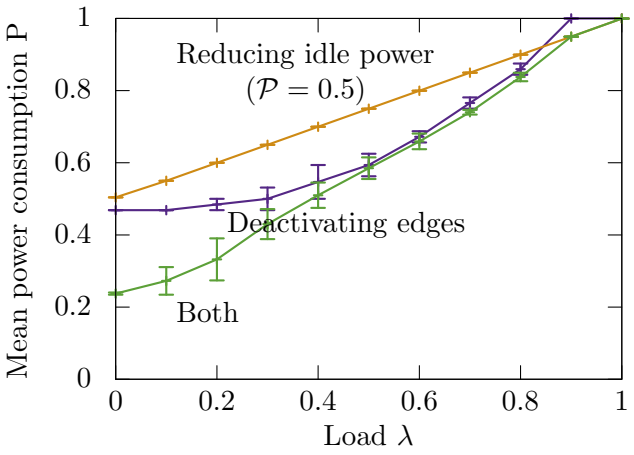

Figure 3.9.: Different ways to conserve power are illustrated for the example of the Hypercube of dimension 4.

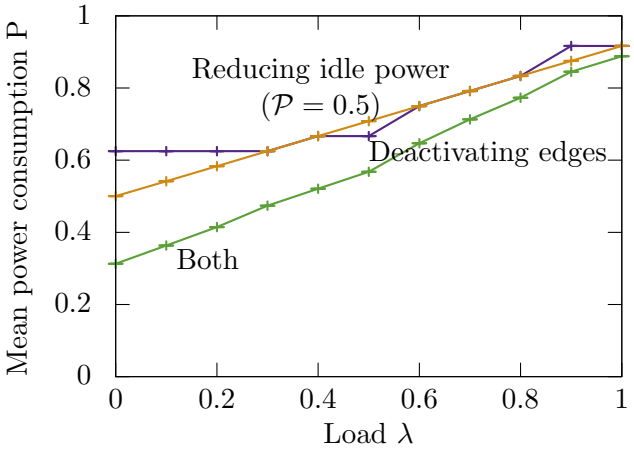

Figure 3.10.: Different ways to conserve power are illustrated for the example of the (4×4)-grid.

queuing delays into the model and test how they interact with power conservation methods.

In this chapter, I showed how a system that is composed of smaller sub-systems can have a different power profile than the sub-systems. In this example the power consumption of the network which consists the power consumption of the edges is closer to the linear power profile than the power profile of the edges, which have a (nearly) binary power profile. To do this the power-cycle durations do not have to be in the same order as request processing times (as needed for chapter 2), but only on the order of changes of the average traffic.

In the rest of the dissertation I apply the same idea to deactivating selected BSs in RAN. I consider the energy consumption of signaling and data transmission independently first (chapters 4 and 5) and later combine all effects in a simulation (chapter 7).

4. Conserving energy in signaling transmissions

In this chapter, I answer the question of how much power can be conserved in radio access networks (RANs) when some base stations (BSs) use cooperative transmissions to signal at extended range. In addition, I analyze the effect of different path-loss exponents and varying number of cooperating BSs.

First, I analytically show how it is possible to cover an area with fewer BSs when their range is extended by cooperation. To do this, I analytically determine the optimal spacing of BSs under the assumption that the BSs can be freely placed. Second, I create and solve an optimization model that describes the power necessary to cover a given area using linear programming. In the optimization model I assume the BSs are already placed and it is only possible to select which of them to activate.

4.1. Introduction

BSs must provide signaling traffic to every location at all times. Hence, they cannot simply be deactivated under low load. A BS can only be deactivated when the neighboring BSs cover its area. Because BSs can only cover the area they are physically close to, methods to extend their ranges are necessary. While different such methods exist, I consider the use of cooperative transmissions (see section 1.4) to extend the ranges of BSs.

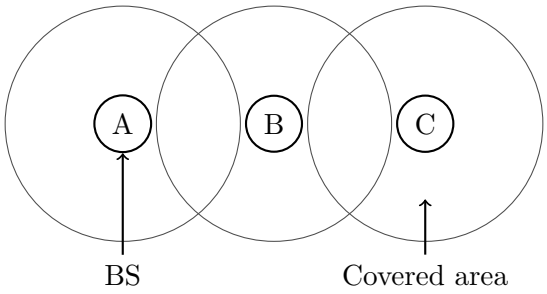

Figure 4.1.: This illustrates three BSs and the area they cover without cooperation.

Figure 4.1 shows the BSs A, B, and C and the area to which they can provide signaling traffic without cooperation. Figure 4.2 shows how the BSs A and C can use cooperation

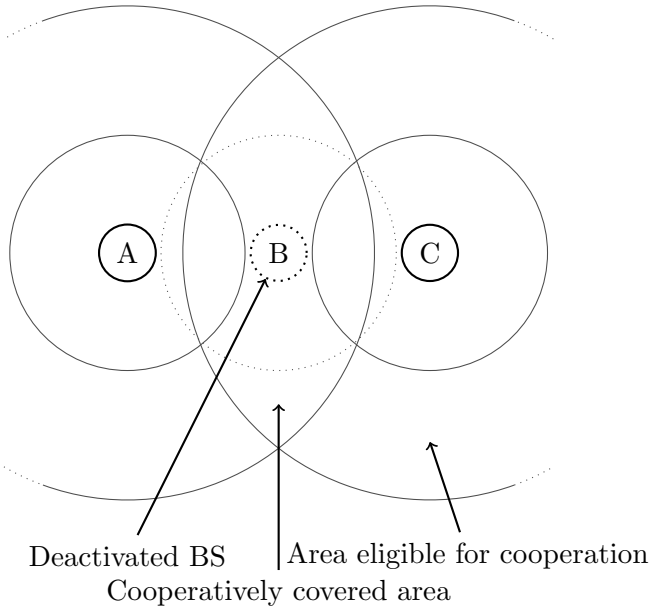

Figure 4.2.: BSs A and C can cooperate to cover the area of BS B (not to scale).

to cover the area previously covered by BS B. This allows BS B to be deactivated and thereby conserve power.

In this chapter, I calculate a theoretically possible increase in covered area under the assumption that the BSs are placed to maximize the covered area when taking cooperation into account. Because BSs are impractical to move in reality, I determine how to deactivate BSs so they can conserve energy when the load is low. I create an optimization model to determine how many BSs, which are placed for non-cooperative coverage, can be deactivated when cooperation is introduced. Up to this point I only consider ergodic capacity. Finally, I briefly compare the area BSs cover when I consider outage capacity instead of ergodic capacity to determine the ranges for signaling.

4.2. Related work

The BS deployment problem is well known and analyzed, for example by Lev-Tov and Peleg [LTP02]. They analyze in which given locations to build BSs, but I determine the best locations when the BSs can be freely placed. Also Lev-Tov and Peleg do not consider cooperative transmissions.

An abstraction that covers both the design time and run-time decisions is the fixed-charge facility location problem [Noz01]. It is also well studied but does not consider cooperation of BSs. While there is other work that analyzes cooperation in RANs [DD08, SSBNS06, SSPS09], it does not consider its use for energy conservation.

Niu et al. [NWGY10] explain how changing the cell sizes (called cell zooming) can be used to reduce the power consumption of RANs. They focus on the effects that changeable cell sizes have on the energy consumption but only briefly state examples how the cell size can actually be changed. The BS cooperation I analyze in this chapter is one of their examples, but they do not provide a detailed analysis.

Richter et al. [RFF09] consider how to minimize the power needed to cover an area with micro and macro cells. They focus on calculating the optimal number of micro BSs. In contrast to this I try to deactivate BSs to conserve power. They assume the power consumption of macro BSs to be static and the power consumption of micro BSs to be dynamic, which is also supported by Arnold et al. [ARFB10]. Using these assumptions they show that in low load scenarios micro BSs are more energy-efficient. This effect is orthogonal to my investigation because both aspects can be combined into a single model.

Sadek et al. [SHL06] propose a protocol to increase the range of BSs by cooperation of user equipments (UEs). They use UEs as cooperation partner and have to consider an additional channel to get the transmission to the partner. I use another BS as cooperation partner and consider only two wireless channels because I assume the BSs are connected by a wired backhaul network. My assumption is realistic (for cooperating BSs) and gives a lower error probability.

Viterbi and Gilhousen [VGZ94] as well as Staehle et al. [SLH+01] make an analysis that is similar to mine for soft handovers. Soft handovers use selection combining, while I consider maximal-ratio combining (MRC).

In the final report of "Understanding the Environmental Impact of Communication Systems" [FDM+09] the authors hint at a method to "Power-down during quiet periods" for GSM that saves around 25% of the consumed energy, but no further references or details are provided.

Lichte et al. [LFK10] consider how cooperation can be used in wireless sensor networks, but focus on different aspects, like building broadcast trees and determining a useful partner for cooperation, while I focus on covering an area with the least amount of power.

Woldegebreal et al. [WK08] analyze how cooperation can increase the range of transmissions in wireless sensor networks. They consider a three-node scenario (sender, relay, receiver) and require the cooperation partner (relay) to receive the transmission using a wireless channel. Woldegebreal et al. focus on reducing the outage probability, while I focus on reducing the power consumption, which is different from radiated power.

Blundon [Blu10] analyzes a variant of the circle-covering problem, which requires every point to be covered by a given number of circles. The key difference is that I only need a location to be covered by several BSs if no BS is close enough to provide connectivity alone.

4.3. Model

In this chapter, I analytically determine the highest distance at which cooperating BSs can be placed so that the BSs can still provide the signaling data rate D_S to every possible point. The radio model I use here considers only the path loss over distance. I do not consider time, load, UEs, and interference in this chapter.

The signaling traffic UEs generate is created by a number of (possibly moving) users. Because the signaling traffic contains only small amounts of data and the need to transmit signaling traffic can arise nearly everywhere I model this as a need to transmit with the signaling data rate D_S to every possible UE location. Because this ignores the number of UEs which want to transmit signaling data, the modeling is only valid as long as the signaling data rate is small enough to not overload BSs.

4.3.1. Non-cooperative range

Using non-cooperative transmission the requirements for signaling explained in chapter 1 define a location u as covered when it closer to a BS B than the signaling range of the BS:

$$\mathrm{d}(u, B) \leq r_S, \tag{4.1}$$

where $\mathrm{d}(u, B)$ is the Euclidean distance between u and B.

4.3.2. Cooperative range

A simple way to guarantee a UE reaches this threshold T is to make sure it receives a mean signal-to-noise ratio (SNR) of at least $T/2$ from two different BSs. I call this type of cooperation *2-cooperation* because it allows two BSs to cooperate and *limited* because I set a lower bound on the received SNR per BS such that receiving such power from two BSs guarantees that the sum of their SNRs is always at least T. Limited cooperation is more restrictive than necessary, but it provides a simple abstraction. The range at which a base station can provide an SNR of at least $T/2$ is $2^{1/\delta}r$ where δ is the path-loss exponent and r is the range of a non-cooperating BSs. Hence, for the special case of $\delta = 2$, the cooperative range is $\sqrt{2}r$. For a limited cooperation of n BSs, every BS needs to provide at least T/n of the mean SNRs.

For *unlimited cooperation*, I remove the lower limit of T/n and require only the sum of the mean SNRs of the cooperating BSs to be higher than the threshold T. Unlike limited cooperation the resulting problem cannot be stated as purely geometric problem. For each possible location of a UE u the definition from the introduction (section 1.4) is equivalent to

$$\sum_{i \in S} \left(\frac{r_S}{\mathrm{d}(u, i)} \right)^{\delta} \geq 1. \tag{4.2}$$

The equivalent form makes it easier to determine the increase in covered area by using cooperation.

4.4. Signaling with optimal BS spacing

In this section, I define different types of cooperation and analytically show how the covered area is increased when cooperation is taken into account for the deployment of the BSs. The size of this area is a proxy for power consumption, because a greater area can be covered with fewer BSs if their cells are larger. In this section, I consider the constraint for connection to be based on ergodic capacity.

In the next subsection, I calculate the cell area without cooperation as a reference for later comparison. Second, I give a simplified cooperative model which describes coverage as a purely geometrical problem. And third, I describe a more powerful abstraction of cooperation, in which optimal deployment of BSs is not a purely geometric problem anymore.

4.4.1. No cooperation

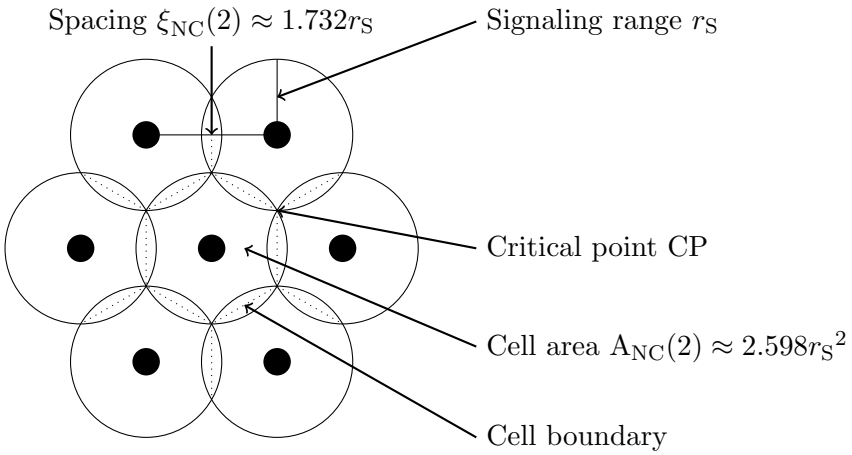

Figure 4.3.: The best non-cooperative coverage is achieved with a hexagonal pattern.

Given my assumptions a single BS can cover the area of a disk. The most efficient pattern to cover an infinite plane with disks is the hexagonal pattern [Ker39, Tot72], which figure 4.3 shows. I define the distance of two neighboring BSs in the hexagonal pattern as *spacing*. The hexagonal pattern allows neighboring BSs to be placed with a spacing of $\xi = \sqrt{3} r_{\mathrm{S}}$ and still cover the infinite plane. Where r_{S} is the signaling range of the BSs. The cell area is uniquely defined for a given spacing ξ as I only consider hexagonal deployments. In general I use $\xi_X(\delta)$ to denote the highest spacing which can provide every location in the plane with signaling traffic, where X is the type of cooperation used and δ is the path-loss exponent of the scenario. A *critical point* CP is a point which receives just enough power to decode the signaling transmission and thus will lose its connectivity when the spacing is increased. This is the point that has the highest distance to any BSs.

Basic geometry limits the signaling cell area of a BS to $\pi {r_S}^2$. But when a larger area has to be covered, cells overlap and the area served by a particular BS of combination of BSs (called *cell area*) $A_{NC}(2)$ is only $3^{3/2}{r_S}^2/2$. In the following subsections, I use this area as a reference value and define the *gain factor* $g_X(\delta)$ of a cooperation type X as the factor by which the cell area A increases compared to not using cooperation.

4.4.2. Limited 2-cooperation

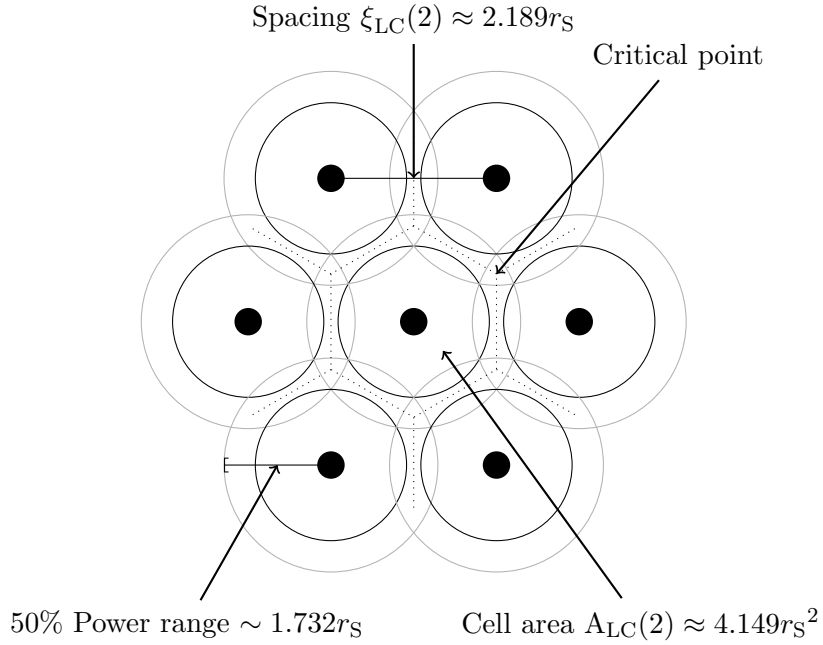

Figure 4.4.: Using limited cooperation of 2 BSs the area each BS is responsible for can be increased by 59.7% when the path-loss exponent δ is 2.

As defined in section 1.4 a location is covered by limited 2-cooperation, when two BSs each provide at least half of the required signal strength. To cover the infinite plane with limited 2-cooperation each point must be closer than r_S to a BS or closer than $\sqrt{2} r_S$ to two different BSs. A possible deployment of BSs is shown in figure 4.4. Using only basic geometry I derive the spacing of $\xi_{LC}(2) = \sqrt{\sqrt{21}+5}\, r_S/\sqrt{2}$. Hence, the cell area is $A_{LC}(2) = 3\sqrt{7}{r_S}^2/4 + 5\sqrt{3}{r_S}^2/4$, which is equivalent to a $g_{LC}(2) = \left(\sqrt{21}+5\right)/6$ increase in cell area over not using cooperation. The BS increases its range by cooperating with the corresponding neighbor at the edges of the cell. Further increasing the spacing will not cover the plane anymore because the critical point would be further away from all BSs than r_S and would not have two BSs in cooperative range.

For path-loss exponents $\delta > 2$ the range at which a UE receives an SNR of at least $T_S/2$ decreases in comparison to the range at which it receives SNR T_S. Hence, the

gain using limited 2-cooperation also decreases for higher path-loss exponents. The cooperative range of $2^{1/\delta} r_S$ and the same geometry used for the special case $\delta = 2$ allow me to calculate the maximum spacing with an arbitrary path-loss exponent δ:

$$\xi_{LC}(\delta) = \frac{\sqrt{\sqrt{3}\sqrt{2^{2/\delta+2} - 1} + 2^{2/\delta+1} + 1}}{\sqrt{2}} r_S. \tag{4.3}$$

This directly results in a cell area of

$$A_{LC}(\delta) = \frac{3\sqrt{2^{2/\delta+2} - 1} + \sqrt{3}2^{2/\delta+1} + \sqrt{3}}{4} {r_S}^2 \tag{4.4}$$

and thus a gain factor over not using cooperation of

$$g_{LC}(\delta) = \frac{\sqrt{3}\sqrt{2^{2/\delta+2} - 1} + 2^{2/\delta+1} + 1}{6}. \tag{4.5}$$

This deployment provides a substantial gain over non-cooperative deployment. Because a proof of its optimality is not evident, I do not know if this deployment is the best deployment possible. While no hexagonal deployment with higher spacing is possible, still other non-hexagonal deployments might need fewer BSs. But because the cell area without cooperation is proven to be optimal [Ker39, Tot72] this is a lower bound for the gain that can be achieved using limited 2-cooperation.

4.4.3. Unlimited 2-cooperation

Figure 4.5 shows a deployment of BSs with a spacing of $\xi_{UC}(2) = \sqrt{6} r_S$ which allows each of the BSs to cover a cell of size $A_{UC}(2) = 3^{3/2} {r_S}^2$. This is a 100% increase in covered area over not using cooperation.

To understand why this deployment covers the complete plane, assume a UE u is connected by unlimited 2-cooperation to the two BSs A and B. Without loss of generality I assume that A is closer to u than B. The following two lemmas can be proven easily:

Lemma 4.1. *Every UE u' that is closer to both BS A and B than u is connected by unlimited 2-cooperation.*

Lemma 4.2. *Every UE u' that is on a line between u and BS A is connected by unlimited 2-cooperation.*

In lemma 4.1 the signal strength of each BSs is higher in u' than in u and, thus, the total signal strength must also be higher.

For lemma 4.2 it is necessary to show that moving closer to the closest BSs increases total signal strength. To prove this it is important that the increase in distance to BS B is at most the decrease in distance to BS A. Therefore, for a path loss exponent of at least 1 (which is true for all practical purposes) the total signal increases from u to u'.

Every point in the plane can be reached from a critical point, which is connected, by using each of the two lemmas once. Figure 4.6 illustrates how starting at a critical point

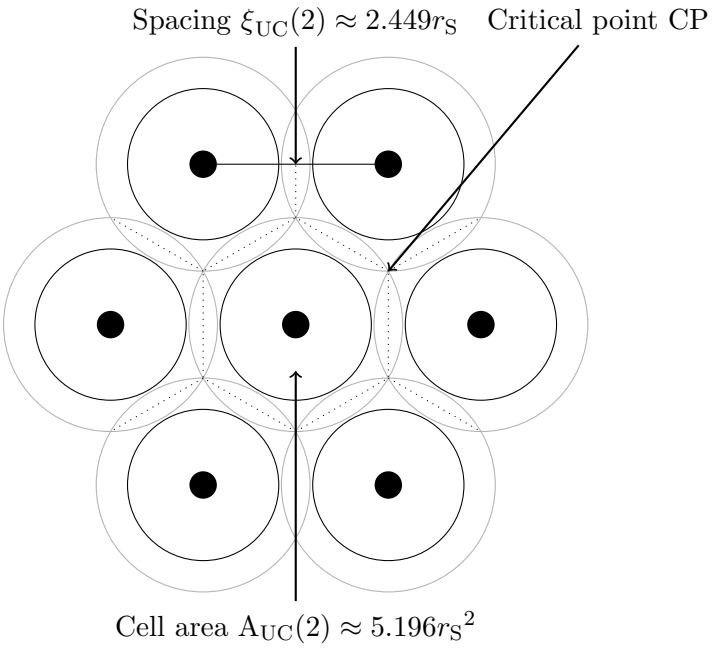

Figure 4.5.: Using unlimited cooperation of 2 BSs doubles the area each BS covers.

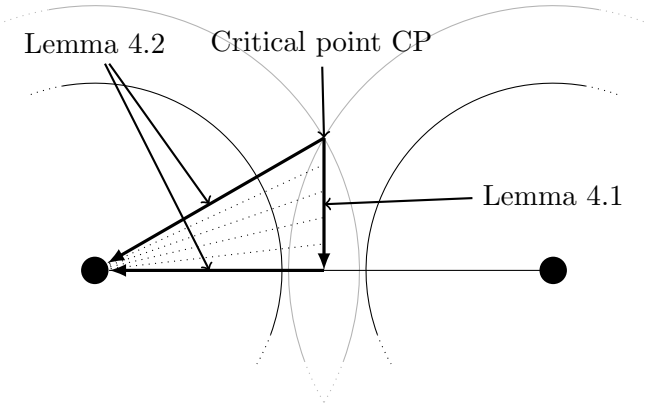

Figure 4.6.: Using lemmata 4.1 and 4.1 to show why the deployment of figure 4.5 covers the plane

and moving along the lines that have coverage granted by the two lemmas reaches any point. Thus, every location in figure 4.5 is covered.

For general path-loss exponents δ it is possible to increase the spacing to $\xi_{\mathrm{UC}}(\delta) = \sqrt{3}\,2^{1/\delta}$ and still cover the plane. This results in a cell area of $A_{\mathrm{UC}}(\delta) = 3^{3/2}\,2^{2/\delta-1}$ and a gain factor of $g_{\mathrm{UC}}(\delta) = 2^{2/\delta}$, which can be shown using geometry and lemmata 4.1 and 4.2.

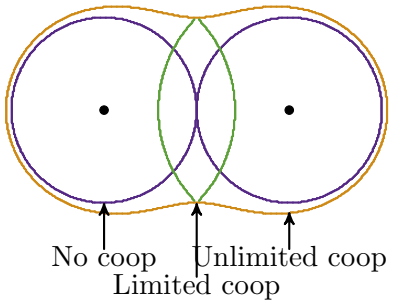

Figure 4.7.: The area that limited cooperation can cover fully contained in the area that unlimited cooperation can cover.

To illustrate the differences between the limited and unlimited cooperation, Figure 4.7 shows the area that can be covered without cooperation, by limited 2-cooperation, and by unlimited 2-cooperation.

4.4.4. Infinite cooperation

The next step to generalize cooperation is to allow more than two BSs to cooperate. I call the cooperation of n BSs n-cooperation. When each BS has to provide at least T_{S}/n SNR, I call it limited n-cooperation and if only the sum of the SNRs has to be greater than T_{S} I call it unlimited n-cooperation.

I do not calculate the gain of n-cooperation for any given n here, but analyze how coverage changes when n approaches infinity. Note that the difference between limited and unlimited cooperation vanishes when n approaches infinity, because the individual threshold a BS has to achieve tends towards zero. I call this *infinite cooperation* and denote it as IC.

As the received power is lowest at the critical points, I calculate the received power for a critical point depending on the spacing. Figure 4.8 shows how I number the BSs as seen from a critical point. For every $i \in \mathbb{N}$ and $j \in \{1, ..., 3i-2\}$ a BS (i,j) exists. Note that this is exactly one third of all BSs. BS (i,j) is located at

$$\text{Location of BS}(i,j) = \left((3i-2-3(j-1)/2)\xi \,\Big/\sqrt{3}\;,\; \xi(j-1)/2\right). \tag{4.6}$$

The distance between the critical point and BS (i,j) is

$$\mathrm{d}(\mathrm{CP}, \mathrm{BS}(i,j)) = \sqrt{3\,j^2 - 9\,i\,j + 9\,i^2 - 3\,i + 1}\cdot \xi \,\Big/\sqrt{3} \tag{4.7}$$

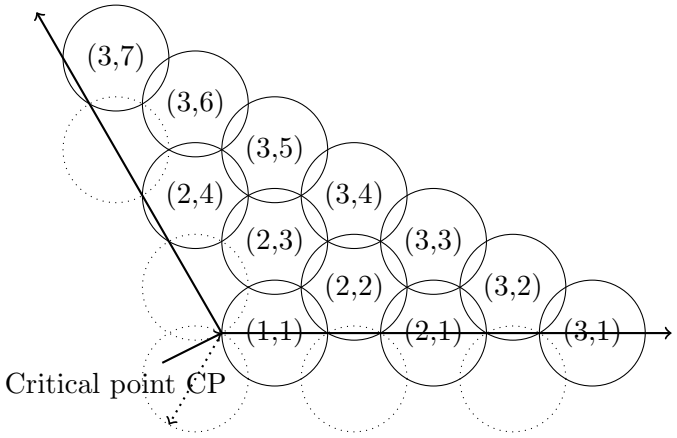

Figure 4.8.: The way I use to number a third of all BSs is rotational symmetric to the other two thirds.

and the expression

$$(3i/2-1)\xi/\sqrt{3} \leq \mathrm{d}(\mathrm{CP},\mathrm{BS}(i,j)) \leq \mathrm{d}(\mathrm{CP},\mathrm{BS}(i,1)) = (3i-2)\xi/\sqrt{3} \tag{4.8}$$

bounds the distance of the critical point CP from BS (i,j) independent of j, where ξ is the spacing of the hexagonal pattern.

For a UE at the critical point to be connected, the following equivalent to unlimited 2-cooperation must hold:

$$3\sum_{i=1}^{\infty}\sum_{j=1}^{3i-2}\left(\frac{r_{\mathrm{S}}}{\mathrm{d}(\mathrm{CP},\mathrm{BS}(i,j))}\right)^{\delta} \geq 1. \tag{4.9}$$

For a path-loss exponent $\delta = 2$ a lower bound for equation 4.9 is

$$\begin{aligned} 3\sum_{i=1}^{\infty}\sum_{j=1}^{3i-2}\left(\frac{r_{\mathrm{S}}}{\mathrm{d}(\mathrm{CP},\mathrm{BS}(i,j))}\right)^{\delta} &\geq 3\sum_{i=1}^{\infty}\sum_{j=1}^{3i-2}\left(\frac{r\sqrt{3}}{\xi(3i-2)}\right)^{2} \\ &= 9{r_{\mathrm{S}}}^{2}\xi^{-2}\sum_{i=1}^{\infty}\sum_{j=1}^{3i-2}\left(\frac{1}{3i-2}\right)^{2} = 9{r_{\mathrm{S}}}^{2}\xi^{-2}\sum_{i=1}^{\infty}\frac{1}{3i-2} \end{aligned} \tag{4.10}$$

As the previous sum is unbounded (see harmonic series), this value will be larger than 1 for every value of ξ. Thus, for a path-loss exponent $\delta = 2$ the covered area and the gain factor are infinite.

For a path-loss exponent $\delta > 2$, however, an upper bound is

$$\begin{aligned} & 3\sum_{i=1}^{\infty}\sum_{j=1}^{3i-2}\left(\frac{r_{\mathrm{S}}}{\mathrm{d}(\mathrm{CP},\mathrm{BS}(i,j))}\right)^{\delta} \leq 3\sum_{i=1}^{\infty}\sum_{j=1}^{3i-2}\left(\frac{\sqrt{3}r_{\mathrm{S}}}{\xi(3i/2-1)}\right)^{\delta} \\ = \;& 2^{-\delta}3^{\delta/2+1}{r_{\mathrm{S}}}^{\delta}\xi^{-\delta}\sum_{i=1}^{\infty}\sum_{j=1}^{3i-2}\left(\frac{1}{3i-2}\right)^{\delta} \\ = \;& 2^{-\delta}3^{\delta/2+1}{r_{\mathrm{S}}}^{\delta}\xi^{-\delta}\sum_{i=1}^{\infty}\left(\frac{1}{3i-2}\right)^{\delta-1} \\ = \;& 2^{-\delta}3^{2-\delta/2}{r_{\mathrm{S}}}^{\delta}\xi^{-\delta}\sum_{i=1}^{\infty}\left(\frac{1}{i-2/3}\right)^{\delta-1} \end{aligned} \tag{4.11}$$

The sum is finite for any $\delta > 2$ according to the convergence of the Hurwitz zeta function [Ivi03] (also known as generalization of the Riemann zeta function) and, thus, the total received SNR will be less than T_{S} for a spacing that is large enough.

I determine the gain factor for the cell area of infinite cooperation by finding the largest spacing s that satisfies equation 4.9:

$$3\sum_{i=1}^{\infty}\sum_{j=1}^{3i-2}\left(\frac{\sqrt{3}\cdot r_{\mathrm{S}}}{\xi\sqrt{3\,j^2-9\,i\,j+9\,i^2-3\,i+1}}\right)^{\delta} \geq 1 \tag{4.12}$$

which is

$$\xi_{\mathrm{IC}}(\delta) = r_{\mathrm{S}}\left(3^{1+\delta/2}Z\right)^{1/\delta} \tag{4.13}$$

with

$$Z = \sum_{i=1}^{\infty}\sum_{j=1}^{3i-2}\left(3\,j^2-9\,i\,j+9\,i^2-3\,i+1\right)^{-\delta/2} \tag{4.14}$$

and results in a covered area A of

$$\mathrm{A}_{\mathrm{IC}}(\delta) = 3^{2/\delta+3/2}{r_{\mathrm{S}}}^2 Z^{2/\delta}/2. \tag{4.15}$$

$\mathrm{A}_{\mathrm{IC}}(\delta)$ is the following multiple of the non-cooperative cell area $\mathrm{A}_{\mathrm{NC}}(\delta) = 3^{3/2}{r_{\mathrm{S}}}^2/2$:

$$g_{\mathrm{IC}}(\delta) = 3^{2/\delta}Z^{2/\delta} \tag{4.16}$$

As no closed form for Z is evident, no closed form is evident for equations 4.15 and 4.16 either.

4.5. Results

The tables 4.1 and 4.2 summarize the results of the previous section. Figure 4.9 shows the gain that can be achieved with different types of cooperation for different values of

Table 4.1.: This shows the coverable area A with path-loss exponent $\delta = 2$ and range $r_S = 1$.

Type of cooperation	Coverable area A(2)	Approx.
No cooperation NC	$3^{3/2}/2$	2.598
Limited 2-cooperation LC	$3\sqrt{7}/4 + 5\sqrt{3}/4$	4.149
Unlimited 2-cooperation UC	$3^{3/2}$	5.196
Infinite cooperation IC	∞	∞

Table 4.2.: The gain factor g describes how much more area can be covered by use of cooperation for a path-loss exponent $\delta = 2$ and in general.

Coop. type	$g(2)$	$g(\delta)$
Limited 2-cooperation LC	$\frac{\sqrt{21}+5}{6}$	$\left(\sqrt{3}\sqrt{2^{2/\delta}-1} + 2^{2/\delta} + 1\right)/6$
Unlimited 2-cooperation UC	2	$2^{2/\delta}$
Infinite cooperation IC	∞	see equation 4.16

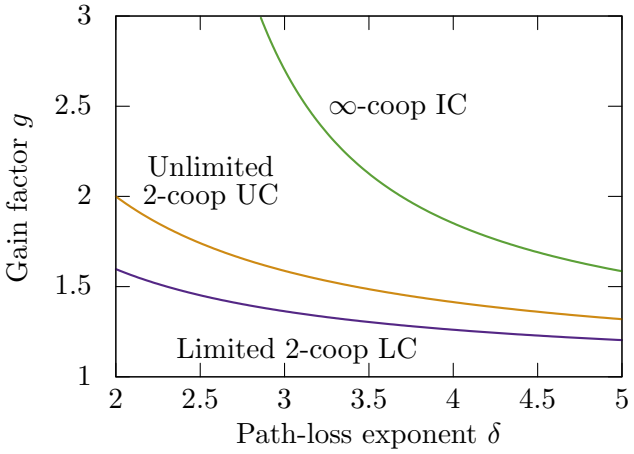

Figure 4.9.: The area that is possible to cover increases when using cooperation.

path-loss exponents: the lower the path-loss exponent, the greater the gain from the use of cooperation. The greater the covered area, the fewer BSs are needed to cover an area and, thus, also less power is consumed.

An alternative deployment that would work under my modeling assumptions is to place *two* BSs at each location and increase the spacing compared to the non-cooperative deployment. Next, I explain why this is not efficient. The two co-located BSs can provide signaling to every point inside their cooperative range together. Hence, for a path-loss exponent of 2 the spacing can be $\sqrt{2}$ times the spacing of non-cooperative deployment. This results in the same number of BSs as the non-cooperative deployment because two BSs are needed at every location. For higher path-loss exponents the co-located deployment is even worse. The reason for this is that the area which is already covered by a single BSs is covered two times with co-located BSs. This shows that it is more efficient to spread signaling BSs instead of co-locating them.

4.6. Deactivating opportunities with fixed BS spacing

In the previous section, I showed that it is possible to cover a larger area with cooperation when BSs are placed specifically for cooperation. In this section, I assume BSs are already placed to cover an area without cooperation. I analyze how cooperation allows some BSs to be deactivated while still covering the area. Again I consider ergodic capacity as the constraint for connectivity.

I create an optimization model with the goal to cover a given area with BSs that have already been placed in a hexagonal pattern. I assume the power consumption of BSs to be binary. Note that this implies that the power consumption does not depend on the network load. A more detailed model [ARFB10] would also have been an alternative, but because I only consider signaling traffic in this chapter this would yield the same results.

As I only consider signaling traffic in this chapter and the power consumption of the BSs is independent of their load, I do not model the capacity of the BSs. In my model the only constraint is to provide the signaling data rate D_{S} to every location of the area. To make sure every location in the scenario is connected I place a fine rectangular grid of UEs in the area to cover. I assume that when each of the UEs is connected, the complete area is covered.

To model cooperation I use the limited n-cooperation as introduced in section 4.4 because it is the more restrictive than unlimited cooperation and, thus, is able to achieve at least the same performance as limited cooperation.

I use the following constraints to calculate the necessary power to cover the given area. Equation 4.17 guarantees that each UE receives a signal high enough to be connected. Because I place the UEs in a fine grid on the area I want to cover, this approximates covering the complete area. I use equation 4.18 to allow only active BSs to transmit signal to the UEs and equation 4.19 to calculate the received signal at the UE: (1) a BS can provide the necessary signal for connectivity to a UE when its distance is smaller than r_{S}, (2) it provides a part of the signaling when further away, but still in cooperative

range, and (3) provides no signal at all, when further away. The optimization target is to minimize the number of active BSs and, thus, power consumption; it is expressed in term 4.20.

$$\forall m \in M : \sum_{b \in B} \mathrm{SNR}[m, b] \geq T_\mathrm{S} \tag{4.17}$$

$$\forall b \in B : \sum_{m \in M} \mathrm{SNR}[m, b] \leq x[b] \cdot |M| \cdot T_\mathrm{S} \tag{4.18}$$

$$\forall m \in M, b \in B : \\ \mathrm{SNR}[m, b] \leq \begin{cases} T_\mathrm{S}, & \text{if } \mathrm{d}(m, b) \leq r_\mathrm{S} \\ \frac{T_\mathrm{S} {r_\mathrm{S}}^\delta}{\mathrm{d}(m,b)^\delta}, & \text{if } r_\mathrm{S} < \mathrm{d}(m, b) \leq r_\mathrm{S} n^{1/\delta} \\ 0, & \text{if } r_\mathrm{S} n^{1/\delta} < \mathrm{d}(m, b) \end{cases} \tag{4.19}$$

$$\min \sum_{b \in B} x[b] \tag{4.20}$$

The optimization problem belongs to the mixed integer linear programming (MILP) class. Next I show how much power can be conserved in a RAN by using the GNU Linear Programming Kit (GLPK LP/MIP Solver, v4.44) and the Gurobi Optimizer 4.0 to solve different scenarios. The ranges in the following plots are the upper and lower bounds of the optimal solution the solver provides.

For all my analyses I consider a fixed rectangular area filled with 2500 UEs in a regular grid and 56 BS located in a hexagonal structure. Each of the 2500 UEs represents only a possible UE location and thus must be able to receive the signaling rate r_S. Figure 4.10 shows the BSs as well as the rectangular area I want to cover.

In real scenarios the ranges of the BSs are larger than strictly necessary to cover an area for different reasons: (1) to compensate for non-homogeneous signal propagation, (2) to allow soft hand-overs and (3) to compensate for misplacement of BSs, because they cannot always be placed in the ideal location. I call the factor by which the real ranges are larger than strictly necessary to cover the area *excess range*. I determine the number of BSs necessary to provide signaling traffic to the area for different values of excess range next.

Figure 4.11 shows that the higher the excess range in a scenario the more power can be conserved using cooperation. Also it shows that allowing up to two or three BSs to cooperate already reduces the power consumption. As I consider only a finite scenario the curves are not smooth and the larger bounds are higher for higher excess ranges as the complexity of MILP is higher. I conclude that cooperation allows BSs to be deactivated even for a low degree of cooperation. Also the gain increases with increasing excess range.

The range extension of BSs using cooperation is influenced by the path-loss exponent. Therefore, I analyze the effect of different path-loss exponents next. I assume the area the BSs can cover is twice as large as necessary to cover the area. That is, the excess

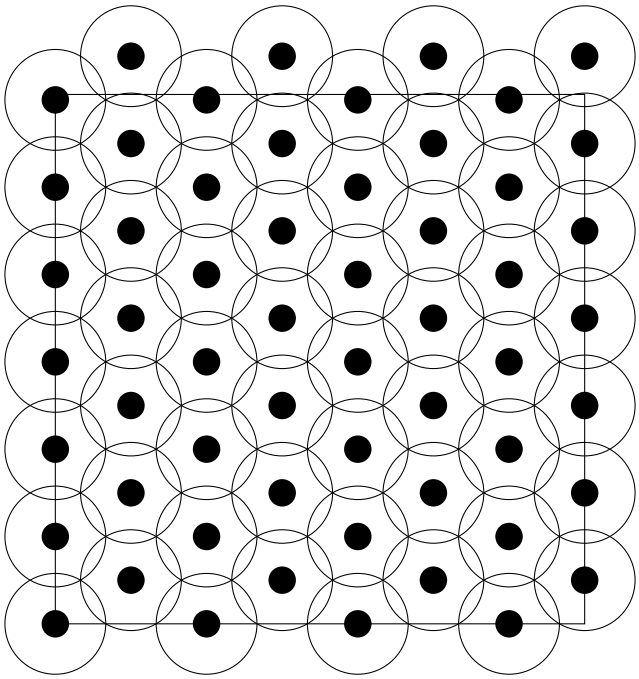

Figure 4.10.: My scenario contains 56 BSs and a rectangular area that the BSs have to cover.

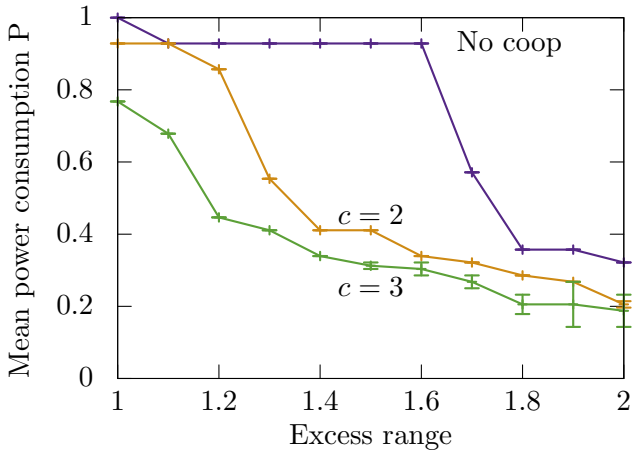

Figure 4.11.: The power necessary to cover the scenario cooperate depends excess range and the number of cooperating BSs c.

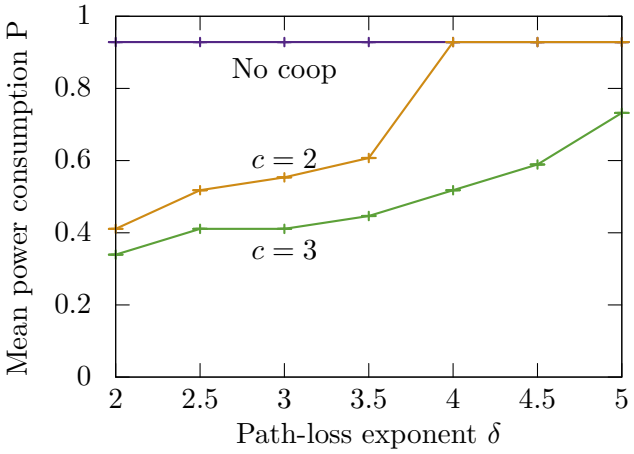

Figure 4.12.: The power necessary to cover the scenario depends on path-loss exponent δ and the number of cooperating BSs c.

range is $\sqrt{2}$ including the necessary adjustment for the different path-loss exponents. Figure 4.12 shows how the different path-loss exponents of the different scenarios influence the potential power savings by cooperation. The area around a BS that is too far away to provide enough power to make a direct reception possible, but still close enough to be usable for cooperation, is small when the path-loss exponent is high. Hence, I conclude that scenarios with a low path-loss exponent are better suited to conserve power by cooperation.

Figure 4.11 shows that cooperation of three BSs allows more power to be conserved than the cooperation of only two BSs. Additionally using three BSs to cooperate is more robust against higher path-loss exponents. How much power is conserved when even more BSs can cooperate is shown in figure 4.13. The power consumption decreases when the degree of cooperation is increased, but the gain becomes insignificant at about four in my scenario. I conclude that it is sufficient to let a few BSs cooperate to conserve power.

4.7. Comparing ergodic and outage capacity

As the channel changes over time, there are two different possibilities to assert its capacity to transfer data: ergodic capacity and outage capacity [MA05]. *Ergodic capacity* is the long-term average data rate that can be transmitted over the channel, assuming the channel state is always known. *Outage capacity* is the highest data rate that can be transmitted over the channel such that the probability for a transmission failure is lower than the *outage probability* O– formally, $\mathbb{P}[\text{SNR} > T] > 1 - O$.

In all earlier sections I considered a lower limit on *ergodic* capacity. In this section, I take a brief look at what happens when I change the requirement for a lower bound on

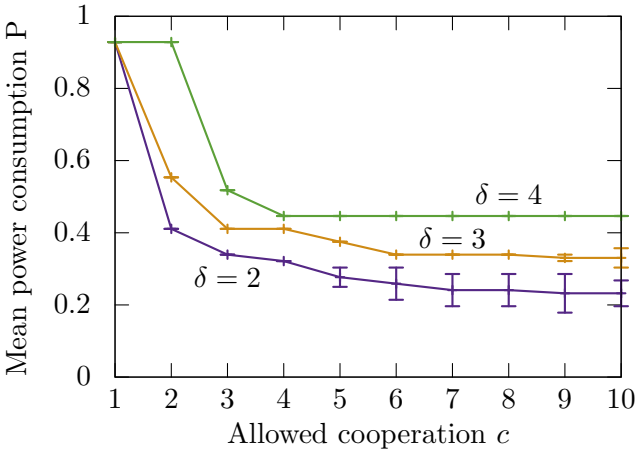

Figure 4.13.: The power necessary to cover the scenario depends on the degree of limited cooperation and path-loss exponent δ.

ergodic capacity to a lower bound on *outage* capacity. To compare both requirements I normalize them in such a way that they cover the same area without cooperation.

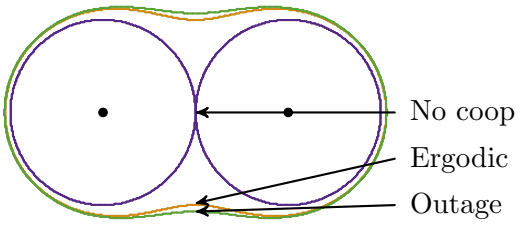

Figure 4.14.: The covered areas differ when defining connectivity by ergodic and by outage capacity thresholds.

Figure 4.14 illustrates the difference between covered area when considering ergodic capacity and outage capacity and unlimited cooperation is used: It shows only a small difference. This provides confidence that the results of my analysis for ergodic capacity will provide similar results when outage capacity is considered.

Figure 4.15 shows the area covered without cooperation, with selection combining, and with MRC. This figure illustrates that the gain ofMRC instead of selection combining is largest when two signals are nearly of equal SNR. Note that the areas also depend on the whether the needed data rate is defined by outage or ergodic capacity.

This is a hint that the results I calculated in sections 4.4 and 4.6 will also be valid when considering outage capacity instead of ergodic capacity.

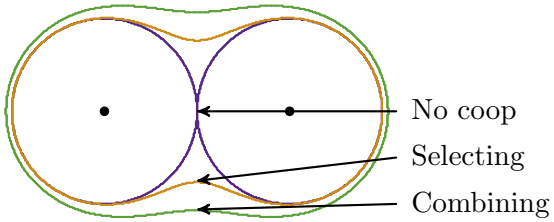

Figure 4.15.: The covered areas depend on the combination method of the UE (selection combining and maximal-ratio combining (MRC)).

4.8. Conclusion

Firstly, I analytically showed how cooperation increases the covered area of a BS. Secondly, I created a model to calculate the power necessary to cover an area with given BSs and analyzed how cooperation reduces the power consumption. My analysis shows that using a few BSs to cooperate on the transmission to each UE is enough to conserve significant amounts of power. Both the analytical and the optimization model show that scenarios with a low path-loss exponent are better suited to conserve power by cooperation.

In the future this has to be verified for more complex channel models and a more detailed form of cooperation to be sure it can be applied in reality. In addition, the necessary protocols to activate and deactivate BSs as well as their cooperation need to be developed.

In this chapter, I quantified that fewer BSs are necessary when cooperation is used to transmit signaling traffic to a given area and how many BSs can be deactivated when already placed BSs can cooperate. Owing to my definition of signaling traffic this is independent of actual user activity. In contrast to this, the BSs activity I determine in the next chapter depends on the user activity to serve data traffic.

5. Conserving energy in data transmissions

In contrast to the previous chapter, the requirement in this chapter is not to provide signaling traffic to every possible location, but to provide data traffic to every active user equipment (UE). Because the requests for data traffic arise randomly in time and space it is not necessary that all base stations (BSs) are always active.

The idea of this chapter is to activate BSs only when necessary to serve data traffic to UEs. In chapter 2 I already analyzed how a single BS can do this when considering power-cycle durations. In this chapter, I do not consider power-cycle durations, but focus on the interaction of neighboring BSs via the areas that both can serve.

The configuration of BSs I determined by optimization at the end of the last chapter describes a configuration that can be used until the load of the network changes. In contrast to this, the configurations I determine in this chapter are determined to provide data traffic to the UEs that are currently active. Therefore, the configurations determined in this chapter change faster, and, thus, I am only interested in their average power consumption.

Just as with signaling, using cooperative transmissions from BSs allows UEs to be reached which are not in range of any single BS. In this chapter, I quantify the reduction of energy consumption when BSs can cooperate compared to not cooperating to provide data traffic to all active UEs.

5.1. Introduction

One way to reduce the energy consumption of the BSs is to decrease the fraction of time they are active. Because all active UEs still have to be covered by active BSs, extending the data range of the BSs reduces the necessary activity of neighboring BSs. In this chapter, I consider the same cooperative transmission used in the previous chapter but to transfer data traffic instead of signaling traffic.

I use the notion of activity probability as a proxy for power consumption in this chapter. The *activity probability* is the fraction of time a BS is serving data traffic to UEs. If the power-cycle duration is low and the BSs have binary power profile, the activity probability is equal to the mean power consumption. One possibility to include the power-cycle durations is to determine the energy consumption from the activity probability analogous to calculations in chapter 2.

To determine the activity probability of data BSs most related work uses simulations. In contrast to this, I determine it analytically. To be able to do this, I assume the BSs are placed in a hexagonal deployment and have a circular radius in which they can provide the requested data rate to the UE. While this is not a realistic model, I only

need this model to determine the *sizes* of overlapping areas and do not consider their actual *shape*. This allows me to contribute an analytic description of the possible gain in energy efficiency when enabling cooperation in a radio access network (RAN).

I determine an approximation of the activity probability. The same method I use can be used to determine upper and lower bounds on the activity probability, but for these bounds to be practically relevant the number of considered BSs has to be impractically high. Also note that the activity probability I determine is only an approximation of the actual value as I ignore some dependencies between probabilities to be able to determine the results analytically. I do not determine the quality of my approximation as the complexity to do this would be comparable to evaluating the upper and lower bounds.

5.2. Related work

Ashraf et al. [ABH11] provide an overview of possible strategies to put small BSs into sleep mode and activate them again when necessary. Ismail and Zhuang [Ism11] describe how energy can be saved on different levels of the network. While they compare the different approaches and provide overviews, I analytically determine the gain which is not specific to any implementation. Other work [BK97, Blu10] describes how to cover an area redundantly.

Wan et al. [WXW11] provide a polynomial-time approximation scheme (PTAS) for the nondeterministic polynomial time (NP)-hard problem of minimum wireless coverage (MWC). MWC seeks to find the minimum number of disks to cover a set of UEs. While this optimally solves a single instance of my problem, I determine the expected number of active BSs, when the UEs are distributed according to a spatial Poisson process. The proof of its NP-hardness is by a transformation from planar 3SAT [Joh82]. While a PTAS exists no *fully* polynomial-time approximation scheme (FPTAS) is known [WXW11].

Hohenberger determines the effects cooperation and queuing on shared area of BSs in his bachelor's thesis [Hoh12]. Based on it we published a paper [HHK13] which compares strategies to assign the UEs to BSs which take queuing effects into account. I do not consider queuing effects in this chapter because I consider a low-load scenario in which these effects are unimportant.

Another group of related work uses stochastic geometry [BB09] to determine outage probabilities, data rates and power consumption under the assumption that both UEs and BSs are placed by Poisson processes. Suryaprakash et al. [SFdSF12], for example, prove that putting BSs into sleep mode is more energy-efficient than varying the available bandwidth. I assume BSs to be placed in a hexagonal grid instead of by a Poisson process and determine activity probabilities of BSs.

Son et al. [SKYK11] and Zhou et al. [ZGY+09] describe heuristics to activate BSs and associate the UEs to BSs. Vereecken et al. [VDC+12] consider a network of macro and femto BSs. In contrast to my analytical results, they provide results by simulations.

Richter et al. [RFF09] determine the optimal number of pico BSs per macro BS by comparing different deployments. Alternatively, the problem can be formulated as an optimization problem. Gonzalez-Brevis et al. [GBGF11] provide an example for such an

optimization.

Conte et al. [CFC11] describe how changing cell sizes and deactivating BSs can conserve energy. Auer et al. [AGD+11, AGG11] focus on detailed power and deployment models and determine their results using simulations. In contrast to this, I use a simpler power and signal model but determine my results analytically.

Marsan et al. [MC09] as well as Oh and Krishnamachari [OK10] describe how the changes in daily load can be approximated and determine the reduction in energy consumption that is possible by adapting to the changes. In contrast to the work I describe in this chapter they do not consider the effects of overlapping areas and the random occurrence of user requests.

5.3. Model

My model consists of an infinite plane with BSs placed in a hexagonal grid, because this is optimal to cover the plane [Ker39]. I assume that a higher layer of signaling BSs exists, which detects the presence and activity of UEs. In the previous chapter I described how these signaling BSs can be placed.

In this chapter, I model only the pico BSs, which provide data to the UEs and do not cover the area with signaling traffic. I assume the macro BSs do not provide data traffic at all and thus the pico BSs need to be able to reach every location. Because pico BSs are usually omnidirectional, I do not consider any directional properties of the antennas. I model only a single moment in time and not the progression of time in this chapter.

The plane is populated with UEs by a 2-dimensional spatial Poisson process with a density λ. Because I assume the UEs to be distributed by a Poisson process, the expected activity $\mathrm{EA}(\mathrm{A}) = 1 - \mathrm{e}^{-\mathrm{A}\lambda}$ is the probability that at least 1 UE is in an area (I call this "activity in the area") of size A.

Each UE must be associated to a single active BS or to two cooperating active BSs.

5.3.1. Cooperation

In this chapter, I consider only limited 2-cooperation because the areas of overlap it creates are simple geometric shapes. Additionally, using unlimited cooperation or a higher degree of cooperation is superior to using limited 2-cooperation. Therefore, the activity probability determined by limited 2-cooperation is an upper bound for the activity probability (and, thus, power consumption) using unlimited cooperation. In this chapter, I assume the path-loss exponent δ to be equal to 2 for simplicity. This results in a cooperative range which is $\sqrt{2}$ times the non-cooperative range. The calculations I do in this chapter are also possible for other path-loss exponents.

For the purpose of this chapter, I consider only the power consumption of the BSs during periods of low load. As a result of this, I assume that every BS is always able to handle the requested load. This allows me to ignore any limits on the number of UEs that are associated to a BS. Depending on the type of traffic this might result in longer transmission (e.g., file transfers) or not (e.g., video stream). When a transmission

takes longer it will stay active longer, and will thus increase the energy consumption. I consider it in chapter 7.

5.3.2. Metrics

I consider two different metrics of power consumption: (1) the activity probability of a BS P and (2) the expected number Q of active BSs per covered area. To fairly compare scenarios with different spacing, I consider both activity probability of a single BS and the expected number of active BSs *per area*. Note that the mean power consumption of a BS is equal to the activity probability under the binary power model and I, thus, use the same symbol.

I denote all functions used with an index "n" for non-cooperative transmission. Because the use of cooperative transmissions allows greater spacing (see chapter 4) I distinguish two different ranges of spacing under cooperation. To directly compare cooperative assignments with non-cooperative assignment I use the same spacing and denote a cooperative assignment as "d" to denote densely placed BSs. As cooperative transmissions allow for a higher spacing I use "s" for sparsely placed BSs with cooperation.

5.4. Analysis

In this section, I determine the probability that a BS has to be active when all UEs have to be covered. The parameters I analyze are the spacing between the BSs and the different transmission technologies. This allows me to determine the gain of using cooperation in a network which is built for non-cooperative transmission and a network that is built for cooperative use.

To determine the activity probability, I divide the plane into areas that summarize all UEs which can be connected to the same set of BSs. From the size of these areas I determine the probability that active UEs are in the area and from that the probability that a BS has to be active.

I first determine the activity probability for spacings which are able to provide service to every location without cooperation. This reflects a RAN which was built without cooperative transmissions in mind. Later I also consider higher spacings which become possible when the network is built for the use of cooperation (see chapter 4). Therefore, I do not need to determine the activity probability for all possible spacings. Determining the results for other spacings would be merely routine work.

5.4.1. No cooperation

Using non-cooperative transmission, each UE has to be within 1 unit distance of an active BS. Because I am interested in the activity probability of a BS I look at the UE from the perspective of a BS.

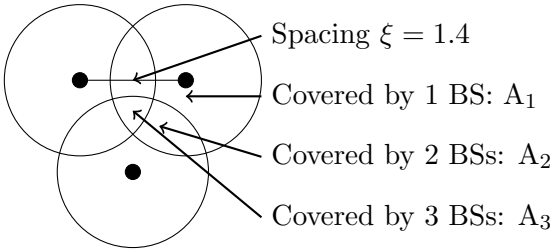

Figure 5.1.: The areas of non-cooperative transmissions with spacing $\xi_{\min} \leq s \leq \xi_{\max}$ overlap.

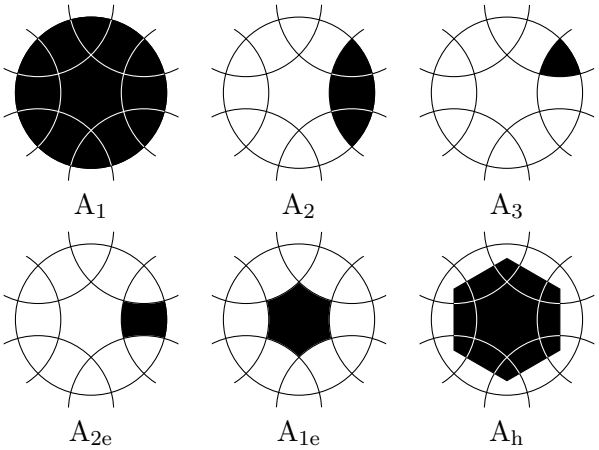

Figure 5.2.: The areas and the names I assign them for non-cooperative transmissions are shown here for a spacing of $\xi = 1.4$.

Areas

In this section, I describe the size of the areas which arise from using non-cooperative transmission. The largest spacing of non-cooperative BSs that still covers the complete plane is $\xi_{\max} = \sqrt{3}$. In general small spacings are not practical because a high number of BSs would be needed to cover a given area. Therefore, arbitrarily small spacings are not practical. I consider $\xi_{\min} = 2/\sqrt{3}$ as the smallest spacing. The reason I use *exactly* this spacing is that further reducing the spacing would create areas of 4 overlapping BSs.

I denote sizes of areas by A and add an index for the number of BSs n that can cover it. I additionally add the label e when the area can be covered by *exactly* n BSs. As a reference, the area a BS covers when each UE is assigned to the closest BS is $\mathrm{A_h}$. Figure 5.1 gives an overview which types of overlapping areas occur and figure 5.2 highlights the areas together with their name.

The size of the area that can be covered by 1 BS is:

$$\mathrm{A}_1 = \pi, \tag{5.1}$$

The size of an area covered by two BSs is:

$$\mathrm{A}_2 = \mathrm{AI}(\xi, 1, 1) = 2 \arccos\left(\frac{\xi}{2}\right) - \frac{\xi\sqrt{2-\xi}\sqrt{\xi+2}}{2}, \tag{5.2}$$

where the size of the intersection of two circles with distance d and radii r and R [Weia] is

$$\begin{aligned}\mathrm{AI}(d, r, R) =& r^2 \arccos\left(\frac{d^2 + r^2 - R^2}{2dr}\right) + R^2 \arccos\left(\frac{d^2 + R^2 - r^2}{2dR}\right) \\ &- \frac{1}{2}\sqrt{(-d+r+R)(d+r-R)(d-r+R)(d+r+R)}.\end{aligned} \tag{5.3}$$

The overlap of 3 BSs is an equilateral circular triangle [Few06] with area:

$$\mathrm{A}_3 = \frac{\sqrt{3}}{4}c^2 + 3\left(\arcsin\left(\frac{c}{2}\right) - \frac{c}{4}\sqrt{4-c^2}\right), \tag{5.4}$$

where c is the distance between two corners of the equilateral circular triangle, which obeys:

$$c^2 = 3 - \frac{\xi^2}{2} - \xi\sqrt{3 - 3\xi^2/4}. \tag{5.5}$$

This results in the size of the areas that can be covered by *exactly* 2 BSs to be:

$$\mathrm{A}_{2\mathrm{e}} = \mathrm{A}_2 - 2\mathrm{A}_3, \tag{5.6}$$

and the area that can be covered by *exactly* one BS:

$$\mathrm{A}_{1\mathrm{e}} = \mathrm{A}_1 - 6\mathrm{A}_{2\mathrm{e}} - 6\mathrm{A}_3. \tag{5.7}$$

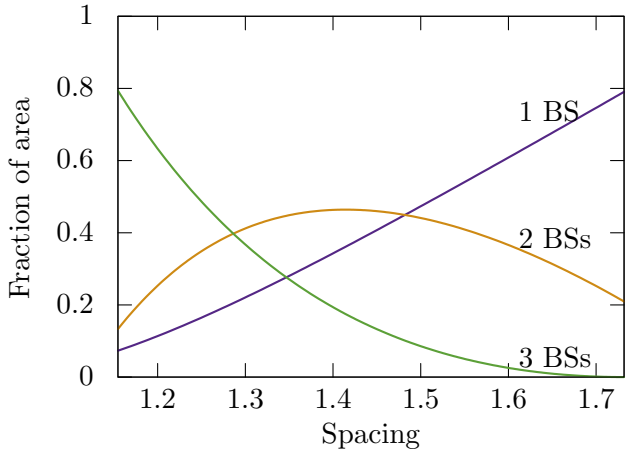

Figure 5.3.: The fraction of the area of the plane that is covered by a given number of BSs depends on the spacing ξ.

To calculate a reference value and to determine the number of BSs per area I additionally need the hexagonal Voronoi area that is closest to a BS:

$$\mathrm{A_h} = \sqrt{3}\xi^2/2. \tag{5.8}$$

Hence, there are $1/\mathrm{A_h}$ BSs per unit area.

Figure 5.3 illustrates which fraction of the plane is covered by *exactly* 1, 2, and 3 BSs. Note that this is different from the fractions of area inside the range of a single BS. The reason for this is that areas $\mathrm{A_{2e}}$ are seen by two BSs and areas $\mathrm{A_3}$ are seen by 3 BSs and are, thus, counted two and three times, respectively, when considering areas each BS "sees".

Activity probability

In this section, I determine (1) the activity probability of a BS and (2) the expected number of active BSs per area for non-cooperative BSs.

As a reference value against which to compare the energy-efficient assignment of UEs I first determine the activity probability when each UE is assigned to its closest BS. When each UE is assigned to its closest BS, a BS has to be active if and only if there is activity in $\mathrm{A_h}$. Hence, the activity probability of a BS under closest assignment is:

$$\mathrm{P_h} = \mathrm{EA}\left(\mathrm{A_h}\right). \tag{5.9}$$

For any activity probability P of a BS the expected number of active BSs per area is:

$$\mathrm{Q} = \mathrm{P}/\mathrm{A_h}. \tag{5.10}$$

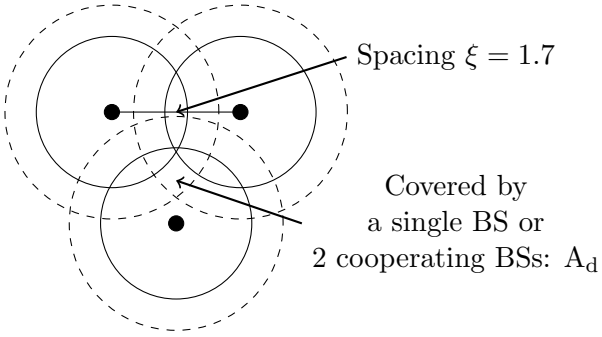

Figure 5.4.: In a dense cooperative deployment, areas can be covered by a single BS or two cooperating BSs. This is valid for spacings $\xi_{\mathrm{minD}} \leq \xi \leq \xi_{\mathrm{maxD}}$.

Determining the exact activity probability of a BS for an energy-efficient assignment is not directly possible because it depends on the activity probability of its neighbors, which in turn depends on the activity probability of their neighbors and so on. Because my scenario is symmetric with respect to all BSs, the activity probability P of all BSs is the same as long as the scheme for assignment of UE considers all BSs equally. Next, I determine an equation for the activity probability of a BS depending on the activity probability of its neighbors. Because all these probabilities are equal assuming a homogenous Poisson process, I solve the equation for it.

Consider a BS B: it has to cover at least its central area A_{1e}. Additionally, consider a neighbor N: If N is active (activity probability for non-cooperative assignment P_n), B can ignore the area which is shared between both B and N (size A_{2e}). If N is not active (probability $1-P_n$), the area has to be covered by B if there is activity. The same argument works for the area A_3: If neither of the two neighbors that also cover the area are active, B has to cover it in case of activity. These arguments allow me to calculate the activity probability P_n of BS B, based on the activity probability of its neighboring BSs. I do not need to recursively continue because from the symmetry of the scenario all BSs have the same activity probability.

Note that using this approach, I ignore the dependencies between activity of neighboring BSs and dependencies between probability of activity in an area and one of its neighboring BSs. For example, if I know that neighbor N is active, the activity probability of B will be slightly lower than P_n because BS N already covers the users in the area shared by B and N. Hence, I only estimate the activity probability and not the exact value.

When the overlapping areas (A_2 and A_3) are small the error done ignoring the dependencies is small. When the size of the overlapping areas is greater the potential for error is greater, but my approximation does not necessarily have to be worse. Because deploying BSs at their maximum range is generally more efficient (as fewer BSs are needed), I consider only spacing close to the highest possible spacing here. For these high spacings the error of my approximation is limited as the size of the overlapping areas is limited.

To determine the activity probability, I encode the activity configurations of the neighbors in the binary digits of variable i: 0 means deactivated and 1 means active. I denote the jth binary digit of i as $i[j] = \lfloor i/2^j \rfloor \bmod 2$. I iterate over all possible configurations of activity of the neighbors and determine the activity probability of BS B for each configurations by computing the size of the area that it has to cover. This results in the following equation:

$$\mathrm{P_n} = \sum_{i=0}^{2^6-1} \prod_{j=0}^{5} \overbrace{i[j]\mathrm{P_n} + \big(1 - i[j]\big)\big(1 - \mathrm{P_n}\big)}^{\text{Probability that neighbor } j \text{ is in state } i[j]} \tag{5.11}$$

$$\cdot \mathrm{EA}\left(\mathrm{A_{1e}} + \mathrm{A_{2e}} \underbrace{\left(\sum_{j=0}^{5} \big(1 - i[j]\big) \right)}_{\text{Number of } \mathrm{A_{2e}} \text{ areas to cover}} + \mathrm{A_3} \underbrace{\left(\sum_{j=0}^{5} \big(1 - i[j]\big)\big(1 - i[j+1 \bmod 6]\big) \right)}_{\text{Number of } \mathrm{A_3} \text{ areas to cover}} \right).$$

Note that simply using the binomial formula to determine the probability for a given number of active neighbors is enough to determine the number of areas $\mathrm{A_{2e}}$ that have to be covered. But this approach does not work for the areas $\mathrm{A_3}$ because not only the number of active neighbors determines the number of areas $\mathrm{A_3}$, but also their relative position has to be taken into account.

Equation 5.11 can be simplified to a polynomial of degree 6 in $\mathrm{P_n}$. While this is too complex for a closed form solution, Newton's method provides me with the means to determine its solution numerically. This solution is the activity probability $\mathrm{P_n}$ of a single BS for the scenario without cooperative transmissions.

While the method just explained determines the same activity probability for each BS, this does not necessarily result in the *minimal* activity probability, when averaged over all BSs because I did not prove that an asymmetric activity pattern does not result in a lower average activity probability. I propose that every asymmetric scheme can be converted to a symmetric scheme by randomly selecting variants of the asymmetric scheme which have been translated or mirrored. But as this is not in the scope of my dissertation I did not pursue this idea. A good starting point to follow up on this idea could be the mathematics of tilings and colorings [Soi08].

5.4.2. Cooperation with densely placed BSs

To compare the activity probability of two association schemes they must have the same spacing. Because cooperative transmissions allow a higher spacing, I call the deployment with the same spacing as the non-cooperative deployment *dense*.

Areas

Figure 5.4 provides an overview of the involved areas when the spacing is between $\xi_{\mathrm{minD}} = \sqrt{8/3}$ and $\xi_{\mathrm{maxD}} = \xi_{\max} = \sqrt{3}$. Figure 5.5 highlights the relevant areas and provides

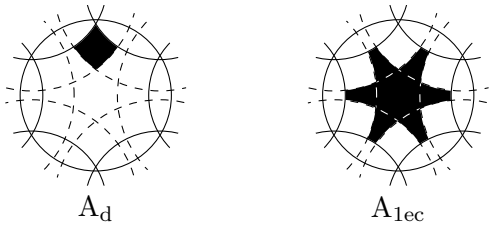

Figure 5.5.: These are the shapes and names of regions for cooperative BSs with dense deployment of spacing $\xi = 1.7$.

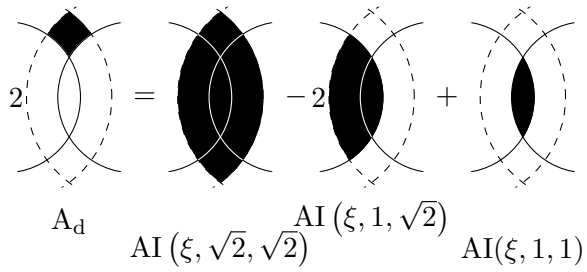

Figure 5.6.: The area A_d can be calculated from overlapping areas of different sizes.

their names. Note that I only consider limited 2-cooperation for dense cooperative deployments due to the more complex shape of areas with unlimited cooperation.

I call the area that can be covered by exactly two cooperating BSs A and B or one other fixed BS C ($A \neq B \neq C$) A_d. It can be calculated from of intersection of circles. Figure 5.6 shows that A_d can be calculated as:

$$A_d = \frac{1}{2}\left(AI\left(\xi, \sqrt{2}, \sqrt{2}\right) - 2AI\left(\xi, 1, \sqrt{2}\right) + AI\left(\xi, 1, 1\right)\right). \tag{5.12}$$

The area that is only covered by a single BS has size:

$$A_{1ec} = A_{1e} - 6A_d. \tag{5.13}$$

Activity probability

The equation for the activity probability with limited cooperation can be determined in the same way the one as without cooperation by additionally considering the area $\mathrm{A_d}$::

$$\begin{aligned} \mathrm{P_d} = \sum_{i=0}^{2^6-1} \prod_{j=0}^{5} i[j]\mathrm{P_n} + \big(1 - i[j]\big)\big(1 - \mathrm{P_n}\big) \\ \mathrm{EA}\Bigg(\mathrm{A_{1ec}} \\ + \mathrm{A_{2e}} \left(\sum_{j=0}^{5} \big(1 - i[j]\big) \right) \\ + \mathrm{A_3} \left(\sum_{j=0}^{5} \big(1 - i[j]\big)\big(1 - i[j+1 \bmod 6]\big) \right) \\ + \underbrace{\mathrm{A_d} \left(\sum_{j=0}^{5} 1 - i[j]i[j+1 \bmod 6] \right)}_{\text{Number of cooperative areas } \mathrm{A_d} \text{ to cover}} \Bigg) . \end{aligned} \tag{5.14}$$

5.4.3. Cooperation with sparsely placed BSs

In addition to the dense deployment, cooperating BSs can also be placed *sparsely*. The reason why this deployment allows all UEs in the plane to be served is the same as already explained in section 4.4.

Areas

The geometric illustration in figure 5.7 shows the areas I consider for cooperative transmissions with a minimum spacing of $\xi_{\mathrm{minS}} = 2$. This lower limit prevents overlap of non-cooperative regions and thus simplifies the calculations. The maximum spacing for limited cooperation is

$$\xi_{\mathrm{maxSL}} = \sqrt{\sqrt{21} + 5} \Big/ \sqrt{2}. \tag{5.15}$$

The same scheme explained here also works for unlimited cooperation with a spacing up to $\xi_{\mathrm{maxSU}} = \sqrt{6}$ (see chapter 4).

The way I determine the sizes of the areas squander some potential to reduce the activity probability because the areas that can actually be reached are larger (compare the size of covered areas by limited and unlimited cooperation in figure 4.7). The benefit of my method is that I can use the same expressions of sparse area sizes for limited and unlimited cooperation and it holds even as I increase the spacing up to ξ_{maxSU}. Note that I squander even more potential for unlimited cooperation than for limited cooperation.

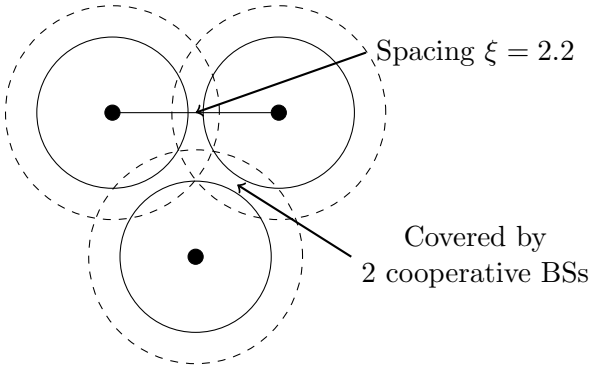

Figure 5.7.: These are the areas in a sparse cooperative deployment (with spacing $\xi_{\mathrm{minS}} \leq \xi \leq \xi_{\mathrm{maxSU}}$ for unlimited cooperation and $\xi \leq \xi_{\mathrm{maxSL}}$ for limited cooperation).

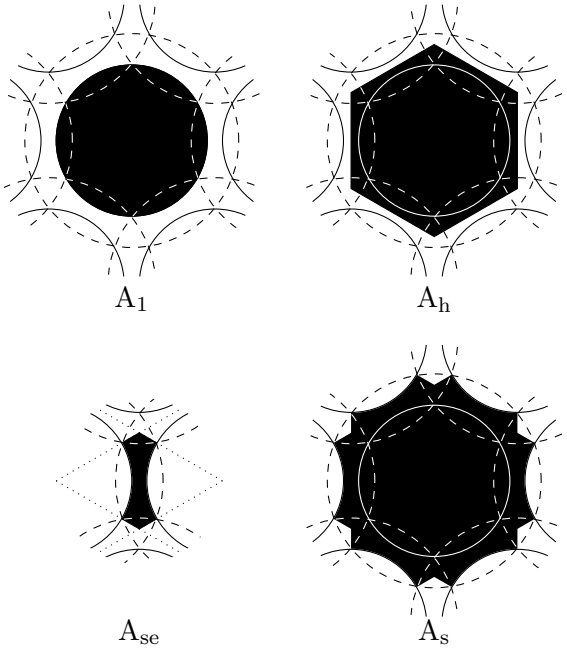

Figure 5.8.: These are the shapes and names of regions for sparse cooperative deployment of BSs with spacing of $\xi = 2.2$.

A lower bound for the size of the area that is covered by exactly two BSs, which is highlighted in figure 5.8, is:

$$A_{se} = \frac{A_h - A_1}{3}. \tag{5.16}$$

Hence, total area for which a BS has to be active is:

$$A_s = A_1 + 6A_{se}. \tag{5.17}$$

Activity probability

The activity of a BS can be directly determined from the activity in the areas, because no association choices have to be made. The activity probability is:

$$P_s = EA(A_s). \tag{5.18}$$

5.5. Numerical results

In this section, I illustrate the analytical results from the previous section with the help of the computer algebra system Maxima 5.28.0.

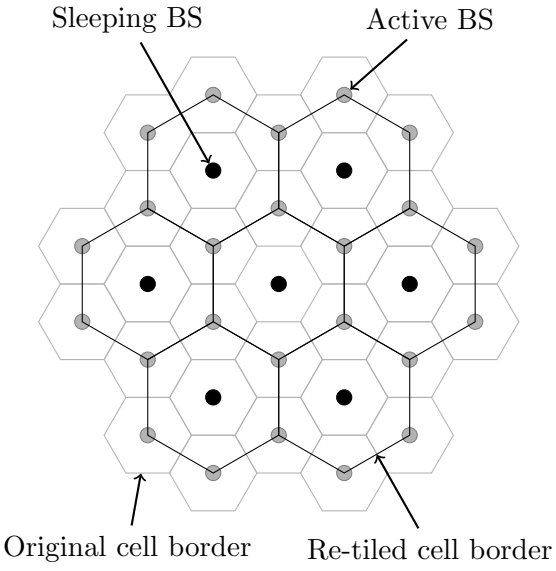

Figure 5.9.: The original hexagonal tiling can be re-tiled to new hexagonal tiling with a $\sqrt{3}$ larger spacing of BSs by deactivating 2 of 3 BSs.

While it is impractical to move BSs to adapt the parameters of a RAN to the current load, it is possible to transform a hexagonal deployment of BSs to another hexagonal deployment with a different spacing simply by removing BSs. For the purpose of energy adaption this just means deactivating and ignoring the removed BSs. Examples for

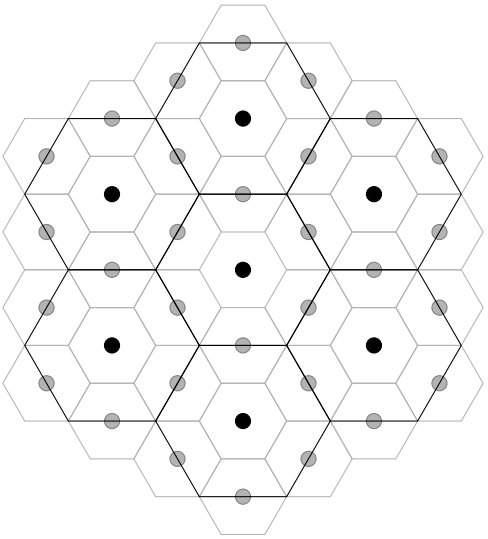

Figure 5.10.: An alternative re-tiling with the double spacing can be achieved by deactivating 3 of 4 BSs.

possible re-tilings of the hexagonal deployments are: (1) removing 2 of 3 BSs results in a hexagonal tiling with a spacing that is larger by a factor of $\sqrt{3}$ (figure 5.9) and (2) removing 3 of 4 BSs results in a hexagonal tiling with a spacing that is larger by a factor of 2 (figure 5.10). While other re-tilings are possible [BSW97], these have higher spacings. These examples show that a hexagonal deployment can be transformed into another hexagonal deployment with higher spacing by deactivating BSs. The factors between spacings for these examples are $\sqrt{3}$ and 2. Note that because each BS has a capacity limit, re-tiling is only possible when the total traffic in the network is low.

Figure 5.11 shows how cooperative schemes reduce the activity probability for a given spacing. Because mean power consumption not only depends on the activity probability, but also on the *number* of placed BSs, this is not a fair comparison. The expected number of active BSs per area Q considers this. Figure 5.12 illustrates how cooperative schemes reduce the expected active BSs per area Q. Non-cooperative schemes have a maximum spacing of $\xi_{\max} = \sqrt{3}$. Cooperative schemes can cover the plane for spacings between $\sqrt{3}$ and 2,but for my analysis it was not necessary to determine the activity probability in this range. I conclude that it is more energy-efficient to operate BSs at a higher spacing, as long as they are still able to serve all UEs with the necessary data rates.

Figure 5.13 shows that the optimal selection of a BS activation scheme depends on the user density. When the user density increases, it converges to the number of BSs per area, because all BSs have to be active in this case.

Figure 5.14 shows the expected number of active BSs per area compared to the non-cooperative assignment. The dense cooperating scheme is always lower than 1 and, hence, better than the non-cooperative assignment. The sparse cooperative assignment

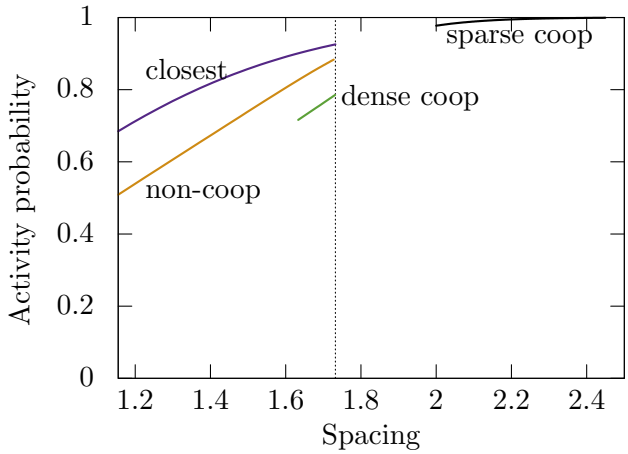

Figure 5.11.: The activity probability of a BS increases, when the spacing is increased.

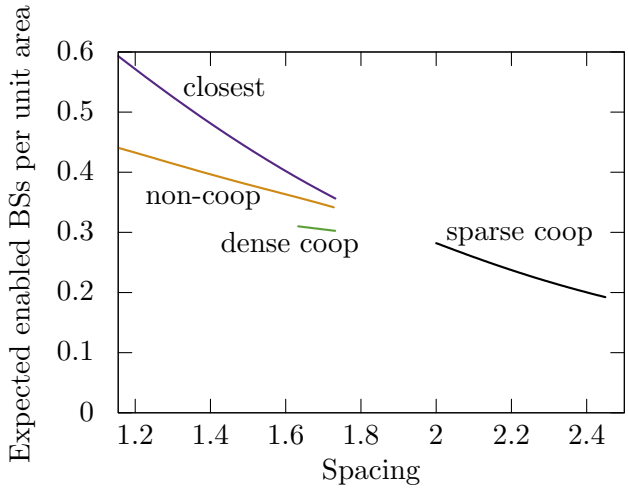

Figure 5.12.: The expected number of active BSs per area decreases when the spacing is increased. The arrow shows a change that can be implemented by deactivating BSs in the network (re-tiling) instead of physically moving BSs.

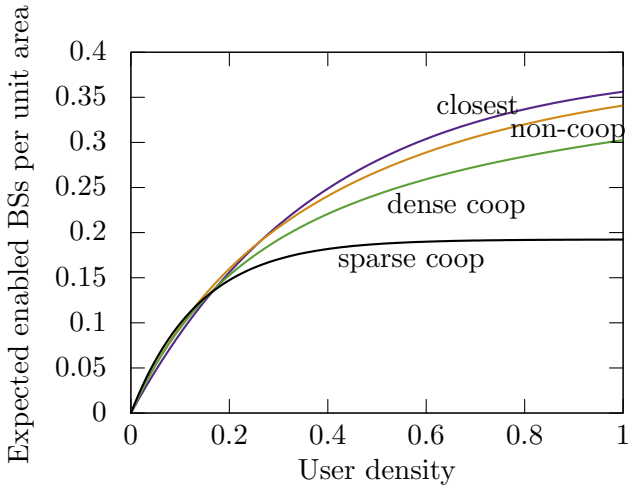

Figure 5.13.: The expected number of active BSs asymptotically approaches a limit for each of the schemes when the user density is increased. In this case all schemes use their optimal spacing.

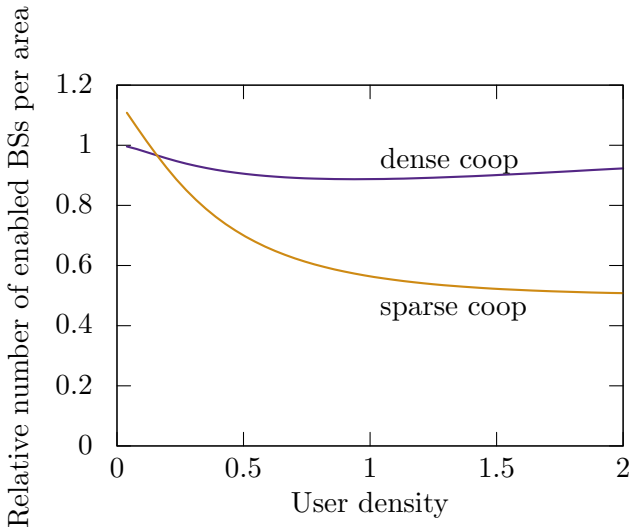

Figure 5.14.: The expected number of active BSs per area for the cooperative schemes at their maximum spacings compared to the non-cooperative scheme. Note that the dense cooperative scheme uses the same spacing and thus the same BSs.

on the contrary has more expected active BSs because a single UE can potentially require activating 2 BSs to cooperatively provide it with data. However, when the user density increases, the sparse deployment becomes more efficient because it uses fewer BSs. Note that this analysis only holds as long as the sparse cooperative BSs can provide enough data rate to all users.

5.6. Conclusion

I first determined the sizes of intersections of ranges for a hexagonal deployment of BSs in a RAN. I then calculated the probability that at least one UE is in such an area under the assumption that the UEs are distributed according to a Poisson process. Based on this I calculated the activity probability of the BSs for cooperative and non-cooperative association schemes. Using this result, I determined the expected number of active BSs per area.

My results show that the reduction in power consumption by allowing cooperative transmissions without changing the spacing of the BSs is between 0% and 11% depending on the load. When additionally changing the spacing the power consumption per area can be reduced by additional 39%. I concluded that fewest active BSs are needed if the BSs are placed as far apart as possible, but still cover the area and are able to provide the necessary data rates. Cooperative transmission from the BSs allows the BSs to be placed further apart than non-cooperative transmission would allow.

Future work will include the limit of data rates BSs can provide and, hence, determine when the assignment schemes will be data rate-limited instead of coverage-limited.

6. Radiated power

In the previous two chapters, I determined how base stations (BSs) can cooperate based on the distance to user equipments (UEs). The distance (together with other factors) determines only the *average* channel quality. In this chapter, I demonstrate how knowledge about the *instantaneous* channel gain can reduce the outage probability and the radiated power. Because the radiated power becomes interference at other UEs it is important to keep it as low as possible, especially in dense urban environments.

To reduce outage probability and radiated power, I compare strategies which select which BSs cooperate based on average and instantaneous channel gains. During the analysis it also becomes clear why BSs selection and transmit power control produce similar results.

6.1. Introduction

An important decision in cooperative radio access networks (RANs) is the selection of the cooperating BSs from all possible BSs. This selection can be based on average or instantaneous channel quality. Basing this decision on instantaneous knowledge cannot be worse and it is reasonable to assume it can be better. Related work has quantified the overhead and the gain of instantaneous knowledge. But most of this work is based on simulations, while I analytically compare BS selection with instantaneous and average channel knowledge. In contrast to other work, I do not consider distances, but directly base my decisions on channel quality, which is also more realistic to know in a real system.

I differentiate between two uses of instantaneous channel knowledge: (1) instantaneously changing the effective radiated power (ERP) of BSs and (2) selecting the BSs that cooperate to transmit the signal to the UE. I ignore the technical requirements (see section 1.4) needed to cooperate and compare the outage probability and the mean ERP. I use ERP because it is proportional to the interference generated at other UEs. The only assumption I make about timing is that BSs which use instantaneous channel knowledge need to be able to track it fast enough.

The contribution of this chapter is the analytic description of outage probability and ERP for cooperating BSs in a RAN with and without instantaneous channel knowledge. Additionally, I show that the highest gain from instantaneous channel knowledge can be achieved when the possibly cooperating BSs have similar average channel gains.

I show that instantaneous selection of cooperating BSs can reduce the ERP while keeping the outage probability the same.

6.2. Related work

Tse and Viswanath [TV05] describe waterfilling, which maximizes the ergodic capacity of a cooperative transmission with instantaneous channel knowledge and limited ERP. In contrast to this, I measure the outage probability and compare it under instantaneous and average channel knowledge.

Hoydis et al. [Hoy11] analyze the optimal fraction of coherence time of a channel to be used to determine the instantaneous channel quality. I assume the instantaneous channel quality is known and analyze how ERP and outage probability can be traded off when using different cooperative schemes.

Park et al. [PSS09] describe how an adaptive use of cooperative and non-cooperative schemes can maximize the network's capacity. Zakhour and Hanly [ZH10] maximize the minimum data rate. Instead of data rate I use the outage probability as a quality metric for the resulting transmission.

Other work considers scenarios with relaying UEs [AK04, LT04, NH04] and multi-hop communication [LFK10]. Biermann et al. [BSC+12] compare how well different back-haul topologies are suited for cooperative transmissions. I assume a list of all possibly cooperating BSs to be available from which a selection can be made. Maaref and Aïssa [MA05] describe the outage and ergodic capacity of multiple-input and multiple-output (MIMO) Systems. I determine the outage probability and the ERP of *cooperating* BSs.

While I assume the channel state to be determined without overhead, Ramprashad and Caire [RC09] consider the overhead of collecting this information in a MIMO System. Similarly, Goldenbaum et al. [GAV11] consider the effect of delayed channel state information. Another group of work [BK12, LSC12] focuses on the practical aspects of cooperative transmissions while I provide analytical results.

6.3. Model

I consider only the connection of a single UE to the RAN and only downlink transmissions from the BSs to the UE. I denote the instantaneous channel gain from BS i to the UE as γ_i and its probability density function as p_i with the mean Γ_i. Note that I do not make any assumption about the distribution of the instantaneous channel gain and, hence, the results are valid for all block-fading fading environments. For simplicity, I assume the BSs to be ordered by decreasing average channel gain Γ.

6.3.1. Effective radiated power

I define the ERP r_i of a single BS to be the power that is transmitted from the sender's antenna including all gains and losses of the sender and its antenna as well as the gain of the receiving antenna. For simplicity, I normalize the ERP to be between 0 and 1. *Power control* is the ability of a BS to change the ERP in the interval $[0, 1]$ for each fading block, while a BS without power control is limited to the values 0 and 1. Note that mean ERP, in addition to being a metric (if the BS adapts the ERP based on the

channel quality), can also be seen as a parameter (if setting ERP to a constant value). When not further specified I assume all BSs transmit at full power.

I do not explicitly consider the number of transmit and receive antennas at each BS, but as long as the signals of different BSs can be additively combined my results also hold for MIMO systems. Because BS clustering and the back-haul network constrain the free selection of BSs, I assume a list of possibly cooperating BSs is available from which the actually cooperating BSs are selected.

6.3.2. Static vs. dynamic association

I compare two different types of cooperative schemes: static and dynamic. Using *static association*, the UE is associated to the c BSs with the highest *average* channel gain. Using *dynamic association*, the UE is associated to the c BSs with the highest *instantaneous* channel gain. I use n to denote the total number of available BSs while c denotes the number of BSs that can actively cooperate.

I assume all users require a fixed minimum data rate. I model this as a threshold T_D of signal-to-noise ratio (SNR) above which the transmission succeeds and below which it fails. The probability that this threshold is not met is called *outage probability* O:

$$\mathbb{P}[\text{Outage}] = \mathbb{P}\left(\underbrace{\frac{r_i\gamma_i}{N}}_{\text{SNR}} < T_\mathrm{D}\right) = \mathbb{P}\left(r_i\gamma_i < T_\gamma\right), \tag{6.1}$$

where N is the mean power of the noise. I denote the outage probability using static association by O_S and using dynamic association by O_D and power control by a "p" in the index. For simplicity, I define the power threshold $T_\gamma = T_\mathrm{D}N$ as the threshold of ERP r and channel gain γ for a successful transmission. This threshold T_γ is system-dependent but a constant for the purpose of my analysis. I denote the sum of the ERP of all BSs averaged over time by R.

6.4. Outage probability

In this section, I describe the outage probability for static and dynamic association when all cooperating BSs transmit at maximum power. While this minimizes the outage probability it will also radiate more power than necessary. These calculations provide a lower bound for the possible outage probability when the ERP is reduced. The calculation of outage probability using static association is not new [Gol05], but I included it in this chapter, using the notation of the rest of the dissertation, as a reference for dynamic association.

6.4.1. Static association

While I focus on cooperative transmissions from the BSs, non-cooperative transmission is the special case of cooperative transmission when only one BS transmits.

Static association without cooperation

Without cooperation and with static association, the lowest outage probability is achieved by associating the UE to the base station with the highest average channel gain. Because I number the BSs in decreasing order of average channel gain, BS 1 has the highest average channel gain.

The outage probability for static selection of a single BS can be calculated as:

$$O_{\mathrm{S}}(1) = \mathbb{P}[\gamma_1 < T_\gamma] = \int_0^{T_\gamma} p_1(x)\mathrm{d}x. \tag{6.2}$$

Static association with two cooperating BSs

If I allow two BSs to cooperate and select these BSs based only on the average channel gain, it is best to select the two BSs with the highest average channel gain to maximize the their sum. Because I number the BS in decreasing order of average channel gain, these are BS 1 and 2. The resulting outage probability with static selection of two BSs is:

$$O_{\mathrm{S}}(2) = \mathbb{P}[\gamma_1 + \gamma_2 < T_\gamma] = \int_0^{T_\gamma} p_1(x_1) \int_0^{T_\gamma - x_1} p_2(x_2)\mathrm{d}x_2\mathrm{d}x_1. \tag{6.3}$$

Static association with c cooperating BSs

If a UE is statically associated to c BSs, it is best to select the c BSs with the highest average channel gain to maximize their sum (note that this only minimizes the outage probability when all quantiles of the distribution with the higher mean are also higher than the corresponding quantiles of the distribution with the lower mean, which is true for the exponential distribution). Because I number the BSs in decreasing order of average channel gain, these are BSs 1 to c. The resulting outage probability with static selection of c BSs is:

$$O_{\mathrm{S}}(c) = \mathbb{P}\left(\sum_{i=1}^{c} \gamma_i < T_\gamma\right) = s_c(T_\gamma), \tag{6.4}$$

where $s_i(t)$ is the probability that the sum of the channel gains of BSs 1 to i is lower than t (i.e., the convolution all c involved γ_i):

$$s_i(t) = \begin{cases} \int_0^t p_i(x_i) s_{i-1}(t - x_i)\mathrm{d}x_i & \text{if } i > 0, \\ 1 & \text{else.} \end{cases} \tag{6.5}$$

Under the Rayleigh fading assumption, the outage probability of static association $O_{\mathrm{S}}(c)$ can be calculated from the Erlang distribution of shape c when all average channel gains are equal. The hypoexponential or generalized Erlang distribution [Ros09] is a generalization which allows the average channels gains to be different. Its cumulative distribution function and probability density function can also be calculated as a special case of the phase-type distribution [Neu81].

6.4.2. Dynamic association

With dynamic association, I select the best c BSs out of n possible BSs based on their instantaneous channel gain. This results in an outage probability of:

$$\begin{aligned} O_{\mathrm{D}}(n,c) &= \mathbb{P}\left(\max_{\tau\in \mathrm{S}_n} \sum_{i=1}^{c} \gamma_{\tau(i)} < T\right) = \sum_{\tau\in \mathrm{S}_n} h_\tau(n,c) \\ &= \sum_{\tau\in \mathrm{S}_n} \mathbb{P}\left(\sum_{i=1}^{c} \gamma_{\tau(i)} < T \wedge \forall i \in \{2,...,n\} : \gamma_{\tau(i-1)} > \gamma_{\tau(i)}\right), \end{aligned} \tag{6.6}$$

where S_n is the symmetric group, that is, the set of all possible permutations τ of the natural numbers $(1,\ldots,n)$ and $\tau(i)$ is the ith element of τ. I use τ to describe the order of instantaneous channel gains. I derive a formula for $O_{\mathrm{D}}(n,c)$ by applying the law of total probability over all possible permutations τ of channel gains using the function h.

The function h calculates the probability that the threshold is not met when the best c BSs cooperate and the order of the instantaneous channel gains is τ. I calculate it from the probability that the sum of the c largest instantaneous channel gains is smaller than T_γ and they are in the order τ (with the function v) and that the $n-c$ other channel gains are in the order τ (with the function q):

$$h_\tau(n,c) = \int_0^{T_\gamma/c} f_{\tau(c)}(x_c) v_\tau(c-1, T_\gamma - x_c, x_c) q_\tau(n, c+1, x_c) \mathrm{d}x_c. \tag{6.7}$$

The function q calculates the probability that the channel gains from BSs i to n (which do not contribute to the combined signal) are in the order τ:

$$q_\tau(n,i,x) = \begin{cases} \int_0^x p_{\tau(i)}(x_i) q_\tau(n, i+1, x_i) \mathrm{d}x_i & \text{if } i \le n, \\ 1 & \text{else.} \end{cases} \tag{6.8}$$

The function v calculates the probability that the channel gains of BSs 1 to i are in order τ and their sum is lower than T_γ. The parameter x is the channel gain of the BS $i{+}1$ and is a lower bound for the channel gain of BS i. The parameter $t = T_\gamma - \sum_{k=i+1}^{c} x_k$ is the channel gain that is left for the BSs 1 to i to not go over threshold T_γ:

$$v_\tau(i,t,x) = \begin{cases} \int_x^{t/i} p_{\tau(i)}(x_i) v_\tau(i-1, t-x_i, x_i) \mathrm{d}x_i & \text{if } i > 0, \\ 1 & \text{else.} \end{cases} \tag{6.9}$$

Note that it is not necessary to split the calculation over all different orders of the $n-c$ not selected BSs, but doing so gives a uniform way to state the probability. My formulation has the property that it does not need any case distinction inside the integrals, for example, for values smaller than 0. This allows easy analytical evaluation of the integrals using a computer algebra system.

In addition to being an abstraction for maximal-ratio combining (MRC) and coherent combining (CC), my formulation can also be used to calculate the outage probability for selection combining [Gol05] when using dynamic association with $c = 1$.

6.4.3. Strictly superior cooperation schemes

In this section, I show which cooperation schemes provide lower outage probability than others for all possible channel situations.

Theorem 6.1. *For natural numbers $a \leq b$ the following holds:*

$$O_S(b) = O_D(b,b) \leq O_D(b,a) \leq O_D(a,a) = O_S(a). \tag{6.10}$$

Proof. $O_{\mathrm{S}}(n) = O_{\mathrm{D}}(n,n)$: Dynamic selection of n out of n BSs is by definition the same as static selection of all n BSs.

$O_{\mathrm{D}}(b,b) \leq O_{\mathrm{D}}(b,a)$: Because both schemes select the best BSs based on instantaneous channel gain, $O_{\mathrm{D}}(b,b)$ will always select the a BSs that $O_{\mathrm{D}}(b,a)$ selects. As all channel gains are positive and the quality of the cooperative transmission depends on the sum of the individual channel gains, $O_{\mathrm{D}}(b,b)$ must be lower than $O_{\mathrm{D}}(b,a)$.

$O_{\mathrm{S}}(b,a) \leq O_{\mathrm{D}}(a,a)$: analogously. □

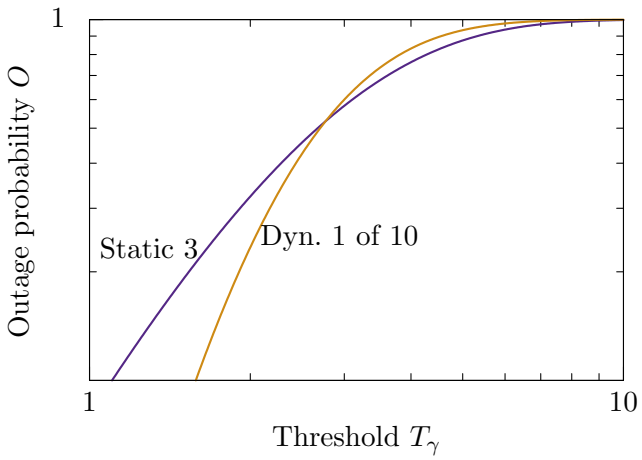

Figure 6.1.: Depending on the threshold each of two cooperative schemes can achieve a lower outage probability at the UE.

This result gives a lower and an upper bound on the outage probability of dynamic association, which are the two bounding static associations. Note that not for all pairs of cooperative schemes one is strictly superior to the other: Whether static association to 3 BSs or dynamic association to 1 of 10 BSs results in a lower outage probability depends on the threshold T_γ. In contrast to the strict superiority of one scheme above the other, Figure 6.1 shows an example in which the threshold determines which scheme provides the lower outage probability.

6.5. Effective radiated power

In this section, I calculate the total ERP (i.e., from all BSs) averaged over time for the different cooperation schemes with and without power control.

Today's BSs have a nearly binary power profile. Hence, a lower transmission power will not directly reduce the power consumption of a BS. However, the power consumption of future BSs which are developed with energy efficiency in mind will depend more on the radiated power than today (see linear power profile in section 1.2). When such BSs are used, the reduction in radiated power will also reduce the consumed power.

However, the main focus of this chapter is that the signal-to-interference-(plus-)noise ratio (SINR) of other nearby transmissions will be higher, when less power is radiated. When the SINR is higher the transmissions are and, thus, provide more opportunities to deactivate BSs.

Without instantaneous channel knowledge, reducing the ERP will increase the outage probability. Hence, there is a trade-off between ERP and outage probability. With instantaneous channel knowledge this is not necessarily the case. That is, with instantaneous channel knowledge, reducing the ERP at the correct times does not increase the outage probability.

Reducing the ERP will result in less interference for other transmissions and thus better channel quality in a multi-user environment. I do not directly quantify the effect on other transmissions in this chapter, but only use ERP to compare different transmission schemes.

6.5.1. Static association

Without instantaneous channel knowledge, it is not possible to adapt the ERP to the actually necessary value. Hence, reducing the ERP can only be done by reducing the ERP in all channel situations.

Calculating the outage probability with reduced ERP can be done with the same equations as for full transmit power: The channel gain with reduced ERP is equal to the channel gain multiplied by the ERP r at full transmit power. This holds both for average and instantaneous values. The total ERP of all BSs is $R_{\mathrm{Sp}}(c) = \sum_{i=1}^{c} r_i$ when BS i transmits with ERP r_i. In the special case of full power transmission this becomes: $R_{\mathrm{S}}(c) = c$. I use these as reference values to compare with dynamic association and instantaneous power control.

The calculation of $R_{\mathrm{Sp}}(c)$ and $O_{\mathrm{Sp}}(n, c)$ enables a calculation of the outage probability for a given allocation of ERPs to the BSs. However, this does not provide a closed formula for the optimal distribution of total ERP to minimize the outage probability.

When the instantaneous channel gains are known it is best to assign as much power to the channel with the highest gain as possible (theorem 6.2). However, this does not hold if only the *averages* of the channel gains are known. That is, allocating all transmit power the the channel with the highest *average* channel gain does not necessarily minimize the outage probability. It is possible to use Newton's method to find a local minimum in the distribution of total ERP. I present an example for this next.

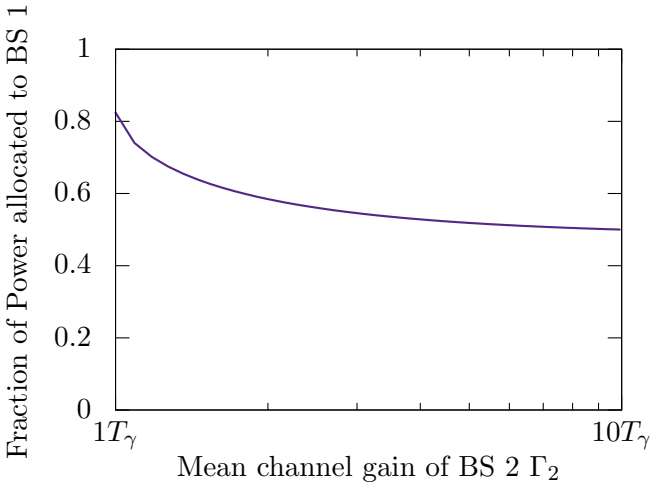

Figure 6.2.: The optimal fraction of power allocated to BS 1 depends on the channel gain of BS 2. The channel gain of BS 1 is $10T_\gamma$.

Figure 6.2 shows how the optimal power distribution between two BSs depends on the channel gain of the BSs. I determined the optimal value using Newton's method. It shows that even when the average channel gain of BS 1 is higher than the average channel gain of BS 2 it is necessary to allocate ERP to BS 2 to minimize the outage probability.

One drawback of this method is that the found local minimum is not always the global minimum: For example using 2 BSs with $\Gamma_1 = 15T_\gamma$ and $\Gamma_2 = 0.995T_\gamma$, a local minimum is at about 0.14 of the power allocated to BS 1, while the global minimum is at 0. While the actual difference in outage probability is small (at least in this example), the problem has to be considered when applying Newton's method to find the optimal allocation of ERP to the BSs. For the remainder of this chapter, I assume all BSs which transmit using static association to transmit at full power. Hence, the power does not need to be distributed, but all BSs transmit at full power.

Figure 6.3 shows that the actual difference is small, but a local minimum is not necessarily a global minimum for a distribution of ERP to the BSs.

6.5.2. Dynamic association

At first I calculate the ERP with dynamic association without power control. That is, the BSs do not transmit at all when the sum of the channel gains will be lower than the necessary threshold. When the threshold can be reached only those BSs transmit which are needed to reach the threshold. With $O_\mathrm{D}(n, 0) = 1$, the expected ERP without power

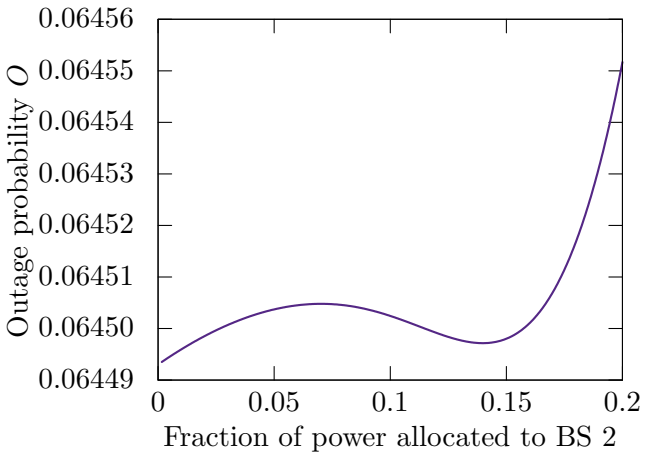

Figure 6.3.: The outage probability depends on the distribution of the ERP. This is an example that the local minimum t 0.14 does not have to be the global minimum at 0. The BSs have a channel gain of $15T_\gamma$ and $0.995T_\gamma$.

control can be written as:

$$\begin{aligned} R_{\mathrm{D}}(n,c) &= \sum_{i=1}^{c} i\mathbb{P}[i \text{ BSs are necessary and sufficient}] \\ &= \sum_{i=1}^{c} i(\overbrace{(1-O_{\mathrm{D}}(n,i))}^{i \text{ BSs are sufficient}} - \overbrace{(1-O_{\mathrm{D}}(n,i-1))}^{i-1 \text{ BSs are not}}) \\ &= \sum_{i=1}^{c} i(O_{\mathrm{D}}(n,i-1) - O_{\mathrm{D}}(n,i)) \\ &= \sum_{i=1}^{c} iO_{\mathrm{De}}(n,i), \end{aligned} \tag{6.11}$$

where $O_{\mathrm{De}}(n,c) = O_{\mathrm{D}}(n,c-1) - O_{\mathrm{D}}(n,c)$ is the probability that exactly c BSs have to cooperate.

With power control, it is possible to transmit only with the ERP necessary to reach the threshold T_γ. The following definition and its use describe how much total ERP is necessary to reach the threshold T_γ and how to distribute it.

I define an allocation of power to be an *on-off allocation* if the following condition is met:

$$\exists j \in \{1,\ldots,n\} : \forall i < j : r_i = 1 \wedge \forall i > j : r_i = 0. \tag{6.12}$$

An intuition of the on-off allocation is that it tries to transmit as much power as possible over the best channels. This not only seems reasonable, but is also the best method to allocate the power. Note that this is only holds when the instantaneous channel gains are known, but not for the average channel gains.

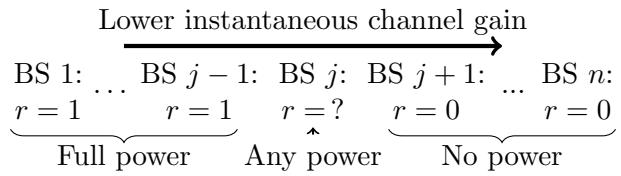

Figure 6.4.: The on-off allocation maximizes the ERP on the channels with the highest instantaneous channel gain.

Figure 6.4 illustrates the on-off power allocation: a BS j exists that transmits at some ERP, all BSs with a higher channel gain transmit at full ERP, and all BSs with a lower channel gain do not transmit. Note that the on-off allocation is uniquely defined for a given total ERP, when all average channel gains Γ_i are different. Else they are unique except for permutation.

Theorem 6.2. *If the instantaneous channel gains γ_i are known, the threshold T_γ is reachable with minimum sum of radiated power if the ERP is distributed according to the on-off allocation sorted by instantaneous channel gains.*

Proof. Assume an allocation A which is not an on-off allocation radiates the least total power and reaches the threshold T_γ at the UE. Select j to be the index of the BS with the lowest channel gain with ERP $r_j > 0$. Because all BSs i with smaller channel gain have $p_i = 0$ the statement $\forall i > j : r_i = 0$ is fulfilled. Hence, there must be a BS k with $k < j$ and $r_k < 1$ (else it would be an on-off allocation).

Now construct a power allocation that reaches the threshold and radiates less power than allocation A and, thus, prove it cannot have been the one that radiates least power. Because I only change the ERP of BS i and k I ignore all others. In allocation A the BS i and k generate a total received channel gain of $r_i\gamma_i + p_k\gamma_k$ at the receiver. Note that $\gamma_k > \gamma_i$ by the way k and j were selected. Select the new values $r_i^* = r_i - \gamma_i/\gamma_k\epsilon$ and $r_k^* = r_k + \epsilon$, with $\varepsilon < \min(r_i, 1 - r_k)$. This results in the same received channel gain $r_i^*\gamma_i + r_k^*\gamma_k$ as the allocation A: $r_i\gamma_i + r_k\gamma_k$. But because $\gamma_i < \gamma_k$ the total ERP is lower. □

Figure 6.5 illustrates how shifting ERP to a BS with higher instantaneous channel gain reduces total ERP, but still reaches the same threshold of channel quality. Note that this is different from waterfilling power allocation [TV05], which maximizes ergodic capacity instead of minimizing ERP for a given outage probability.

An implication from the optimality of the on-off allocation is the following: In a set of cooperating BSs with instantaneous power control, all BSs that transmit will transmit at full power, with the single exception of the BS with the lowest instantaneous channel gain. Hence, adapting the ERP will only make a difference at one BS, namely the BS that is actively transmitting and has the lowest instantaneous channel gain.

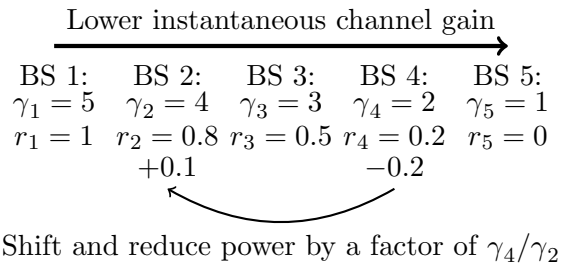

Figure 6.5.: Allocating ERP from BS 4 to BS 2 reduces the total ERP as the increase at BS 2 is lower than the decrease at BS 4, while the signal strength at the receiver does not change. This is an illustration of theorem 6.2

The mean ERP for a single BS which can instantaneously adapt its power can be calculated as:

$$R_{\mathrm{Dp}}(1,1) = \int_T^\infty p(x) \overbrace{T_\gamma / x}^{\text{ERP}} \mathrm{d}x. \tag{6.13}$$

Calculating the optimal allocation of power in a given situation with instantaneous channel knowledge can be derived from the on-off allocation. The mean ERP if selecting 1 out of n BSs is:

$$R_{\mathrm{Dp}}(n,1) = \sum_{\tau \in \mathrm{S}_n} \int_{T_\gamma}^\infty p_{\tau(1)}(x_1) \frac{T_\gamma}{x_1} q_\tau(n,2,x_1) \mathrm{d}x_1, \tag{6.14}$$

where the function q_τ is from section 6.4. For $c > 1$ BSs the following formula describes the mean total ERP:

$$R_{\mathrm{Dp}}(n,c) = R_{\mathrm{Dp}}(n,c-1) + \sum_{\tau \in \mathrm{S}_n} u_\tau(n,c,1,T_\gamma,T_\gamma), \tag{6.15}$$

where the function u describes the ERP if c of n BSs are necessary and sufficient to not be in outage and instantaneous channel gains are in order τ:

$$u_\tau(n,c,i,t,x) = \begin{cases} \int_{t/(c-i+1)}^{\min(t,x)} p_{\tau(i)}(x_i) u(n,c,i+1,t-x_i,x_i) \mathrm{d}x_i & \text{if } i < c, \\ \int_t^x f_{\tau(i)}(x_i) \underbrace{(c-1+t/x_i)}_{\text{ERP}} q_\tau(n,c+1,x_i) \mathrm{d}x_i & \text{else,} \end{cases} \tag{6.16}$$

where t is the amount of channel gain that the BSs i to c have to provide. The channel gain of BS i must be smaller than the channel gain of BS $i-1$, which is represented as parameter x. The formula includes the calculation of a minimum in the boundaries of the integral, which must be replaced with a case distinction to evaluate it. This makes evaluation more complex, but I was unable to find a solution without such a distinction.

Bounds for $R_{\mathrm{Dp}}(n,c)$, which are easier to evaluate than the exact formula, can be expressed with the help of the on-off allocation:

$$(c-1)O_{\mathrm{De}}(n,c) \leq \sum_{\tau \in \mathrm{S}_n} u_\tau(n,c,T_\gamma,1,T_\gamma) \leq cO_{\mathrm{De}}(n,c). \tag{6.17}$$

The lower bound ignores the power radiated from the BS that is not transmitting at full power. The upper bound considers it to transmit at full power. Note that it is only necessary to replace the exact terms of u_τ in equation 6.15 by the boundaries when three or more BSs cooperate because the exact terms for one and two BSs are easy to evaluate.

6.6. Results

In this section, I describe implications from the results of the previous section. Now, I assume the instantaneous conditions of the channel to be described by Rayleigh fading, making the instantaneous channel gains exponentially distributed [TV05]. The results shown in this section are analytical results derived with the help of the computer algebra system Maxima 5.27.0 (and not results of simulations).

6.6.1. Scenarios with high gain

First, I explain why the benefit of dynamic association becomes greater the more similar the average channel gains are. Dynamic association will always select the BSs with the highest instantaneous channel gain to cooperate and, hence, will always use the best BSs. In contrast to this, static association selects the BSs based on average channel knowledge and, hence, can make non-optimal selections of BSs.

If the order based on the average and the instantaneous channel gains of all BSs is the same, both methods will select the same BSs. More precisely, if and only if the c BSs with highest average channel gains also are the c BSs with the highest instantaneous channel gains, both schemes will pick the same BSs.

The probability that one exponentially distributed random variable is smaller than another is $\mathbb{P}[X < Y] = \frac{\lambda_X}{\lambda_X+\lambda_Y}$ [Ros09]. Hence, the probability that the order of two channel gains is different for the average and instantaneous values is smaller when their expected values are further apart. Given this, the probability that static association selects the best BSs is:

$$\begin{aligned}\mathbb{P}[\text{Static selects best BSs}] &= \mathbb{P}\left[\bigwedge_{\substack{1\leq i\leq c\\ c<j\leq n}} \gamma_i \geq \gamma_j\right] = \mathbb{P}\left[\bigwedge_{c<j\leq n}\left(\min_{1\leq i\leq c}\gamma_i\right) \geq \gamma_j\right] \\ &= \int_0^\infty p_\Sigma(y) \prod_{j=c+1}^{n} \int_0^y p_j(x)\mathrm{d}x\mathrm{d}y,\end{aligned} \tag{6.18}$$

where $f_\Sigma(y)$ is the probability density function of $\min_{1 \le i \le c} \gamma_i$, which is exponentially distributed with a mean of $\Gamma_\Sigma = 1/\left(\sum_{i=1}^{c} \frac{1}{\Gamma_i}\right)$.

While dynamic association will always select the best BSs it will not guarantee a successful transmission, but if static and dynamic association select the same BSs their outage state will be the same. When they select different BSs there are two possibilities: (1) if the dynamic association is in outage, the static association must also be in outage, because it cannot select better channels and (2) if the dynamic association is not in outage, the static association may or may not be in outage, depending on the selection of BSs. Hence, if the static association selects non-optimal BSs it does not mean that it performs worse than the dynamic association, but it is a necessary condition to do so. I will analyze the full effect next.

I conclude that the gain from using instantaneous knowledge to associate UEs to BSs is greatest if the average channel gains are similar. Or put differently: using instantaneous knowledge for association is unnecessary if the average channel gains are very different.

6.6.2. Outage probability and channel gain

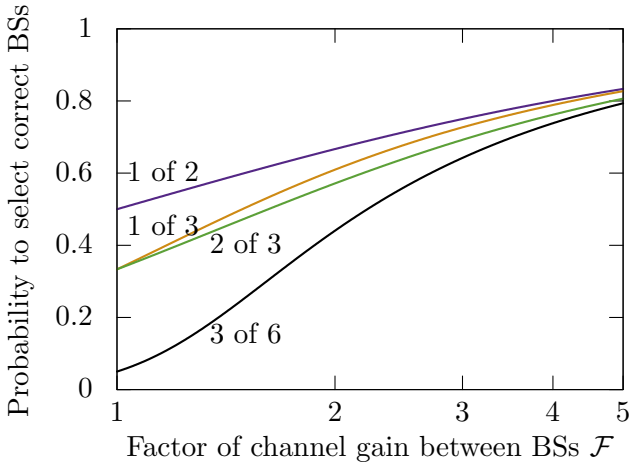

Figure 6.6.: The probability that static association selects the correct BSs depends on factor $\mathcal{F}$ by which the average channel gains of the BSs are apart.

Figure 6.6 shows the probability that static association selects the best c of n BSs if the average channel gains of the BSs are a geometric progression with the common ratio $1/\mathcal{F}$. The average channel gain of BS i is $10T_\gamma/\mathcal{F}^i$. For example, for a path-loss exponent of 2 and successive BSs being at the double distance $\mathcal{F}$ equals 4. This means that on the left side of the plot the average channel gains are close to each other while they are different on the right side. The figure shows that the probability to select non-optimal BSs is lower the more different the average channel gains are.

In RANs the users with the worst data rates are on the edge cells. That is, they are

close to the boarder between two cells and thus approximately equally far from two BSs. Therefore, the UEs which have the worst data rate benefit the most from cooperation because they usually have several BSs with similar average channel gain in range.

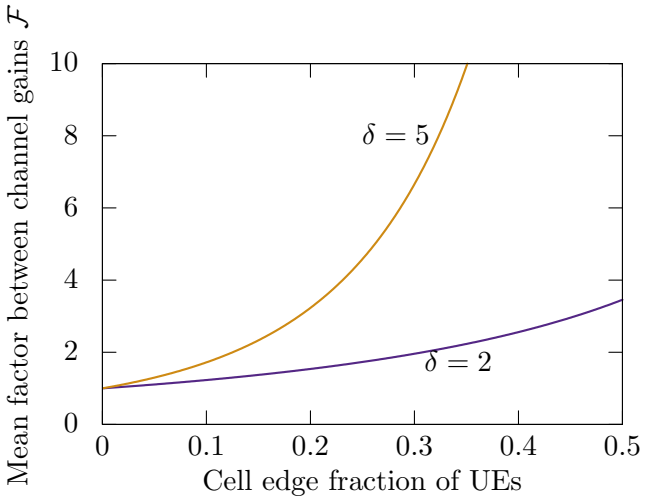

Figure 6.7.: The average factor between channel gains for UEs at the cell edge is close to 1. When the fraction of UEs considered as belonging to the edge increases the factor between channel gains $\mathcal{F}$ increases.

To get an understanding of the relevant sizes of $\mathcal{F}$ consider UEs on a line between two BSs (as in figure 1.8). Figure 6.7 shows the factor between the average channel gains $\mathcal{F}$ depending on the fraction of UEs considered to belong to the cell edge. It shows that even for relatively large fractions of considered locations of UEs the average factor between channel gains is low. It also shows that for higher path-loss exponents δ the factor $\mathcal{F}$ is higher for the same fraction.

Figure 6.8 illustrates the effect that instantaneous BSs selection is better than static selection especially if the channel gains are similar. Additionally, it shows that dynamically selecting 2 out of 3 BSs is nearly as good as always using 3 BSs in terms of outage probability.

Figure 6.9 shows the effect of Theorem 6.1: Some schemes provide lower outage probability than others for all possible thresholds. Note that the dynamic association to 2 out of 3 BSs is closer to the static selection of 3 BSs than to the static selection of 2 BSs. This shows that selecting 2 out of 3 BSs generates nearly the same outage probability as always using all 3 BSs, but is better than statically selecting 2 BSs.

Figure 6.10 shows the difference in ERP between static and dynamic association becomes smaller the more different the channel gains of the cooperating BSs are. These results match the results which determine the probability to select the correct BSs (figure 6.6). Note that figure 6.10 does not show anything about the efficiency of the compared schemes because they have different outage probabilities. I relate the ERP to

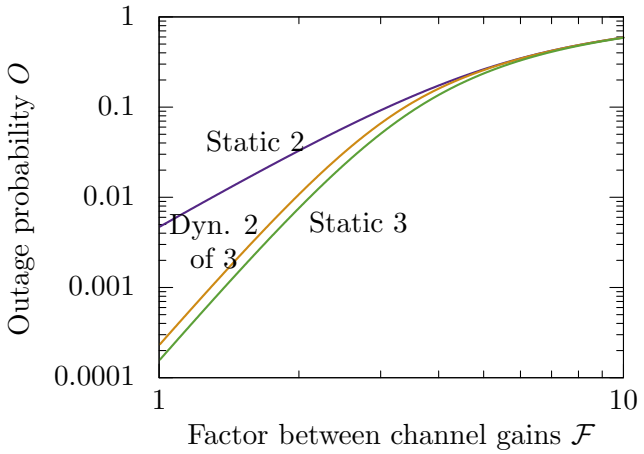

Figure 6.8.: The outage probability of different cooperation schemes depends on the factor between individual average channel gains $\mathcal{F}$.

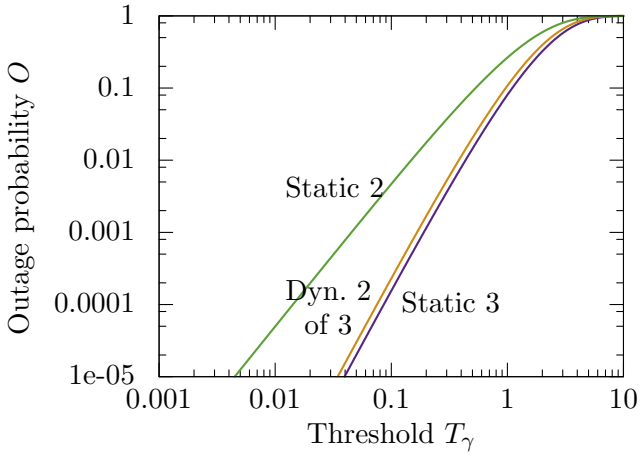

Figure 6.9.: Outage probability of static and dynamic association of 3 BSs depends on the thresholds T_γ at the UE.

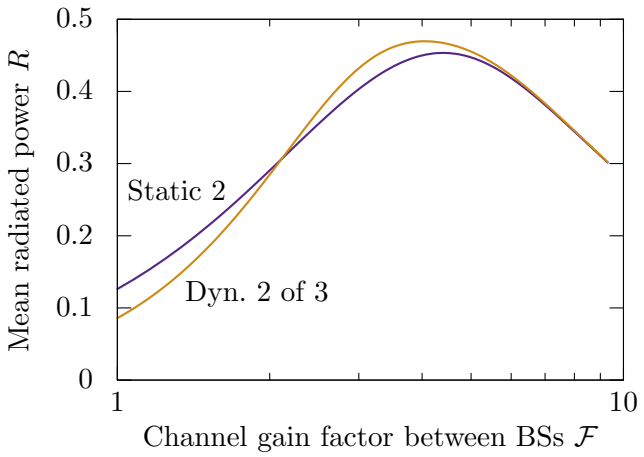

Figure 6.10.: The mean ERP of cooperative schemes with power control depends on factors between the channel gain of individual BSs.

the outage probability in the next section.

6.6.3. Effective radiated power

The following plots show how the distribution of ERP to two different BSs influence the outage probability if the total ERP is fixed.

Figure 6.11 shows that distributing the ERP between BSs can reduce the outage probability. Note that this is in contrast to theorem 6.2 that distributes power optimally if the instantaneous channel conditions are known. I have not been able to find a closed form solution for the optimal ERP distribution without instantaneous channel knowledge. However it is possible to determine the optimum numerically.

More power is radiated if a higher threshold has to be reached at the receiver. Because a higher SNR at the receiver has to be reached. Figure 6.12 however, shows that when the threshold reaches a certain level, further increasing the threshold reduces the mean ERP, because the channel is in outage more and more often and no power is radiated in this case. This illustrates that considering only the ERP without the outage probability is not meaningful.

6.6.4. Relating ERP and outage probability

For each individual transmission scheme a trade-off between ERP and outage probability exists. Finding the most energy-efficient scheme (in J/bit) is not helpful, because as the most energy-efficient scheme will only transmit when the channel is extremely good. This would result in unusably low data rates. Hence, I do not compare the energy efficiency of the different cooperative schemes.

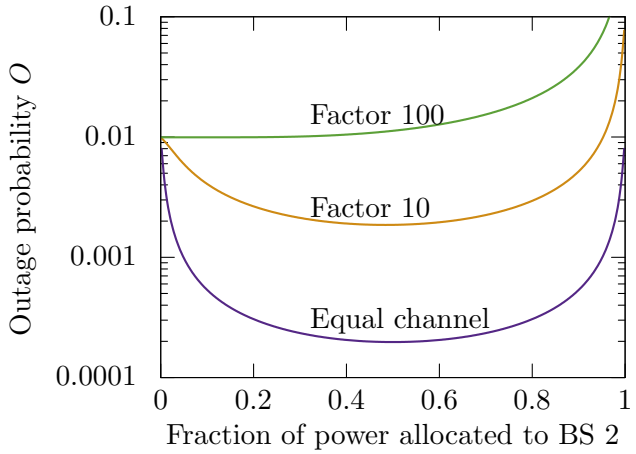

Figure 6.11.: The outage probability of different cooperative schemes depends on the distributions of ERP between the two available BSs. Here the average channel gain of BS 1 is fixed and the average channel gain of BS 2 is lower by a factor of 1 (10 and 100, resp.).

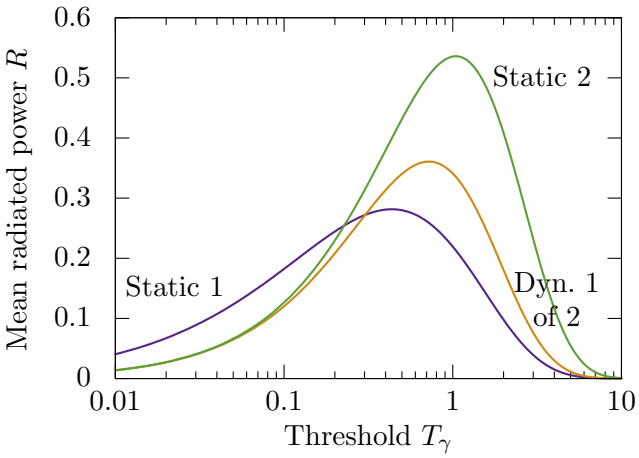

Figure 6.12.: The ERP of different cooperation schemes with power control depends on the receive thresholds at the UE. Note that for high thresholds the channel is in outage more often and the power control prevents all transmissions in this case.

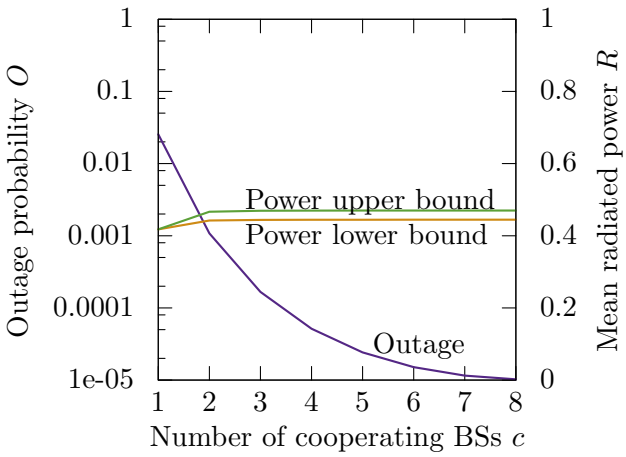

Figure 6.13.: The outage probability and mean ERP depend on the number of cooperating BSs.In this scenario dynamic association is used to determine the BSs best suited for cooperation from a pool of 8 BSs with equal average channel gain.

Figure 6.13 shows how outage probability and ERP are related. The outage probability can be decreased greatly by letting more BSs cooperate. Because dynamic association only uses the additional BS when necessary, the ERP is kept low. Additionally, figure 6.13 illustrates the quality of the bounds for ERP.

Figure 6.14 summarizes both ERP and outage probability for a scenario with 10 BSs in the range of the UE. The static association with power adaption achieves the lowest ERP because it only transmits if the channel is not in outage; but as its outage probability is about 63% in case $c = 1$ it does not transmit most of the time. The static association with power adaption and the dynamic association become more similar the higher the allowed number of cooperating BSs gets and becomes the same at $c = 10$ because both select all 10 BSs to cooperate and determine the necessary ERP using the on-off allocation.

I conclude that the use of instantaneous channel knowledge to select the cooperating BSs provides a large reduction in ERP, while additionally using power control reduces the ERP further but not by as much. Figure 6.14 also shows that a higher number of cooperating BSs does not radiate much more power than non-cooperative transmission if instantaneous channel knowledge can be used to control the ERP.

6.7. Conclusion

In this chapter, I provided the means to calculate the outage probability and ERP for cooperative schemes with and without instantaneous channel knowledge. The results allow me to quantify the gain of providing cooperative BSs with instantaneous channel

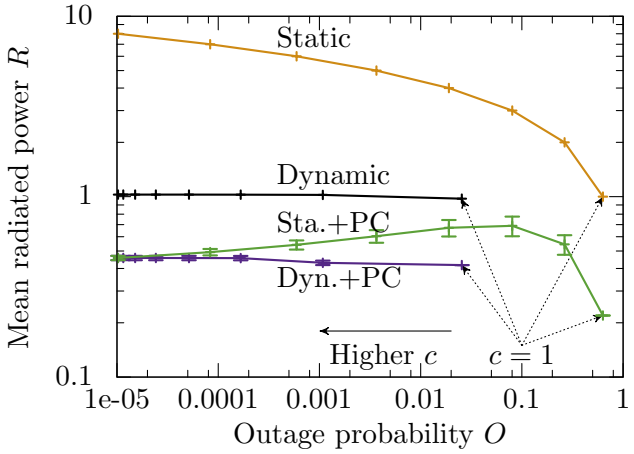

Figure 6.14.: The different schemes for static and dynamic association have different outage probabilities and ERP. In this scenario 10 BSs are available, which all have an average channel gain of T_γ. The intervals mark upper and lower bounds. The lines connect the different cooperative degrees of the same scheme: static and dynamic association with and without power control (PC). The cooperative degree starts at $c = 1$ on the right and increases to the left by 1 for each drawn point.

knowledge. The formulas I provides can be used to quantify the different trade-offs between ERP, outage probability, and number of cooperating BSs.

The on-off allocation and the evaluations show that using instantaneous channel knowledge to select the cooperating BSs is similar to using it for power control because it only makes a differences at one BS. I also showed that instantaneous channel knowledge provides the greatest gain over average channel knowledge if the average channel gains of possibly cooperating BSs are the same.

Possible future work includes extending the two special cases (average channel knowledge and instantaneous channel knowledge) to a continuum of cases, that is, quantifying the effect of non-perfect and delayed channel knowledge. Moreover, it would be interesting to determine the average factor channel gains of BSs $\mathcal{F}$ experimentally.

7. Network simulation

In the previous chapters, I determined the the latency and power consumption of radio access networks (RANs) analytically. I analyzed effects auch as power-cycle durations and cooperative transmission individually. The next step is to combine all effects in a single model and determine how they interact. However, determining a solution to a single model that contains all aspects is impractical. Therefore, I show the results of an event-based simulation in this chapter.

The goal of this chapter is to determine if the methods to conserve power, which I described in the previous chapters can be applied to a more realistic scenario. As a realistic scenario I choose the dense urban scenario of the mobile working group of the GreenTouch consortium. This scenario is based on the 3GPP E-UTRA simulation model [3GP10]. Because I try conserve energy with a sleep mode, I consider only the scenario with low load. In this scenario the potential gain from deactivating is greatest. The work of this chapter was partially funded by the GreenTouch consortium. For the work I present in this chapter Till Hohenberger supported me by implementing large parts of the simulation. Additionally I thank Daniel Roeske for finding and removing bugs in the simulator.

7.1. Introduction

In this chapter, I analyze how to decide which base stations (BSs) should be active. I compare both local and central algorithms to make this selection. In contrast to related work, I do not develop new heuristics, but take algorithms for which a analytic analyses exists and apply them to a realistic simulation. These algorithms are in part those which I explained in the earlier chapters.

I consider the local "accumulate and fire" algorithm and its variants (see section 2.3.3). They consider power cycle durations, but have no notion of interaction between BSs. I also analyze a set cover approach (see chapter 5), which considers interaction between BSs but has no notion of power cycle durations. In the theoretical analysis of these algorithms some effects were ignored (interaction of BSs and power cycles, resp.). This means that assumptions of the theoretical analyses do apply here. However, the theoretical analyses provide hints at their performance.

In contrast to the theoretical analyses, I consider all these effects for all algorithms in this chapter. The contribution of this chapter is to determine which of the algorithms performs best in a realistic scenario. Additionally, I compare both the achieved data rates and the energy consumption of scenarios with less macro BSs (see chapter 4).

7.2. Related work

Fehske et al. [FFMB11] describe the global energy consumption of RANs. Other authors [CYZK11, HHA+11, HA13, MMSL11, OKLN11] provide overviews and explain different ideas how to reduce the energy consumption of RANs. Others [ABH11, BZB10, CZB10, HB11] categorize approaches differently (e.g., by network layer or needed BSs capabilities). All overview chapters describe the problems and possible approaches, while I pick one possible approach (sleep mode) to reduce the energy consumption and compare different methods to use it.

Cai et al. [CXY+03], Hossain et al. [HMJ11], Falconetti et al. [FHG13] and Niu et al. [NWGY10] implement a scenario similar to ours in a simulator and compare different versions of their own heuristics. In contrast to their work I compare theoretically well understood algorithms to each other and to the optimum. This gives me confidence that my algorithms also behave well in scenarios I have not simulated.

Other simulations consider only macro BSs [BEK+10] or more different types of BSs [AGD+11, CFC11]. They determine the achieved data rates for different network configurations. Their work does not consider how to determine which BSs to activate and which to put to sleep.

McLaughlin et al. [MG11] determine which transmission modes for multi antenna systems are more efficient and use only short sleep cycles (on time scales of milliseconds). They determine the effect for only a single BS while I focus on the interaction of BSs. Also, I use only a single mode and determine how to use this sleep mode with longer sleep cycles (on time scales of seconds).

Others [SFdSF12, TGAA13] analyze similar models analytically. While their results are analytical instead of based on simulations, the models they use are simpler than ours and for example do not include activation and deactivation times. Others [LCB11, VDC+12] use such simplified models to formulate the problem as an integer linear problem and solve it optimally. Because my model is too complex to find optimal solutions I simulate heuristics instead.

Ericson [Eri11] determines the energy consumption of RANs analytically by estimating the fraction of time a BS will spend in sleep mode for different protocols. Both Marsan et al. [MC09, MCCM12, MM11, RRMF13] and Oh and Krishnamachari [OK10] determine the fraction of time a BS can be switched off from the daily traffic patterns. In contrast to this, I compare different algorithms which adapt the activity of BSs depending on the channel quality to nearby user equipments (UEs).

Others [FMF10, MFF10, RFGB10, SER11] compare different strategies to place BSs and different numbers of macro and pico BSs. They try to reduce the energy consumption under maximum load (where a sleep mode is not needed), while I try to reduce the energy consumption under low load in RANs where BSs need to be put into sleep mode. Son and Krishnamachari [SK12] determine how distributing the load of BSs differently and adapting the processing speed of BSs can reduce energy consumption. While they base their energy-conservation method on speed scaling I use a sleep mode.

To reduce the total energy consumption of a RAN it is possible it reduce the energy consumption of each BS on its own. This can be done by changing the timing of

transmissions [HAH11, ZHS04], the number of active radio units [HA11, SDA+11], and the interaction of sub-components of a BS [MA11]. Torrea-Duran et al. [TDD12] consider adapting the transmission parameters to the channel conditions to conserve energy. While all these methods also reduce the energy consumption of RANs, I consider the possibilities on a higher layer of abstraction, where I coordinate the sleep modes of BSs instead of reducing the energy consumption of each BS individually.

Further approaches to the problem of reducing energy consumption of RANs include: Auctions to distribute radio resources [CMP09] and mechanism design for pricing schemes [CA11]. These two approaches assume that the users of a RANs are (greedy) adversaries which do not agree on the best usage of the resources of a RAN. While they seek ways to achieve an agreement, I assume this is dictated by the RANs. While I only consider a single RAN, Ismail and Zhuang [Ism11] describe how different RANs can cooperate to reduce the total energy consumption.

7.3. Model

My model is based on the 3GPP E-UTRA Model [3GP10] and is used in all system-wide simulations of the mobile working group of the GreenTouch consortium. Refer to Blume et al. [BAWB13] for a more detailed model description. This scenario is the default scenario for dense urban (DU) areas with low load for the year 2020 of the mobile working group of the GreenTouch consortium.

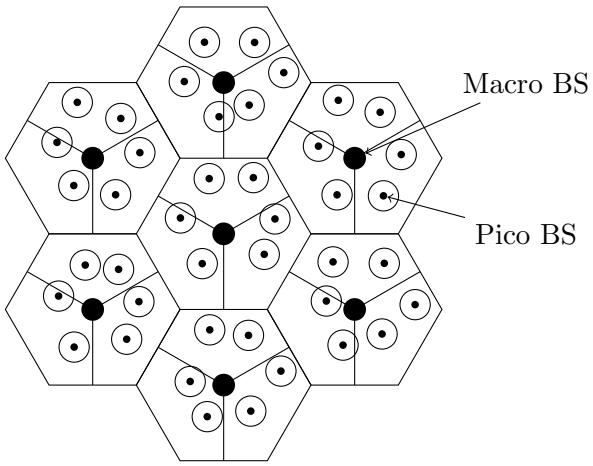

Figure 7.1.: The scenario consists of 7 macro BSs and two pico BSs per macro BS sector.

7.3.1. BS deployment

The model consists of a scenario of seven macro BSs with three sectors each; Figure 7.1 illustrates this deployment. Cells look slightly different due to antenna lobe considerations in the simulation and due not have sharp edges due to shadowing. If nothing else

is said, I model 2 pico BSs per sector. They radiate and consume less power and have only a single omnidirectional antenna.

The macro BSs are placed in a hexagonal deployment (figure 7.1). Each sector covers one third of the BS' coverage. Both scenarios, DU and sparse dense urban (SDU), consist of 7 macro BSs; the distance between the macro BSs is 500 m in DU and 1000 m in SDU.

The pico BSs' deployment depends on the positions of hotspots. At first, each hotspot will be covered by one pico BS. The remaining pico BSs will be distributed uniformly at random, equally into each sector of a macro BS.

7.3.2. UE deployment and load generation

I model only active UEs and only downlink data transfers, because they will generate most traffic in a real network (caused by internet video [Cis13]). Hence, I do not model signaling traffic in this chapter. I consider all UEs to be stationary during data transfers.

The model consists of a fixed number of hotspots. The positions of hotspots are independently and randomly selected with equal probability for each location. Two thirds of the total UE placed at hotspots with a 2-dimensional Gaussian distribution (independent multivariate normal distribution with standard deviation $\sigma_x = \sigma_y = 100\,\mathrm{m}$). The remaining traffic is generated by a homogenous Poisson process. The DU scenario has a size of $2.16\,\mathrm{km}^2$ (see figure 7.1) with two hotspots, and $8.66\,\mathrm{km}^2$ with eight hotspots for the SDU scenario.

The density of the Poisson process that creates the UEs demands is constant over time. To reflect the changing load over the course of a day, I let the simulation run with different densities, but do not change it during a single run. UEs demand a 2 MiB large transmission. If the transmission is not finished within 4 seconds they abort the transmission. These values are used, for example, by the GreenTouch consortium which based its decision on the 3GPP simulation parameters [BAWB13].

7.3.3. Radio model

To determine the average channel quality between a BS and UEs I consider the path-loss based on distance, the antenna gain (based on both horizontal and vertical angle), log-normal shadowing with a 0.5 correlation from a UE to all BSs, a constant wall-penetration loss (to compensate for walls which I do not model explicitly), receiver noise, and interference from all other transmitting BSs of the same type.

The decision whether a UE can be served from a BS depends not only on the channel between them but also on the interference from other BSs and the number of other UEs which are also connected to the BS. For this purpose I assume every BS has constant radiated power, if it transmits to at least one UE and zero radiated power otherwise.

To determine the achieved data rate from the mean signal-to-interference-(plus-)noise ratio (SINR) I use a look-up table. It was generated based on detailed channel models by members of the GreenTouch consortium, but is not publicly available. It includes multiple-input and multiple-output (MIMO) gain and fading effects. For modeling abstractions I denote it as a function dr mapping SINR to data rate. If n UEs are connected

to a BS each gets $1/n$-th of the radio resources and, thus, $1/n$-th of the data rate it would get if it were the only UE at the BSs.

To prevent border effects I implemented a wrap-around. That is, assuming the hexagonal deployment of BSs is infinite, each UE considers the 7 closest BSs. I map the 7 closest BSs to the 7 BSs in my model.

7.3.4. Power model

Each BS consumes a constant idle power (macro: 132 W, pico: 4 W) when it is active and linearly scales its power consumption in the load to maximum load (macro: 638 W, pico: 11 W). Mean is the fraction of power consumed when idle is $\mathcal{P} \approx 0.21$ for macro BSs and $\mathcal{P} \approx 0.36$ for pico BSs. The power profiles of the BSs are not assumed to be binary in this simulation, because BSs in the year 2020 are assumed to be closer to the linear power profile.

Note that pico BSs always transmit at full power or not at all, which makes their power profile binary. Because the same holds for each of the three sectors of a macro BSs their load changes in steps of one third of the maximum load. I assume the time it takes to change a BS from activated to sleep $\mathcal{A}$ (and the other way $\mathcal{Z}$) is constant (1 s) if not specified otherwise. I assume BSs do not transmit during activation and deactivation and, thus, also do not create interference.

7.3.5. Cooperative diversity

In my model I allow BSs to cooperatively transmit to UEs to increase the data rate. I extended the GreenTouch model by cooperative diversity to determine its effects. When two or more BSs cooperate they transmit the same data to the UE. I assume the BSs are synchronized so that the signal at the UE constructively interferes. Hence, the received signal of a cooperative transmission is equivalent to a single transmission with a signal that has the sum of all signal strengths. This is an abstraction of joint processing used in Coordinated Multipoint transmission of Long Term Evolution Advanced (LTE-A) [SK10], which is an implementation of coherent combining [Gol05].

I assume that cooperation can only be done with BSs of the same type, because pico- and macro BSs operate on different frequencies. As the cooperating BSs can assign different fractions of their radio resources to a UE I have to determine the final data rate. Let $C = \{\mathrm{BS}_1, \ldots, \mathrm{BS}_n\}$ denote the set of cooperating BSs and $F = \{f(\mathrm{BS}_1), \ldots, f(\mathrm{BS}_n)\}$ the set of fractions of the radio resources these BSs assign to a given UE. Let $\min_i(K)$ describe the i-th order statistic (i.e, the i-th smallest element of a set K) and define $\min_0(K) = 0$ for all K. The function dr maps SINR values to data rate (using the table described above). The data rate D_D when using cooperation is computed as:

$$d = \sum_{i=1}^{n} \overbrace{\left(\min_{i}(F) - \min_{i-1}(F)\right)}^{\text{Fraction of time exactly } i \text{ BSs cooperate}} \cdot \mathrm{dr}\left(\overbrace{\frac{\sum\limits_{f(i)\geq \min\limits_{i}(F)} \text{Signal}(i)}{N + \underbrace{\sum\limits_{f(i)<\min\limits_{i}(F)} \text{Signal}(i)}_{\text{Interference}}}}^{\text{Sum SINR of i cooperating BSs}}\right). \tag{7.1}$$

For two signals with the fraction of radio resources the BSs provide $F_1 > F_2$ this simplifies to:

$$d = \overbrace{(F_1 - F_2)\mathrm{dr}\left(\frac{\text{Signal}(1)}{N + \text{Signal}(2) + \text{Other signals}}\right)}^{\text{Only BS 1 transmits}} + \overbrace{F_2 \cdot \mathrm{dr}\left(\frac{\text{Signal}(1) + \text{Signal}(2)}{N + \text{Other signals}}\right)}^{\text{Cooperation of BS 1 and 2}}. \tag{7.2}$$

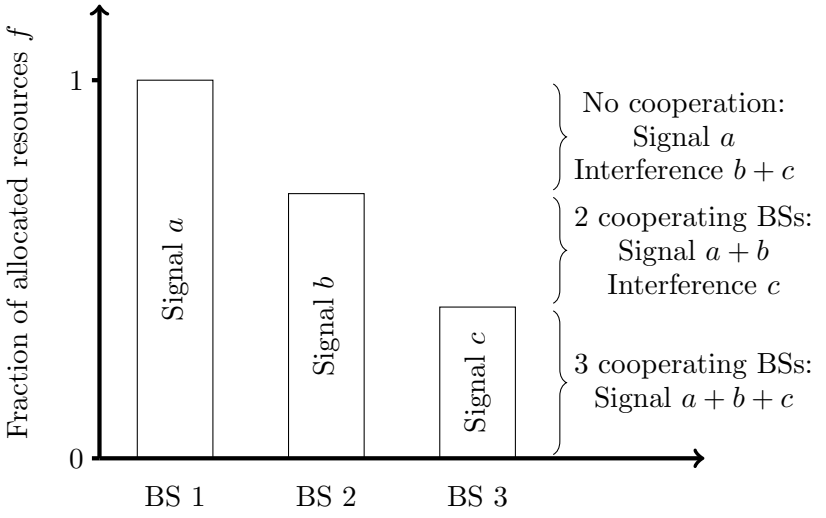

Figure 7.2.: Combining of cooperative transmissions results in a higher data rate than using only a single transmission.

This means that each BS cooperates as much as it can (depending on the fraction of time it is working on this transmission) with the others. Figure 7.2 illustrates how the fractions of the cooperating BSs determine the final data rate. While this model is more complex than needed for the current simulations (with maximum of two cooperating BSs) it allows future simulations to take advantage from more complex forms of cooperation (with more than two cooperating BSs).

Reducing the effective radiated power (ERP) as described in chapter 6 is only partially considered in this chapter. The function dr that maps the SINR to the data rate includes

effects such as MIMO transmissions and multi-user diversity. This includes transmitting data over the best channel and not transmitting over bad channels. However I do not consider the effect of reduced ERP as reduced interference. To do this it would be necessary to simulate the instantaneous channel conditions, which would increase the time need to run the simulations.

7.4. Algorithms

This section introduces the algorithms that choose which BSs will be activated and which will be deactivated or stays unchanged. In all simulations, I do not change the state of macro BSs dynamically to guarantee that signaling traffic can be received.

The theoretical analyses of the algorithms have no notion of overlap of cells or power-cycle durations which need to be considered in reality. For each algorithm I describe how I implemented the algorithm to incorporate these effects. While these changes would make the theoretical analysis far more complex their implementation changes only slightly and I can determine the results from my simulation.

The strategies are split into two main parts. (1) Making the (de-)activation decision (called *policy*); (2) assigning the UEs to BSs (called *scheme*) when a UEs creates a new demand or a BS changes its state (i.e., creating handovers).

For the association scheme I only simulate a greedy approach that assigns each UE to the *active* BS with the highest SINR. Handovers are possible to change the assignment if a BSs with a higher SINR becomes active.

To decide which BSs cooperate I determine the SINR between the best BS and the UE. If the second best SINR is less than a fixed factor (default value 2) worse (I call this *coop factor*), these two BSs cooperate. Formally: $c_f \min_{n-1}(\text{SINR}) \geq \min_{n}(\text{SINR})$, where c_f is the coop factor. Furthermore, the second best BS will only join the cooperation if it would be idle otherwise. This means that I do not activate BSs specifically to cooperate but only cooperate when a cooperation partner happens to be active.

Next I describe the (de-)activation policies I compare.

7.4.1. Always on and always off

The simplest policies always-on and always-off keep all pico BSs switched on or off, respectively. The UEs will be assigned to the BS (macro or pico) with the highest SINR. Because dr is monotonic this is equivalent to deciding by data rate. I denote Always-on by OnP and Always-off by OffP. I also used the always-on policies as a reference in previous chapters. In earlier chapters the always-off policy was not feasible because it would not have served any requests. However, because I only apply the activation policies to pico BS, macro BS can still sever requests.

Osagami [Oso05] describes the analytical background for the queuing effects of strategies in which idle BSs (servers) help others to process their jobs without the use of sleep modes. While Osagami's analysis is helpful to understand the fundamental interactions

of load sharing, the methods are too complex to be used in scenarios of larger scale. Hence, I compare it to others in my simulation.

7.4.2. Greedy and accumulate & fire

The greedy policy GP activates a pico BS B as soon as it is the BS with the highest SINR for a UE that demands a transmission. I assume I can estimate SINRs even for BSs in sleep mode. Once no UEs are assigned to a pico BS, it is deactivated immediately.

As BSs need time to activate, during activation a UE is assigned to the already *active* BS with the highest SINR and is reassigned once the activation has finished. Once B is activated, all UEs best served by B will be assigned to it. Note that the all assignment orders result in the same assignment. In case that B finishes its activation and no UE will be optimally served by B it will deactivate again immediately. The deactivation is delayed if the BS can support another BS via cooperative transmission.

Generalizing the greedy policy leads to the *accumulate and fire* policy AP(k) [CX07]. Using this policy, the pico BS only activates if the *on threshold* noted as k of UEs which would be assigned to it if it were active is exceeded. Once there is no UE connected to a pico BS it will deactivate. I denote it by AP(k). The accumulate and fire policy can cause cascading effects: when a BS finishes its transmission, the interference of other channels changes. Other UEs are reassigned and other BSs are (de-)activated. Note that the greedy policy, described earlier, is just the special case AP(1). Both the greedy policy and the accumulate and fire policy are direct applications from policies from chapter 2.

In comparison to the always on/off strategies, the accumulate and fire policy tries to make better use of the knowledge about active UEs but coordinate the activity of neighboring BSs. Hence, a BS can be activated only to discover that the UE which it was activated to serve has already been served by another BS. This can lead to increased energy consumption.

7.4.3. Set cover

Another approach to decide on (de-)activation decisions is to understand it as a set cover problem. That is, I look for the smallest subset of all BSs that is able to serve all UEs. I use the standard notation of set cover problems as used by Vazirani [Vaz01]. I have a universe U of UEs as well as a set of covering sets S of pico BSs. While U simply contains all active UEs, the set for a BS in S is built by analyzing which of the UEs in U can be served by which BS, so that the data rate requirement of the UE can be fulfilled. In terms of activity of BSs that means, I search for the smallest set of active BSs such that each UE achieves at least an given SINR.

To keep the set cover approach as simple as possible, I define a threshold of SINR over which I consider the connection as valid. During this calculation I do not update the interference calculation because this would change the possible associations and has undetermined effects for the set cover algorithms. While the activation algorithm ignores these changes in interference they are of course considered during simulation to compute the resulting download times.

I determine a solution to the set cover problem using a greedy approach [Vaz01]. Since the macro BSs are able to serve UEs as well, it is unnecessary to cover all UEs with pico BSs. Hence, I stop the set cover approach once the first 90% of all UEs are connected. I selected 90% as the remaining 10% can be served by the macro BSs in my model. For other scenarios this value has to be newly estimated. Similar to the accumulate and fire policy, a UE is assigned to the best *active* BS during state changes of BSs. I use SP(C) to describe the set cover algorithm based on the currently active UEs. The SP(C) policy one possibility to solve the problem of covering all UEs, which I analyzed in chapter 5.

As the set cover approach has no notion of activation times, the solution it determines may be outdated when the BSs reach their states. In addition, the set cover approach ignores the fact that the data rates of UEs share BSs decreases. These effects may lead to higher energy consumption and download times, which I quantify in my simulation.

Variants The set cover policy can be varied by changing the algorithm that determines the solution to the set cover problem (only greedy in my case) and the elements of the sets to cover (i.e., the contents of the sets U, S). The algorithm described above considers *currently* active UEs. Because this can change very fast, while the BSs might not be able to react that fast, it can be useful to use the set cover to cover the *average* number of UEs in a given area. In effect, this approach tries to cover areas where many requests are generated and does not try to adapt to the currently active UEs. The idea behind this variant is that it results in fewer power cycles and, thus, can better cope with long power cycles.

In this variation I split up the whole coverage area into smaller parts, each covered by a subset of BSs. U then represents the average number of UE arrivals in one of these areas. The subsets of BSs, able to serve UEs in each particular area, are defined as S.

I use SP(A) to refer to the set cover algorithm using the average UE arrival to select the active BSs.

7.5. Results

In this section, I present the results of simulating the algorithms using the model described earlier. To do so, I implemented an event-based simulation in OMNeT++.

I assume the state change durations $\mathcal{A}$ and $\mathcal{Z}$ are 1 second unless further specified. I arbitrarily selected this value but determine the results of my simulation for other values as well.

The following figures show the mean power consumption of all macro and pico BSs in watts on the X-Axis in relation to the mean download duration in seconds on the Y-Axis. Each point in the plot, stating power consumption and download duration, consists of two 95% confidence intervals for power and download duration (without Bonferroni correction). I consider a policy to be good if it results in low download times and low power consumption.

Figure 7.3 shows how the different strategies trade off the mean download duration and total power consumption. As expected, the always-on policy OnP results in the

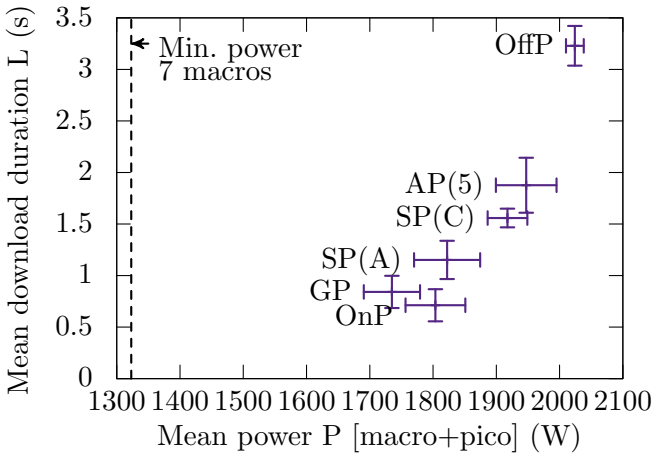

Figure 7.3.: Comparing policy performance for 20% of the average load, dense urban (DU) deployment, 42 pico BSs, 7 macro BSs, pico BSs in hotspots and random deployment, and 2 hotspots.

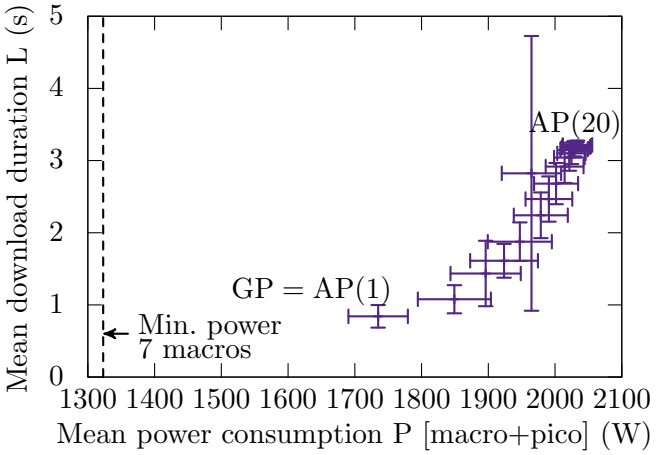

Figure 7.4.: Comparing accumulate and fire thresholds k for DU deployment, 42 pico BSs, 7 macro BSs, pico BSs deployed in hotspot and random deployment, 2 hotspots.

lowest download duration. However, the always-on policy consumes less power than the always-off policy. The reason for this is that the pico BSs can transmit the requested data to the UEs more efficiently than the macro BSs. In this scenario the greedy policy GP and the always-on policy OnP provide the best trade-offs between power consumption and download duration. The always-off policy OffP, on the contrary, has the highest download duration and highest power consumption. This shows that with these power consumption parameters it is not generally a good idea to disable as many pico BSs as possible and off-load the traffic onto the macro BSs.

As already determined analytically in section 2.6, figure 7.4 illustrates that higher activation thresholds in the accumulate and fire policy AP(k) increase the latency. However, here a higher threshold also increases power consumption as requests are not waiting, but assigned to a macro BS which serves the requests and consumes more power than the pico BS. Therefore, I will compare different ideas how to reduce the power consumption of macro BSs next.

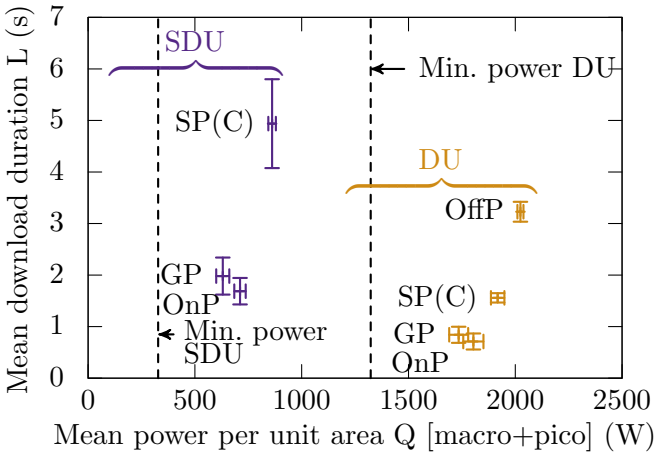

Figure 7.5.: Comparing different strategies in SDU and DU deployment, 42/168 pico BSs, 7 macro BSs, pico BSs in hotspots and random deployment, and 2/8 hotspots.

Figure 7.5 shows how the algorithms perform if the distance between macro BSs is doubled. Using the sparse deployment SDU, the download durations using the always-off policy and set cover policy SP(C) are higher but the greedy policy GP and the always-on policy OnP fulfill the demands of the UEs with the pico BSs nearly as good as the using the normal spacing. Important here is that all policies consume less power in the sparse SDU placement than in the normal placement DU. However, as it would not be possible to change the spacing at after construction of the RAN, it is necessary to determine if it also can support high load. Figure 7.6 shows that the sparse SDU deployment has longer download durations than the normal DU deployment. Next, I will explain why cooperation does not reduce the data rate, but increases the energy consumption in this

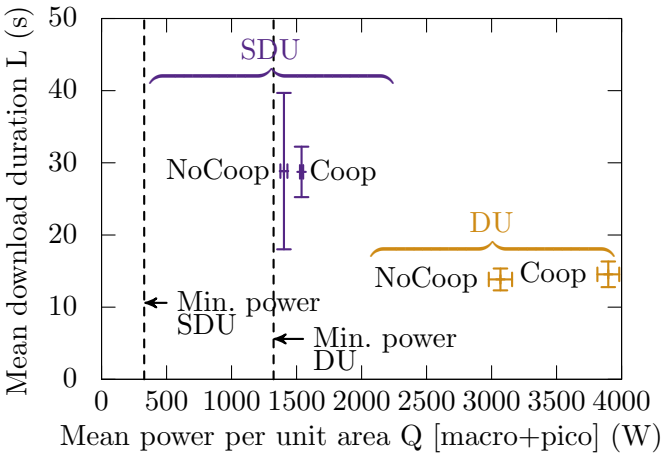

Figure 7.6.: Comparing the always-on policy OnP for 140% of the average load in SDU and DU deployment, 42/168 pico BSs, 7 macro BSs, pico BSs in hotspot and random deployment, and 2/8 hotspots.

scenario.

In contrast to the analysis of chapter 5 I assume the request size to be constant instead of the duration (this is a good model for bulk transfers, while the other is a good model for phone calls). Also cooperation has the following effects: (1) the SINR the UE receives is higher and thus the download duration is lower and (2) as the second BS that cooperates has a lower signal its energy-efficiency is lower than the first cooperating BSs. Thus, data is transmitted at a lower mean efficiency and the total energy consumption is higher. The time gained by faster downloads does not allow conserving the energy additionally spent using cooperation. Figure 7.6 shows that when BSs cooperate the power consumption increases from the increased amount of transmitted data, but the download durations do not decrease as the cooperation partners cannot increase the data rate significantly.

I also simulate scenarios in which I deactivate either 1 or 6 of the macro BSs and keep them in place in figure 7.7. This allows turning them on again when necessary during high load and still conserves power during times of low load. Deactivating macro BS is one of the methods to conserve power which I describe in chapter 4. While it does not recreate a hexagonal deployment it is also an example for re-tiling (see figures 5.9 and 5.10). Using the exact method described earlieer is not possible in a scenario with only 7 BSs. Note that I do not use the activation and deactivation strategies for the macro BSs but keep them in a single state during a simulation run. Figure 7.7 shows that deactivating macro BSs conserves large amounts of power while the pico BSs can help to provide data to the UEs.

I compare the results for different power-cycle durations in figure 7.8. It shows that faster power cycles lead to both lower power consumption and lower download durations

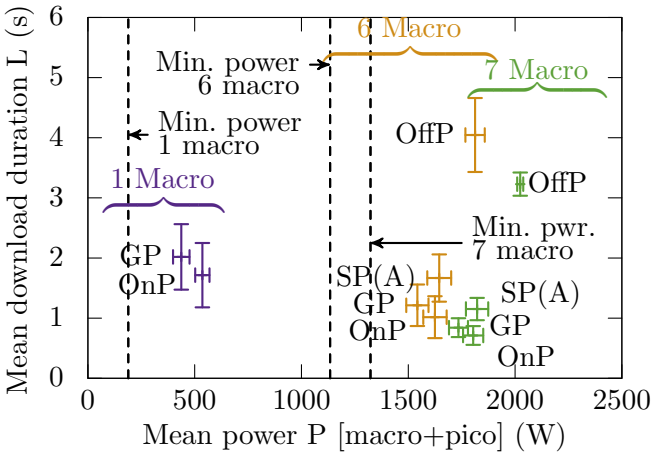

Figure 7.7.: Comparing performance when varying the number of macro BSs; 20% of the average load, DU deployment, 42 pico BSs, pico BSs in hotspot and random deployment, and 1 hotspot per sector. Always-off policy with one macro BS has a mean download duration of about 5 s and is clipped from the plot.

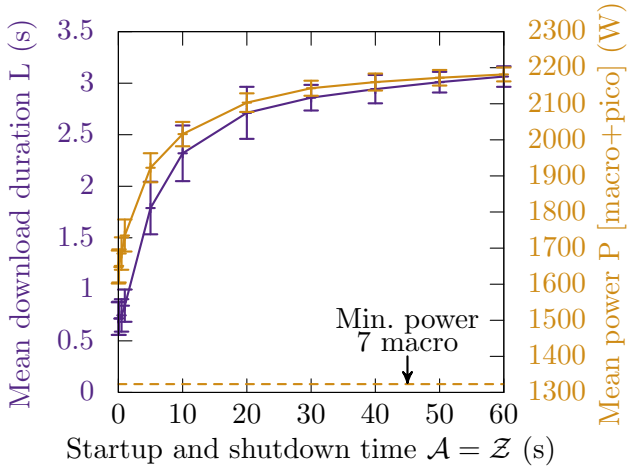

Figure 7.8.: Deactivation times for 20% of the average load in DU deployment, 42 pico BSs, 1 macro BSs, pico BSs in hotspot and random deployment, 2 hotspots, with the greedy policy GP.

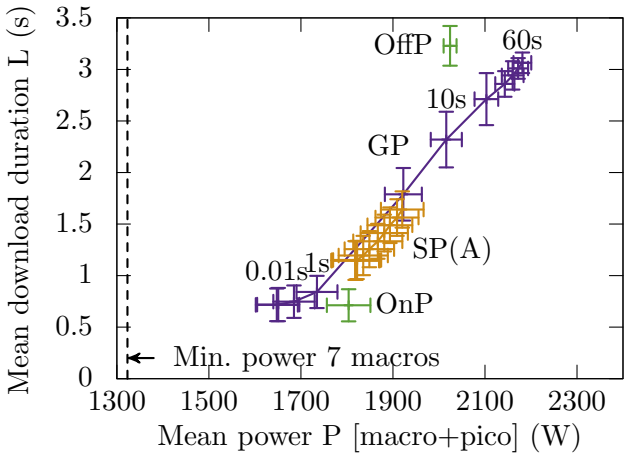

Figure 7.9.: Deactivation durations for 20% of average load, DU, 42 pico BSs, 1 macro BSs, pico BSs in hotspot and random deployment, and 2 hotspots. Durations are set for both activation $\mathcal{A}$ and deactivation durations $\mathcal{Z}$.

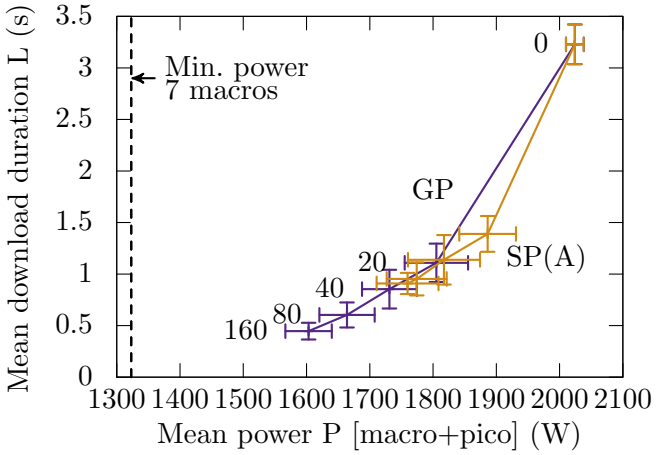

Figure 7.10.: Comparing different number of pico BSs for DU deployment, 1 macro BS, pico BSs deployed in hotspot and random deployment, and 2 hotspots.

with the greedy policy. Figure 7.9 shows the same information, and additionally adds the SP(A), the always-on, and always-off policies as a comparison. The set cover policy based on average arrival rates SP(A) is better suited to cope with high power cycle times than the greedy policy GP, but only for low power-cycle durations does the greedy policy consume less power than the always-on policy. This shows that it is important to determine and reduce the power-cycle durations of future BSs.

Figure 7.10 shows that increasing the number of pico BSs can reduce both the power consumption as well as the latency. This is an interesting effect, but it is necessary to determine if the modeling assumptions still hold in scenarios with a high number of pico BSs and how it interacts with fewer macro BSs.

7.6. Conclusion

I described different theoretically analyzed algorithms to deactivate BSs in RANs and simulated them in a realistic scenario. My results show that statically deactivating and sparser deployment of macro BSs can conserve power and still provide high data rates. An important parameter which describes the trade-off between power and data rate is the power-cycle duration of a BS. Low power cycle times reduce both power consumption and download durations. In contrast to the previous chapter, cooperation does not reduce the power consumption in this chapter as I compare them with strategies with the same deployment of BSs just without cooperation.

I conclude that activating and deactivating BSs can conserve power in realistic scenarios. Deactivating macro BSs reduces the power consumption more than any strategy to deactivate pico BSs. Future work will need to include other scenarios as well as other deployment strategies for pico BSs. Especially interesting are scenarios with a higher number of pico BSs. Assigning UEs to BSs not based on signal-to-noise ratio (SNR) but on energy efficiency is a potential next step. It is also necessary to determine the power-cycle duration and power profiles of future BSs.

8. Final thoughts

In this chapter, I first summarize the content of this dissertation. I then outline which future work is necessary to continue my work to reduce the energy consumption of radio access networks (RANs). In the end, I present a final conclusion.

8.1. Summary

In my dissertation I compared ways to reduce the energy consumption of RANs on the network level. To conserve energy, I considered using fewer base stations (BSs) and deactivating idle BSs. Deactivating idle BSs reduces energy consumption because their power consumption is still considerable when idle. To deactivate BSs, I assumed BSs can be put into a low-power sleep mode.

The large difference in required radio resources between signaling and data traffic motivated the idea of a RAN in which signaling and data traffic are split. Splitting the traffic allows large-range macro BSs to detect the requests of user equipments (UEs) and to activate, on demand, low-range pico BSs, which serve the requests. Hence, the split of signaling and data traffic allows the energy consumption of RANs to be adapted to the load.

In chapter 2 I analyzed an abstract queuing system of a single server with a sleep mode on its own. I quantified the effect of power-cycle durations on energy consumption and latency. Moreover, I analyzed the trade-off between energy consumption and latency for Poisson arrivals and the competitive ratio for worst-case arrivals.

When considering a single server, the latency increases approximately linearly with the power-cycle durations. The energy consumption asymptotically approaches its maximum energy consumption when the power-cycle durations increase. While this is intuitively clear, I provided an analytic derivation for all power-cycle durations. It shows that considering each individual BS with a sleep mode only reduces the energy consumption if the power-cycle durations are low.

In chapter 3 I showed that considering a network as a whole (instead of each device individually) allows the energy consumption to be reduced even if power-cycle durations are high. To illustrate the idea, I used an example of wired networks to express the general idea of network-level energy-saving methods without having to consider the complex effects of wireless transmissions. For the rest of the dissertation, I applied the idea of network-level energy saving to RANs.

Both for signaling and data BSs, increasing the range allows having fewer active BSs per area and thereby conserving energy when the load is low. One technique to increase the range is to use cooperative transmissions from several BSs to a single UE. This allows

the BSs to reach a UE that is not in range of any individual BS. This requires more energy per transmission but allows deactivating of other BSs. Because the energy offset for deactivating a BS under low load is higher than the additionally consumed energy, this will in total conserve energy.

I described how much energy can be conserved in RANs when BSs cooperatively transmit to extend their signaling range in chapter 4. I analyzed the effect for different path-loss exponents and a varying number of cooperating BSs. The analytical results show that the area a BS covers can be significantly increased when cooperative transmissions are introduced. Moreover, scenarios with a low path-loss exponent are best suited for cooperation. I concluded that it is reasonable to let a few BSs cooperate to detect requests of UEs to reduce the overall energy consumption of RANs.

Cooperative transmissions can also be used to reduce the number of active pico BSs which serve the data traffic. I analytically quantified the reduction in activity when pico BSs can cooperate in chapter 5. Reducing the fraction of time BSs are active reduces energy consumption. Placing the BSs at the optimal distance for cooperation additionally reduces energy consumption.

When BSs cooperate to serve data to a UE, more BSs will be transmitting and, thus, create more interference for other UEs. The interference can be dedreased by selecting which BSs actually transmit data to a UE based on instantaneous channel knowledge instead of average channel knowledge. This allows the effective radiated power (ERP) to be at nearly non-cooperative levels and have the outage probability of cooperative transmissions. The greatest gain (both in terms of outage probability and ERP) from using instantaneous channel knowledge is achieved when the average channel gains of all possibly cooperating BSs are the same. From chapter 6 I conclude that using cooperative transmission based on instantaneous channel knowledge nearly removes the drawbacks that cooperative transmissions introduce.

In chapter 7 I developed an event-based simulation of a RAN based on the specifications agreed upon within the mobile working group of the GreenTouch consortium. I used this simulation to analyze the behavior of a RAN when all the effects, which I studied analytically earlier, interact. The policies to conserve energy have no notion of some effects (e.g., power-cycle durations or BS interaction), but have to cope with them in the simulation. The results show that they are able to conserve significant amounts of energy while keeping the quality of service high.

8.2. Future work

The simulation I described in chapter 7 only describes a single scenario (a dense urban environment under low load). While this is the scenario with the greatest potential to conserve energy, it is necessary to consider other scenarios as well. For example in rural areas, where most BSs are needed to cover the area, conserving energy by deactivating BSs is harder. One possibility is to use the cooperative techniques, which I described in chapter 4, to increase the spacing and reduce the energy consumption. My work only showed the general applicability and potential gains, but this needs to be tested in

realistic environments. In general, the simulation should be extended to include signaling traffic and movement of users.

I did not analyze uplink transmissions and interactive traffic. While both of them are not as prevalent as video streaming, RANs have to support them. Because they have stricter requirements a detailed analysis for these types of traffic is needed.

Traffic with less strict requirements (e.g., updates) can be processed when this is possible with less energy. This introduces additional complexity because the UEs and the RANs have different interests but need to coordinate the transmissions anyway.

As both the theoretical analysis (chapter 2) and the simulation (chapter 7) showed, the power-cycle duration is an important parameter to determine the energy consumption of a RAN. Therefore, it is important to have good predictions of the power-cycle durations of future BSs. Also reducing the power-cycle durations will reduce the energy consumption and increase data rates of the RANs. In addition to determining the power-cycle durations, the power profile of future BSs are important. Power profiles which are closer to the linear profile can be combined with BS deactivation. Hence, it is necessary for future work to determine the power profile of future BSs and also reduce it.

To make the simulations more realistic a more detailed user model is needed. Because user distributions are usually only approximated by a Poisson process, a deeper analysis of real user locations is needed. This can for example be achieved by a more detailed model of the hotspots and studies to determine their sizes and traffic in reality. This is also valid for the very simplified model of demands that I used in chapter 7.

Moreover, it is necessary to define the interfaces and protocols which BSs can use to (de-)activate each other. Because potentially many BSs will be deactivated most of the time it is necessary that the activation methods do not consume much energy. This is also valid for the backbone network which the BSs might use to transmit the activation signals.

While cooperative transmission from BSs are part of the Long Term Evolution Advanced (LTE-A) standard, it was not designed to reduce the energy consumption. To use cooperation to reduce the energy consumption it is necessary to allow BSs to go into sleep modes. The protocols to allow this need to be developed.

As I showed in chapter 6, selecting the cooperating BSs based on instantaneous channel conditions enables a reduction of the radiated power. Radiated power becomes interference at other receivers. With less interference, transmissions are finished faster and thus consume less energy. Hence, future work is needed to make sure the transmissions of BSs can quickly adapt to changing channel conditions.

8.3. Conclusion

There is no single method to reduce the energy consumption of RANs to the minimum. A mix of methods to conserve energy on different levels is needed. These include reducing the maximum power consumption of BSs, making the power profile of BSs more linear, adapting the activity of BSs to the load in the network and implementing low-power signaling.

A RAN that is based on a new technology can be both more energy-efficient and provide a higher quality of service (QoS) to the end user. But in a given system there is usually a trade-off between energy consumption and QoS. Also, a system that is very energy-efficient but cannot provide the needed service to the end users is not very useful. Hence, it is not only necessary to understand the behavior of RANs but also what the users demand. Examples are: daily traffic patterns, locations of hotspots, and the mix of voice and data traffic.

A very important factor that will influence both the QoS and energy consumption of RANs is the power-cycle duration. For low power-cycle durations a simple greedy on/off policy provides good results. Deactivating macro BSs during low load is more important than deactivating pico BSs. However, for future BSs neither their power-cycle durations nor their power profiles are known. This information is necessary, if one wants to make informed decisions about methods to conserve power.

A. Proofs for the stretch metrics

In this chapter, I describe the technical details of the proofs referenced in chapter 3. Because all sums, products, maxima, minima, and averages in this chapter are over "$d \in \mathrm{D}$" I will drop them for readability. To further improve readability I drop the "(d)" after x,y and ϕ. I use the following short-hand notations to reduce the number of indexes used in the proofs:

- $x(d) := \mathrm{L_C}(d)$,
- $y(d) := \mathrm{L_{C_L}}(d)$, and
- $\max(x) := \max_{d \in \mathrm{D}}(x(d))$.

To follow the proofs it is helpful to consider x and y as vectors containing the latencies of the demands in the two considered configurations.

A.1. Equality of stretch metrics for the geometric mean

The two metrics S^{GS} and S^{SG} are equal (referred to in section 3.3.3):

$$\begin{aligned} S^{\mathrm{GS}} &= \mathrm{geo}\left(\frac{x}{y}\right) \\ &= \left(\prod (x/y)^{\phi}\right)^{1/\sum \phi} \\ &= \left(\frac{\prod x^{\phi}}{\prod y^{\phi}}\right)^{1/\sum \phi} \\ &= \frac{\left(\prod x^{\phi}\right)^{1/\sum \phi}}{\left(\prod y^{\phi}\right)^{1/\sum \phi}} \\ &= \frac{\mathrm{geo}(x)}{\mathrm{geo}(y)} = S^{\mathrm{SG}}. \end{aligned} \tag{A.1}$$

A.2. Possible orders of metrics

In this section, I present examples using which each possible order of the stretch metrics is achieved. I refer to them in section 3.4.1. I denote the examples for the latencies using the latency-minimizing configuration $\mathrm{L}(\mathrm{C_L})$ and the latencies using energy-conserving configuration $\mathrm{L}(\mathrm{C})$ as $[\mathrm{L}(\mathrm{C_L})] \to [\mathrm{L}(\mathrm{C})]$.

Possible orders with two demands:

- $S^{\mathrm{SM}} > S^{\mathrm{AS}} > S^{\mathrm{SA}}$: $[3, 2] \to [3, 7]$,
- $S^{\mathrm{SM}} > S^{\mathrm{SA}} > S^{\mathrm{AS}}$: $[2, 1] \to [3, 1]$,
- $S^{\mathrm{AS}} > S^{\mathrm{SM}} > S^{\mathrm{SA}}$: $[2, 1] \to [2, 5]$, and
- $S^{\mathrm{AS}} > S^{\mathrm{SA}} > S^{\mathrm{SM}}$: $[2, 1] \to [2, 2]$.

Orders in which $S^{\mathrm{SA}} > S^{\mathrm{SM}}$ and $S^{\mathrm{SA}} > S^{\mathrm{AS}}$ (which include $S^{\mathrm{SA}} > S^{\mathrm{SM}} > S^{\mathrm{AS}}$ and $S^{\mathrm{SA}} > S^{\mathrm{AS}} > S^{\mathrm{SM}}$) are not possible with only two demands. I prove this in the next section.

The last two possible orderings are (with 3 demands):

- $S^{\mathrm{SA}} > S^{\mathrm{SM}} > S^{\mathrm{AS}}$: $[1, 2, 7] \to [1, 3, 9]$ and
- $S^{\mathrm{SA}} > S^{\mathrm{AS}} > S^{\mathrm{SM}}$: $[1, 2, 5] \to [1, 3, 6]$.

A.3. Impossible orders

There is no configurations in which $S^{\mathrm{SA}} > S^{\mathrm{AS}}$ and $S^{\mathrm{SA}} > S^{\mathrm{SM}}$ hold at the same time with exactly two unweighted demands when all values are positive. I refer to this in section 3.4.1. This includes the two orders $S^{\mathrm{SA}} > S^{\mathrm{SM}} > S^{\mathrm{AS}}$ and $S^{\mathrm{SA}} > S^{\mathrm{AS}} > S^{\mathrm{SM}}$.

Proof. Assume a configurations for two demands exists in which $S^{\mathrm{SA}} > S^{\mathrm{AS}}$ and $S^{\mathrm{SA}} > S^{\mathrm{SM}}$ holds. Thus, the two equations

$$S^{\mathrm{SA}} > S^{\mathrm{AS}} \Leftrightarrow \frac{x_1\phi_1 + x_2\phi_2}{y_1\phi_1 + y_2\phi_2} > \frac{x_1\phi_1}{y_1(\phi_1 + \phi_2)} + \frac{x_2\phi_2}{y_2(\phi_1 + \phi_2)} \tag{A.2}$$

and

$$S^{\mathrm{SA}} > S^{\mathrm{SM}} \Leftrightarrow \frac{x_1\phi_1 + x_2\phi_2}{y_1\phi_1 + y_2\phi_2} > \frac{\max(x_1, x_2)}{\max(y_1, y_2)} \tag{A.3}$$

must hold. Without loss of generality I assume $x_1 \geq x_2$ and look at the three cases:

1. $y_1 = y_2$: Simplifying $S^{\mathrm{SA}} > S^{\mathrm{AS}}$ gives $0 > 0$ ↯
2. $y_1 < y_2$: Simplifying $S^{\mathrm{SA}} > S^{\mathrm{AS}}$ gives $x_2\frac{y_1}{y_2} > x_1$, which leads to

$$x_2 \geq x_2\frac{y_1}{y_2} > x_1 \geq x_2 \text{ ↯} \tag{A.4}$$

3. $y_1 > y_2$:
   a) Simplifying $S^{\mathrm{SA}} > S^{\mathrm{AS}}$ gives $x_2\frac{y_1}{y_2} < x_1$ and
   b) Simplifying $S^{\mathrm{SA}} > S^{\mathrm{SM}}$ gives $x_2\frac{y_1}{y_2} > x_1$ ↯

□

A.4. Bounds between metrics

In this section, I prove that the bound

$$A \leq \text{skew(C)} \cdot \text{skew}(\text{C}_\text{L}) \cdot B \tag{A.5}$$

holds for all combinations of A and B from the five metrics. I improve the bound where possible. I refer to the bounds in section 3.4.1. The following bounds hold for strictly positive values (which is a useful assumption for latencies):

$$\frac{\max(x)}{\max(y)} \leq \max\left(\frac{x}{y}\right) \tag{A.6}$$

$$\frac{\min(x)}{\min(y)} \leq \max\left(\frac{x}{y}\right) \tag{A.7}$$

$$\max\left(\frac{x}{y}\right) \leq \frac{\max(x)}{\min(y)} \tag{A.8}$$

$$\frac{\min(x)}{\max(y)} \leq \min\left(\frac{x}{y}\right). \tag{A.9}$$

The following proofs either use transitivity to directly show the result or are proofs by contradiction. The proofs of contradiction start by assuming the opposite is true and construct an inequality chain from it which leads to a contradiction.

A.4.1. Maximum of stretches

I start by showing that the maximum of the stretches S^{MS} is always higher than the other four metrics.

Proof of $S^{\textsf{SM}} \leq S^{\textsf{MS}}$, Assume: $S^{\text{MS}} < S^{\text{SM}}$

$$\frac{\max(x)}{\max(y)} \overbrace{\leq}^{A.6} \max\left(\frac{x}{y}\right) < \frac{\max(x)}{\max(y)} \lightning \tag{A.10}$$

Proof of $S^{\textsf{AS}} \leq S^{\textsf{MS}}$, Assume: $S^{\text{MS}} < S^{\text{AS}}$ and define $z := \max\left(\frac{x}{y}\right)$

$$z = \max\left(\frac{x}{y}\right) < \frac{\sum \frac{x\phi}{y}}{\sum \phi} \leq \frac{\sum z\phi}{\sum \phi} = \frac{z\sum \phi}{\sum \phi} = z \lightning \tag{A.11}$$

Proof of $S^{\textsf{SA}} \leq S^{\textsf{MS}}$, Assume: $S^{\text{MS}} < S^{\text{SA}}$ and define $z := \max\left(\frac{x}{y}\right)$

$$z = \max\left(\frac{x}{y}\right) < \frac{\sum x\phi}{\sum y\phi} \overbrace{\leq}^{x \leq zy} \frac{\sum zy\phi}{\sum y\phi} = \frac{z\sum y\phi}{\sum y\phi} = z \lightning \tag{A.12}$$

Note that this one half of Cauchy's Third Inequality [Ste04].

Proof of $S^{\mathsf{GS}} \leq S^{\mathsf{MS}}$, Transitivity of $S^{\mathrm{GS}} \leq S^{\mathrm{AS}} \leq S^{\mathrm{MS}}$

A.4.2. Stretch of maximum

Now I show that $A \leq \mathrm{skew}(y) S^{\mathrm{SM}}$ holds for any of the metrics A I described.

Proof of $S^{\mathsf{MS}} \leq \mathsf{skew}(y) S^{\mathsf{SM}}$, Assume: $\mathrm{skew}(y) S^{\mathrm{SM}} < S^{\mathrm{MS}}$

$$\frac{\max(x)}{\min(y)} = \overbrace{\frac{\max(y)}{\min(y)}}^{\mathrm{skew}(y)} \overbrace{\frac{\max(x)}{\max(y)}}^{S^{\mathrm{SM}}} < \max\left(\frac{x}{y}\right) \overbrace{\leq}^{A.8} \frac{\max(x)}{\min(y)} \lightning \tag{A.13}$$

Proof of $S^{\mathsf{SA}} \leq \mathsf{skew}(y) S^{\mathsf{SM}}$, Transitivity of $S^{\mathrm{SA}} \leq S^{\mathrm{MS}} \leq \mathrm{skew}(y) S^{\mathrm{SM}}$

Proof of $S^{\mathsf{AS}} \leq \mathsf{skew}(y) S^{\mathsf{SM}}$, Transitivity of $S^{\mathrm{AS}} \leq S^{\mathrm{MS}} \leq \mathrm{skew}(y) S^{\mathrm{SM}}$

Proof of $S^{\mathsf{GS}} \leq \mathsf{skew}(y) S^{\mathsf{SM}}$, Transitivity of $S^{\mathrm{GS}} \leq S^{\mathrm{AS}} \leq \mathrm{skew}(y) S^{\mathrm{SM}}$

A.4.3. Stretch of average

Now I show the inequalities with S^{SA} on the right side.

Proof of $S^{\mathsf{MS}} \leq \mathsf{skew}(x)\,\mathsf{skew}(y) S^{\mathsf{SA}}$, Assume: $\mathrm{skew}(x)\,\mathrm{skew}(y) S^{\mathrm{SA}} < S^{\mathrm{MS}}$

$$\begin{aligned} \frac{\max(x)}{\min(y)} = \frac{\max(x)}{\min(x)} \frac{\max(y)}{\min(y)} \frac{\sum \min(x)\phi}{\sum \max(y)\phi} &\leq \frac{\max(x)}{\min(x)} \frac{\max(y)}{\min(y)} \frac{\sum x\phi}{\sum y\phi} \\ &< \max\left(\frac{x}{y}\right) \overbrace{\leq}^{A.8} \frac{\max(x)}{\min(y)} \lightning \end{aligned} \tag{A.14}$$

Proof of $S^{\mathsf{SM}} \leq \mathsf{skew}(x) S^{\mathsf{SA}}$, Assume: $\mathrm{skew}(x) S^{\mathrm{SA}} < S^{\mathrm{SM}}$

$$\frac{\max(x)}{\max(y)} = \frac{\max(x)}{\min(x)} \frac{\sum \phi \min(x)}{\sum \phi \max(y)} \leq \frac{\max(x)}{\min(x)} \frac{\sum \phi x}{\sum \phi y} < \frac{\max(x)}{\max(y)} \lightning \tag{A.15}$$

Proof of $S^{\mathsf{AS}} \leq \mathsf{skew}(x)\,\mathsf{skew}(y) S^{\mathsf{SA}}$, Transitivity of $S^{\mathrm{AS}} \leq S^{\mathrm{MS}} \leq \mathrm{skew}(x)\,\mathrm{skew}(y) S^{\mathrm{SA}}$

Proof of $S^{\mathsf{GS}} \leq \mathsf{skew}(x)\,\mathsf{skew}(y) S^{\mathsf{SA}}$, Transitivity of $S^{\mathrm{GS}} \leq S^{\mathrm{AS}} \leq \mathrm{skew}(x)\,\mathrm{skew}(y) S^{\mathrm{SA}}$

A.4.4. Average of stretches

Now I show the inequalities with S^{AS} on the right side.

Proof of $S^{\mathsf{MS}} \leq \mathsf{skew}(x)\,\mathsf{skew}(y)S^{\mathsf{AS}}$, Assume: $\operatorname{skew}(x)\operatorname{skew}(y)S^{\mathrm{AS}} < S^{\mathrm{MS}}$

$$\begin{aligned}\frac{\max(x)}{\min(y)} \overbrace{\leq}^{A.9} \frac{\max(x)}{\min(x)}\frac{\max(y)}{\min(y)}\min\left(\frac{x}{y}\right) &\leq \frac{\max(x)}{\min(x)}\frac{\max(y)}{\min(y)}\frac{\sum x\phi/y}{\sum \phi} \\ &< \max\left(\frac{x}{y}\right) \overbrace{\leq}^{A.8} \frac{\max(x)}{\min(y)} \lightning \end{aligned} \tag{A.16}$$

Proof of $S^{\mathsf{SM}} \leq \mathsf{skew}(x)S^{\mathsf{AS}}$, Assume: $\operatorname{skew}(x)S^{\mathrm{AS}} < S^{\mathrm{SM}}$

$$\frac{\max(x)}{\max(y)} \overbrace{\leq}^{A.9} \frac{\max(x)}{\min(x)}\min\left(\frac{x}{y}\right) \leq \frac{\max(x)}{\min(x)}\frac{\sum \phi x/y}{\sum \phi} < \frac{\max(x)}{\max(y)} \lightning \tag{A.17}$$

Proof of $S^{\mathsf{SA}} \leq \mathsf{skew}(x)\,\mathsf{skew}(y)S^{\mathsf{AS}}$, Transitivity of $S^{\mathrm{SA}} \leq S^{\mathrm{MS}} \leq \operatorname{skew}(x)\operatorname{skew}(y)S^{\mathrm{AS}}$

Proof of $S^{\mathsf{GS}} \leq S^{\mathsf{AS}}$, Generalized inequality of arithmetic and geometric means to weighted means [Ste04].

A.4.5. Geometric mean of stretches

Now I show the inequalities with S^{GS} on the right side.

Proof of $S^{\mathsf{MS}} \leq \mathsf{skew}(x)\,\mathsf{skew}(y)S^{\mathsf{GS}}$, Assume: $\operatorname{skew}(x)\operatorname{skew}(y)S^{\mathrm{GS}} < S^{\mathrm{MS}}$

$$\begin{aligned}\frac{\max(x)}{\min(y)} \overbrace{\leq}^{A.9} \frac{\max(x)}{\min(x)}\frac{\max(y)}{\min(y)}\min\left(\frac{x}{y}\right) &\leq \frac{\max(x)}{\min(x)}\frac{\max(y)}{\min(y)}\left(\prod\left(\frac{x}{y}\right)^{\phi}\right)^{\frac{1}{\sum\phi}} \\ &< \max\left(\frac{x}{y}\right) \overbrace{\leq}^{A.8} \frac{\max(x)}{\min(y)} \end{aligned} \tag{A.18}$$

Proof of $S^{\mathsf{SM}} \leq \mathsf{skew}(x)\,\mathsf{skew}(y)S^{\mathsf{GS}}$, Transitivity of $S^{\mathrm{SM}} \leq S^{\mathrm{MS}} \leq \operatorname{skew}(x)\operatorname{skew}(y)S^{\mathrm{GS}}$

Proof of $S^{\mathsf{AS}} \leq \mathsf{skew}(x)\,\mathsf{skew}(y)S^{\mathsf{GS}}$, Transitivity of $S^{\mathrm{AS}} \leq S^{\mathrm{MS}} \leq \operatorname{skew}(x)\operatorname{skew}(y)S^{\mathrm{GS}}$

Proof of $S^{\mathsf{SA}} \leq \mathsf{skew}(x)\,\mathsf{skew}(y)S^{\mathsf{GS}}$, Transitivity of $S^{\mathrm{SA}} \leq S^{\mathrm{MS}} \leq \operatorname{skew}(x)\operatorname{skew}(y)S^{\mathrm{GS}}$

A.5. Latency in rings

The following holds in an n-circle when an edge is deactivated:

$$\lim_{n\to\infty} S^{\mathrm{GS}} = \frac{2}{\sqrt{\mathrm{e}}}. \tag{A.19}$$

This is approximately 1.21. I refer to this in section 3.3.3.

A.5.1. Limit in odd-length rings

Proof. First consider the case that n is odd. In the latency-minimizing configurations $\mathrm{C_L}$ (the circle) the geometric mean of the latencies is:

$$\mathrm{geo}(\mathrm{L}_{\mathrm{C_L}}) = \sqrt[n(n-1)]{\left(\prod_{i=1}^{(n-1)/2} i\right)^{2n}}. \tag{A.20}$$

And in the energy-minimizing configurations C (the path) it is:

$$\mathrm{geo}(\mathrm{L_C}) = \sqrt[n(n-1)]{\prod_{s=1}^{n}\prod_{d=1,d\neq s}^{n} |d-s|}. \tag{A.21}$$

Thus, I need to determine:

$$\lim_{n\to\infty} S^{\mathrm{GS}} = \lim_{n\to\infty} \sqrt[n(n-1)]{\frac{\prod_{s=1}^{n}\prod_{d=1,d\neq s}^{n} |d-s|}{\left(\prod_{i=1}^{(n-1)/2} i\right)^{2n}}} = \lim_{n\to\infty} \left(\frac{\prod_{s=1}^{n-1} s!}{(((n-1)/2)!)^n}\right)^{\frac{2}{n(n-1)}}. \tag{A.22}$$

I replace n with $2m+1$ and use Stirling's approximation $n! \approx \left(\frac{n}{\mathrm{e}}\right)^n$. I only need to consider terms that do not approach 1 after the exponentiation with $2/(2m(2m+1))$.

$$\lim_{m\to\infty} \left(\frac{\prod_{s=1}^{2m} s!}{(m!)^{2m+1}}\right)^{\frac{2}{2m(2m+1)}} = \lim_{m\to\infty} \left(\frac{\prod_{s=1}^{2m} \left(\frac{s}{\mathrm{e}}\right)^s}{\left(\frac{m}{\mathrm{e}}\right)^{m(2m+1)}}\right)^{\frac{2}{2m(2m+1)}} \tag{A.23}$$

Using $\prod_{s=1}^{2m} \mathrm{e}^{-s} = \mathrm{e}^{-m(2m+1)}$ I get

$$\lim_{m\to\infty} \left(m^{-m(2m+1)}\mathrm{e}^{m(2m+1)}\mathrm{e}^{-m(2m+1)} \prod_{s=1}^{2m} s^s\right)^{\frac{2}{2m(2m+1)}}. \tag{A.24}$$

A Stirling-like series for the hyperfactorial [Weib] is $\prod_{i=1}^{n} i^i \approx \mathrm{e}^{-n^2/4} n^{n(n+1)/2}$, where again all terms have been dropped that approach 1 in the end result.

$$\lim_{m\to\infty} \left(m^{-m(2m+1)}\mathrm{e}^{-(2m)^2/4}(2m)^{2m(2m+1)/2}\right)^{\frac{2}{2m(2m+1)}} = \frac{2}{\sqrt{\mathrm{e}}} \tag{A.25}$$

□

A.5.2. Limit in even-length rings

Proof. In case of even length, the formula for the path is the same, but the formula for the circle is

$$\sqrt[n(n-1)]{\left(\left(\prod_{i=1}^{n/2-1} i\right)^2 \frac{n}{2}\right)^n}. \tag{A.26}$$

Replacing n with $2m$ and inserting Sterling's approximation I get

$$\lim_{m\to\infty}\left(\frac{\prod_{s=1}^{2m-1} s!^2}{\left((m-1)!^2 m\right)^{2m}}\right)^{\frac{1}{2m(2m-1)}} = \lim_{m\to\infty}\left(\frac{\prod_{s=1}^{2m-1}\left(\frac{s}{\mathrm{e}}\right)^s}{\left(\frac{m-1}{\mathrm{e}}\right)^{(m-1)2m}}\right)^{\frac{2}{2m(2m-1)}}. \tag{A.27}$$

Using $\prod_{s=1}^{2m-1}\mathrm{e}^{-s} = \mathrm{e}^{m-2m^2}$ I get

$$\lim_{m\to\infty}\left((m-1)^{-(m-1)2m}\mathrm{e}^{(m-1)2m}\mathrm{e}^{m-2m^2}\prod_{s=1}^{2m-1} s^s\right)^{\frac{2}{2m(2m-1)}}. \tag{A.28}$$

Using the same Stirling-like approximation of the hyperfactorial the result is:

$$\lim_{m\to\infty}\left((m-1)^{-(m-1)2m}\mathrm{e}^{(m-1)2m}\mathrm{e}^{m-2m^2}\mathrm{e}^{-(2m-1)^2/4}(2m-1)^{(2m-1)2m/2}\right)^{\frac{2}{2m(2m-1)}} = \frac{2}{\sqrt{\mathrm{e}}}. \tag{A.29}$$

□

B. Bibliography

[3GP10] 3GPP. Evolved Universal Terrestrial Radio Access (E-UTRA); Further advancements for E-UTRA physical layer, 2010.

[ABH11] I. Ashraf, F. Boccardi, and L. Ho. Sleep Mode Techniques for Small Cell Deployments. *IEEE Communications Magazine*, (August):72–79, 2011.

[ADD+93] I. Althöfer, G. Das, D. Dobkin, D. Joseph, and J. Soares. On Sparse Spanners of Weighted Graphs. *Journal of Discrete & Computational Geometry*, 9(1):81–100, 1993.

[AGD+11] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. Imran, D. Sabella, M. Gonzalez, O. Blume, and A. Fehske. How Much Energy is Needed to Run a Wireless Network? *IEEE Wireless Communications Magazine*, 18(5):40–49, October 2011.

[AGG11] G. Auer, V. Giannini, and I. Godor. Cellular Energy Efficiency Evaluation Framework. *Proceedings of the IEEE Vehicular Technology Conference Spring*, (June):1–6, 2011.

[AK04] P. Anghel and M. Kaveh. Exact Symbol Error Probability of a Cooperative Network in a Rayleigh-Fading Environment. *IEEE Transactions of Wireless Communications*, 3(5):1416–1421, September 2004.

[Alb09] S. Albers. Algorithms for Energy Saving. *Efficient Algorithms*, pages 173–186, 2009.

[ARFB10] O. Arnold, F. Richter, G. Fettweis, and O. Blume. Power Consumption Modeling of Different Base Station types in Heterogeneous Cellular Networks. *Proceedings of the Future Network & Mobile Summit*, pages 1–8, 2010.

[Aus12] Australien Energy Market Operator. *Economic Outlook Information Paper*. 2012.

[AWT09] L. Andrew, A. Wierman, and A. Tang. Optimal Speed Scaling under Arbitrary Power Functions. *ACM SIGMETRICS Performance Evaluation Review*, 37(2):39–41, October 2009.

[BAWB13] O. Blume, A. Ambrosy, M. Wilhelm, and U. Barth. Energy Efficiency of LTE Networks Under Traffic Loads of 2020. In *Proceedings of the International Symposium on Wireless Communication Systems*, 2013.

[BB09] F. Baccelli and B. Blaszczyszyn. *Stochastic Geometry and Wireless Networks*. 2009.

[BBEE08] A. Beck, S. Borst, B. Ensor, and J. Esteban. New Challenges in Content Dissemination Networks. *Bell Labs Technical Journal*, 13(3):5–12, 2008.

[BBPD99] L. Benini, A. Bogliolo, G. Paleologo, and G. De Micheli. Policy Optimization for Dynamic Power Management. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 18(6):813–833, June 1999.

[BCD07] P. Baptiste, M. Chrobak, and C. Dürr. Polynomial Time Algorithms for Minimum Energy Scheduling. In *Proceedings of the European Conference on Algorithms*, pages 136–150. Springer-Verlag, 2007.

[BEK+02] P. Bohrer, E.N. Elnozahy, T. Keller, M. Kistler, C. Lefurgy, C. McDowell, and R. Rajamony. The Case for Power Management in Web Servers. In *Power Aware Computing*, volume 62. 2002.

[BEK+10] O. Blume, H. Eckhardt, S. Klein, E. Kuehn, and W. Wieslawa. Energy Savings in Mobile Networks Based on Adaptation to Traffic Statistics. *Bell Labs Technical*, 15(2):77–94, 2010.

[BEY05] A. Borodin and R. El-Yaniv. *Online Computation and Competitive Analysis*. Cambridge University Press, 2005.

[BGN+10] J. Berral, Í. Goiri, R. Nou, F. Julià, J. Guitart, R. Gavaldà, and J. Torres. Towards Energy-Aware Scheduling in Data Centers using Machine Learning. *Proceedings of the International Conference on Energy-Efficient Computing and Networking*, 2:215, 2010.

[BK97] A. Bezdek and W. Kuperberg. Circle Covering with a Margin. *Periodica Mathematica Hungarica*, 34(1):3–16, 1997.

[BK12] K. Balachandran and J. Kang. An Analysis of Uplink Base Station Cooperation with Practical Constraints. *IEEE Transactions on Wireless Communication*, 11(3):1056–1065, March 2012.

[Blu10] W. Blundon. Multiple Covering of the Plane by Circles. *Mathematika*, 4(01):7–16, February 2010.

[Bol98] G. Bolch. *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. Wiley-Interscience Publication. Wiley, 1998.

[BPV08] R. Buyya, M. Pathan, and A. Vakali. Content Delivery Networks. *Lecture Notes In Electrical Engineering*, 49(1):418, 2008.

[Bre13] N. Bredenbals. *Energy-Efficient Queuing with Delayed Activation and Deactivation*. Master's thesis, Universität Paderborn, 2013.

[BSC+12] T. Biermann, L. Scalia, C. Choi, H. Karl, and W. Kellerer. CoMP Clustering and Backhaul Limitations in Cooperative Cellular Mobile Access Networks. *Journal of Pervasive and Mobile Computing*, 8(5):662–681, March 2012.

[BSW97] M. Bernstein, N. Sloane, and P. Wright. On Sublattices of the Hexagonal Lattice. *Discrete Mathematics*, 170:29–39, 1997.

[BZB10] O. Blume, D. Zeller, and U. Barth. Approaches to Energy Efficient Wireless Access Networks. *Proceedings of the International Symposium on Communications, Control and Signal Processing*, (March):2–6, 2010.

[CA11] A. Chorppath and T. Alpcan. Mechanism Design for Energy Efficiency in Wireless Networks. In *International Symposium of Modeling and Optimization of Mobile, Ad Hoc, and Wireless Networks*, pages 389–394, May 2011.

[CC95] L. Cai and D. Corneil. Tree Spanners. *SIAM Journal on Discrete Mathematics*, 8(3):359, 1995.

[CCLL07] H. Chan, W. Chan, T. Lam, and L. Lee. Energy Efficient Online Deadline Scheduling. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, pages 795–804, 2007.

[CFC11] A. Conte, A. Feki, and L. Chiaraviglio. Cell Wilting and Blossoming for Energy Efficiency. *IEEE Wireless Communications Magazine*, 18(5):50–57, 2011.

[CH08] A. Corliano and M. Hufschmid. Energieverbrauch der Mobilen Kommunikation. *Bundesamt für Energie, Ittigen, Switzerland, Techical Report*, 2008.

[CHS08] H. Claussen, L. Ho, and L. Samual. An Overview of the Femtocell Concept. *Bell Labs Technical Journal*, 13(1):221–245, 2008.

[Cis13] Cisco. Cisco Visual Networking Index: Forecast and Methodology , 2012 – 2017. 2013.

[CLMW07] J. Chan, T. Lam, K. Mak, and P. Wong. Online Deadline Scheduling with Bounded Energy Efficiency. *Proceedings of International Conference Theory and Applications of Models of Computation*, pages 416–427, 2007.

[CLP09] S. Chechik, M. Langberg, and D. Peleg. Fault-Tolerant Spanners for General Graphs. *SIAM Journal on Computing*, 39(7):3403–3423, 2009.

[CMN09a] L. Chiaraviglio, M. Mellia, and F. Neri. Energy-Aware Backbone Networks: A Case Study. *Proceedings of the IEEE International Conference on Communications Workshops*, 2009.

[CMN09b] L. Chiaraviglio, M. Mellia, and F. Neri. Reducing Power Consumption in Backbone Networks. *Proceedings of the IEEE International Conference on Communications*, 2009.

[CMP09] C. Comaniciu, N. Mandayam, and H. Poor. Radio Resource Management for Green Wireless Networks. *Proceedings of the IEEE Vehicular Technology Conference Fall*, September 2009.

[CSB+08] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang, and S. Wright. Power Awareness in Network Design and Routing. In *Proceedings of the IEEE Conference on Computer Communications*, 2008.

[CSF12] A. Capone, A. Santos, and I. Filippini. Looking Beyond Green Cellular Networks. *Proceedings of the Conference on Wireless On-demand Network Systems and Services*, pages 127–130, 2012.

[CX07] Y. Chen and F. Xia. Stochastic Modeling of Dynamic Power Management Policies and Analysis of Their Power-Latency Tradeoffs. *Technical Report Series NCL-EECE-MSD-TR-2007-123*, 2007.

[CXSY09] Y. Chen, F. Xia, D. Shang, and A. Yakovlev. Fine-grain Stochastic Modelling of Dynamic Power Management Policies and Analysis of Their Power–Latency Tradeoffs. *IET Software*, 3(6):458, 2009.

[CXY+03] S. Cai, L. Xiao, H. Yang, J. Wang, and S. Zhou. A Cross-Layer Optimization of the Joint Macro-and Picocell Deployment with Sleep Mode for Green Communications. *Tsinghua University Initiative Science Research Program*, 2003.

[CYZK11] T. Chen, Y. Yang, H. Zhang, and H. Kim. Network Energy Saving Technologies for Green Wireless Access Networks. *IEEE Wireless Communications*, 18(5), 2011.

[CZB10] L. Correia, D. Zeller, and O. Blume. Challenges and Enabling Technologies for Energy Aware Mobile Radio Networks. *IEEE Communications Magazine*, 48(11):66–72, 2010.

[CZXL11] Y. Chen, S. Zhang, S. Xu, and G. Li. Fundamental Trade-Offs on Green Wireless Networks. *IEEE Communications Magazine*, 49(6):30–37, June 2011.

[DD08] T. Drezner and Z. Drezner. Locating Base Stations for Mobile Servers. *Proceedings of the Americas Conference on Information Systems*, page 209, 2008.

[De 01] G. De Micheli. Comparing System Level Power Management Policies. *IEEE Design & Test of Computers*, 18(2):10–19, 2001.

[DE00] P. Dankelmann and R. Entringer. Average Distance, Minimum Degree, and Spanning Trees. *Journal of Graph Theory*, 33(1):1–13, January 2000.

[DG77] J. Doyle and J. Graver. Mean Distance in a Graph. *Discrete Mathematics*, 1977.

[Eri11] M. Ericson. Total Network Base Station Energy Cost vs. Deployment. *Proceedings of the IEEE Vehicular Technology Conference Spring*, May 2011.

[Far10] H. Farhangi. The Path of the Smart Grid. *IEEE Power and Energy Magazine*, 2010.

[FDM+09] C. Forster, I. Dickie, G. Maile, H. Smith, and M. Crisp. Understanding the Environmental Impact of Communication Systems. *Ofcom final report*, (April), 2009.

[Few06] M. Fewell. Area of Common Overlap of Three Circles. *Defence Science and Technology Organsiation, Edinburgh (Australia) Maritim Operation Division*, 2006.

[FFMB11] A. Fehske, G. Fettweis, J. Malmodin, and G. Biczok. The Global Footprint of Mobile Communications: The Ecological and Economic Perspective. *IEEE Communications Magazine*, 49(8):55–62, August 2011.

[FHG13] L. Falconetti, L. Hévizi, and I. Gódor. Sleep Mode Control for Low Power Nodes in Heterogeneous Networks. *Proceedings of the International Symposium on Wireless Communication Systems*, pages 321–325, 2013.

[Fis07] G. Fischer. Next-generation Base Station Radio Frequency Architecture. *Bell Labs Technical Journal*, 12(2):3–18, 2007.

[FMF10] A. Fehske, P. Marsch, and G. Fettweis. Bit per Joule Efficiency of Cooperating Base Stations in Cellular Networks. *Proceedings of the IEEE Globecom Workshops*, pages 1406–1411, December 2010.

[FMK+10] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin. Diversity in Smartphone Usage. *Proceedings of the International Conference on Mobile Systems, Applications, and Services*, page 179, 2010.

[FMXY12] X. Fang, S. Misra, G. Xue, and D. Yang. Smart Grid — The New and Improved Power Grid: A Survey. *IEEE Communications Surveys & Tutorials*, 14(4):944–980, 2012.

[FN04] P. Fryzlewicz and G. Nason. A Haar-Fisz Algorithm for Poisson Intensity Estimation. *Journal of Computational and Graphical Statistics*, 13(3):621–638, September 2004.

[FRKR05] W. Felter, K. Rajamani, T. Keller, and C. Rusu. A Performance-Conserving Approach for Reducing Peak Power Consumption in Server Systems. *Proceedings of the International Conference on Supercomputing*, page 293, 2005.

[FS07] H. Frey and D. Simplot. Localized Topology Control Algorithms for Ad Hoc and Sensor Networks. In *Handbook of Applied Algorithms: Solving Scientific, Engineering, and Practical Problems*. John Wiley and Sons, 2007.

[FSC11] R. Fantini, D. Sabella, and M. Caretti. Energy Efficiency in LTE-Advanced Networks with Relay Nodes. *Proceedings of the IEEE Vehicular Technology Conference Spring*, pages 1–5, May 2011.

[GAV11] M. Goldenbaum, R. Akl, and S. Valentin. On the Effect of Feedback Delay in the Downlink of Multiuser OFDM Systems. In *Proceedings of the Conference on Information Sciences and Systems*, 2011.

[GBGF11] P. Gonzalez-Brevis, J. Gondzio, and Y. Fan. Base Station Location Optimization for Minimal Energy Consumption in Wireless Networks. *Proceedings of the IEEE Vehicular Technology Conference Spring*, pages 1–5, 2011.

[GBW95] R. Golding, P. Bosch, and J. Wilkes. Idleness is Not Sloth. *Proceedings of the USENIX Winter Conference*, 1995.

[Gol05] A. Goldsmith. *Wireless Communications*. Cambridge University Press, 2005.

[HA11] M. Hedayati and M. Amirijoo. Reducing Energy Consumption through Adaptation of Number of Active Radio Units. *Proceedings of the IEEE Vehicular Technology Conference Spring*, pages 1–5, May 2011.

[HA13] T. Han and N. Ansari. On Greening Cellular Networks via Multicell Cooperation. *IEEE Wireless Communications*, (February):82–89, 2013.

[HAH11] H. Holtkamp, G. Auer, and H. Haas. On Minimizing Base Station Power Consumption. In *Proceedings of the IEEE Vehicular Technology Conference Fall*, pages 1–5, September 2011.

[HB11] Z. Hasan and H. Boostanimehr. Green Cellular Networks: A Survey, Some Research Issues and Challenges. *IEEE Communications Surveys & Tutorials*, 13(4):524–540, 2011.

[HCM12] C. Hoymann, W. Chen, and J. Montojo. Relaying Operation in 3GPP LTE: Challenges and Solutions. *IEEE Communications Magazine*, 50(2):156–162, February 2012.

[Hey69] D. Heyman. Optimal Operating Policies for M/G/1 Queueing Systems. *Annals of Physics*, 54(2):362–382, 1969.

[HHA+11] C. Han, T. Harrold, S. Armour, I. Krikidis, S. Videv, P. Grant, H. Haas, J. Thompson, I. Ku, C. Wang, T. Le, M. Nakhai, J. Zhang, and L. Hanzo. Green Radio: Radio Techniques to Enable Energy-Efficient Wireless Networks. *IEEE Communications Magazine*, 49(6):46–54, June 2011.

[HHK] M. Herlich, T. Hohenberger, and H. Karl. Activation Strategies for Low-Power Radio Access Networks. *In preparation.*

[HHK13] T. Hohenberger, M. Herlich, and H. Karl. Trade-Off between Latency and Coverage in Cooperative Radio Access Networks. In *Proceedings of the International Conference on Advanced Networks and Telecommunication Systems*, 2013.

[HK] M. Herlich and H. Karl. Analytic Quantification of Outage Probability and Radiated Power of Cooperative Base Stations. *In preparation.*

[HK11a] M. Herlich and H. Karl. Reducing Power Consumption of Mobile Access Networks with Cooperation. In *Proceedings of the International Conference on Energy-Efficient Computing and Networking*, pages 77–86, New York, USA, 2011. ACM.

[HK11b] M. Herlich and H. Karl. The Trade-Off between Power Consumption and Latency in Computer Networks. In Vicente Casares-Giner, Pietro Manzoni, and Ana Pont, editors, *Proceedings of the Networking Workshops*, volume 6827 of *Lecture Notes in Computer Science*, pages 273–280. Springer Berlin / Heidelberg, 2011.

[HK12] M. Herlich and H. Karl. Average and Competitive Analysis of Latency and Power consumption of a Queuing System with a Sleep Mode. In *Proceedings of the International Conference on Future Energy Systems: Where Energy, Computing and Communication Meet*, pages 14:1—-14:10, New York, NY, USA, 2012. ACM.

[HK13] M. Herlich and H. Karl. Energy-Efficient Assignment of User Equipment to Cooperative Base Stations. In *Proceedings of the International Symposium on Wireless Communication Systems*, 2013.

[HMJ11] M. Hossain, K. Munasinghe, and A. Jamalipour. An Eco-Inspired Energy Efficient Access Network Architecture for Next Generation Cellular Systems. *IEEE Wireless Communications and Networking Conference*, pages 992–997, March 2011.

[HNS09] M. Hewitt, G. L. Nemhauser, and M. W. P. Savelsbergh. Combining Exact and Heuristic Approaches for the Capacitated Fixed-Charge Network Flow Problem. *INFORMS Journal on Computing*, 22(2):314–325, September 2009.

[Hoh12] T. Hohenberger. *Queuing Latency at Cooperative Base Stations.* Bachelor's thesis, Universität Paderborn, 2012.

[Hoy11] J. Hoydis. Optimal Channel Training in Uplink Network MIMO systems. *IEEE Transactions on Signal Processing*, 59(6):2824–2833, 2011.

[HS89] D. Hochbaum and A. Segev. Analysis of a Flow Problem with Fixed Charges. *International Journal of Networks*, 19(3):291–312, May 1989.

[HYT05] S. Hashimoto, T. Yachi, and T. Tani. A New Stand-Alone Hybrid Power System with Wind Turbine Generator and Photovoltaic Modules for a Small-Scale Radio Base Station. *IEEJ Transactions on Power and Energy*, 125(11):1041–1046, 2005.

[IP05] S. Irani and K.R. Pruhs. Algorithmic Problems in Power Management. *ACM SIGACT Newsletter*, 36(2):63–76, 2005.

[ISG02] S. Irani, S. Shukla, and R. Gupta. Competitive Analysis of Dynamic Power Management Strategies for Systems with Multiple Power Saving States. *Proceedings of the Design, Automation and Test in Europe Conference and Exhibition*, pages 117–123, 2002.

[Ism11] M. Ismail. Network Cooperation for Energy Saving in Green Radio Communications. *IEEE Wireless Communications Magazine*, (October):76–81, 2011.

[Ivi03] A. Ivić. *The Riemann Zeta-Function: Theory and Applications.* Dover Books on Mathematics. Dover, 2003.

[JLR78] D. Johnson, J. Lenstra, and A. Rinnooy Kan. The Complexity of the Network Design Problem. *International Journal of Networks*, 8(4):279–285, 1978.

[Joh82] D. Johnson. The NP-Completeness Column: An Ongoing Guide. *Journal of Algorithms*, 3(2):182–195, 1982.

[Kah11] D. Kahneman. *Thinking, Fast and Slow.* Farrar, Straus and Giroux, 2011.

[KAK10] D. Kutscher, B. Ahlgren, and H. Karl. Information-Centric Networking. In *Dagstuhl Seminar*, pages 1–17, 2010.

[KAK+11] D. Kilper, G. Atkinson, S. Korotky, S. Goyal, P. Vetter, D. Suvakovic, and O. Blume. Power Trends in Communication Networks. 17(2):275–284, 2011.

[KCLMT12] P. Kling, A. Cord-Landwehr, and F. Mallmann-Trenn. Slow Down and Sleep for Profit in Online Deadline Scheduling. *Proceedings of the Mediterranean Conference on Algorithms*, 7659:234–247, 2012.

[Ker39] R. Kershner. The Number of Circles Covering a Set. *American Journal of Mathematics*, 61(3):665–671, 1939.

[KF91] D. Khang and O. Fujiwara. Approximate Solutions of Capacitated Fixed-Charge Minimum Cost Network Flow Problems. *Networks*, 21(6):689–704, October 1991.

[Kin93] J. Kingman. *Poisson Processes*. Oxford Science Publications. Clarendon Press, 1993.

[KLZ09] R. Kwan, C. Leung, and J. Zhang. Proportional Fair Multiuser Scheduling in LTE. *IEEE Signal Processing Letters*, 16(6):461–464, June 2009.

[KW97] M. Kouider and P. Winkler. Mean Distance and Minimum degree. *Journal of Graph Theory*, 25(1):95–99, May 1997.

[LCB11] J. Lorincz, A. Capone, and D. Begušić. Optimized Network Management for Energy Savings of Wireless Access Networks. *Computer Networks*, 55(3):514–540, February 2011.

[Lee91] L. Leemis. Nonparametric Estimation of the Cumulative Intensity Function for a Nonhomogeneous Poisson Process. *Journal of Management Science*, 37(7):886–900, 1991.

[LFK10] H. Lichte, H. Frey, and H. Karl. Fading-Resistant Low-Latency Broadcasts in Wireless Multihop Networks: The Probabilistic Cooperation Diversity Approach. In *Proceedings of the ACM international Symposium on Mobile Ad Hoc Networking and Computing*, pages 101–110. ACM, 2010.

[LGP12] J. Lorincz, T. Garma, and G. Petrovic. Measurements and Modelling of Base Station Power Consumption under Real Traffic Loads. *MDPI Journal of Sensors*, 12(4):4181–310, January 2012.

[Li09] F. Li. Competitive Scheduling of Packets with Hard Deadlines in a Finite Capacity Queue. *IEEE Conference on Computer Communications*, pages 1062–1070, April 2009.

[LLH+13] B. Lannoo, S. Lambert, W. Heddeghem, M. Pickavet, F. Tudelft, G. Koutitas, H. Niavis, A. Certh, M. Till, A. Fischer, H. de Meer, P. Ulanc, T. Papaioannou, N. Viet, T. Plagemann, and J. Aracil. Network of Excellence in Internet Science D8.1. Overview of ICT Energy Consumption. 2013.

[LLT+09] T. Lam, L. Lee, H. Ting, I. To, and P. Wong. Sleep with Guilt and Work Faster to Minimize Flow Plus Energy. *Automata, Languages and Programming*, pages 665–676, 2009.

[LRH10] U. Lee, I. Rimac, and V. Hilt. Greening the Internet With Content-Centric Networking. *Proceedings of International Conference on Energy-Efficient Computing and Networking*, 2010.

[LRKH11] U. Lee, I. Rimac, D. Kilper, and V. Hilt. Toward Energy-Efficient Content Dissemination. *IEEE Network Journal*, 2011.

[LS12] G. Lin and S. Soh. Power-Aware Routing in Networks with Delay and Link Utilization Constraints. *Proceedings of the IEEE Conference on Local Computer Networks*, pages 272–275, 2012.

[LSC12] D. Lee, H. Seo, and B. Clerckx. Coordinated Multipoint Transmission and Reception in LTE-Advanced: Deployment Scenarios and Operational Challenges. *IEEE Communication Magazine*, (February):148–155, 2012.

[LT04] J. Laneman and D. Tse. Cooperative Diversity in Wireless Networks: Efficient Protocols and Outage Behavior. *IEEE Transactions on Information Theory*, 50(12):3062–3080, December 2004.

[LTP02] N. Lev-Tov and D. Peleg. Exact Algorithms and Approximation Schemes for Base Station Placement Problems. *Proceedings of the Scandinavian Workshop on Algorithm Theory*, pages 15–27, 2002.

[LWS+10] K. Loa, C. Wu, S. Sheu, Y. Yuan, M. Chion, D. Huo, and L. Xu. IMT-Advanced Relay Standards. *IEEE Communications Magazine*, 48(8):40–48, 2010.

[MA05] A. Maaref and S. Aissa. Closed-Form Expressions for the Outage and Ergodic Shannon Capacity of MIMO MRC Systems. *IEEE Transactions on Communications*, 53(7):1092–1095, July 2005.

[MA11] V. Mancuso and S. Alouf. Reducing Costs and Pollution in Cellular Networks. *IEEE Communications Magazine*, 49(8):63–71, 2011.

[MBL13] J. Malmodin, P. Bergmark, and D. Lundén. The Future Carbon Footprint of the ICT and E&M Sectors. In *Proceedings of the International Conference on Information and Communication Technologies for Sustainability*, pages 12–20, 2013.

[MC09] M. Marsan and L. Chiaraviglio. Optimal Energy Savings in Cellular Access Networks. *Proceedings of the IEEE Communications Workshops*, 2009.

[MCCM12] M. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo. Multiple Daily Base Station Switch-Offs in Cellular Networks. *International Conference on Communications and Electronics*, pages 245–250, August 2012.

[MF11] P. Marsch and G. Fettweis. *Coordinated Multi-Point in Mobile Communications: From Theory to Practice*. Coordinated Multi-point in Mobile Communications: From Theory to Practice. Cambridge University Press, 2011.

[MFF10] P. Marsch, A. Fehske, and G. Fettweis. Increasing Mobile Rates While Minimizing Cost per Bit-Cooperation vs. Denser Deployment. *Proceedings of the International Symposium on Wireless Communication Systems*, 2010.

[MG11] S. McLaughlin and P. Grant. Techniques for Improving Cellular Radio Base Station Energy Efficiency. *IEEE Wireless Communications*, (October):10–17, 2011.

[MLOH10] J. Manner, M. Luoma, J. Ott, and J. Hämäläinen. Mobile Networks Unplugged. *Proceedings of the International Conference on Energy-Efficient Computing and Networking*, page 71, 2010.

[MM11] M. Marsan and M. Meo. Energy Efficient Wireless Internet Access with Cooperative Cellular Networks. *Computer Networks*, 55(2):386–398, February 2011.

[MML+10] J. Malmodin, Å. Moberg, D. Lundén, G. Finnveden, and N. Lövehagen. Greenhouse Gas Emissions and Operational Electricity Use in the ICT and Entertainment & Media Sectors. *Journal of Industrial Ecology*, 14(5):770–790, October 2010.

[MMSL11] G. Micallef, P. Mogensen, H. Scheck, and J. Louhi. Reversing the Energy Trend in Mobile Networks. *Proceedigns of the IEEE Vehicular Technology Conference Fall*, (1):7–11, 2011.

[MW84] T. Magnanti and R. Wong. Network Design and Transportation Planning: Models and Algorithms. *Transportation Science*, 18(1):1, 1984.

[Nel95] R. Nelson. *Probability, Stochastic Processes, and Queueing Theory: The Mathematics of Computer Performance Modelling*. Springer, 1995.

[Neu81] M. Neuts. *Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach*. Phoenix Edition Series. Dover Publications, 1981.

[NH04] A. Nosratinia and T. Hunter. Cooperative Communication in Wireless Networks. *IEEE Communication Magazine*, 42(10):74–80, October 2004.

[Noz01] L. Nozick. The Fixed Charge Facility Location Problem with Coverage Restrictions. *Transportation Research Part E: Logistics and Transportation Review*, 37(4):281–296, August 2001.

[NP02] G. Norman and D. Parker. Formal Analysis and Validation of Continuous-Time Markov Chain Based System Level Power Management Strategies. *Proceedings of the IEEE International High-Level Design Validation and Test Workshop*, pages 45–50, 2002.

[NWGY10] Z. Niu, Y. Wu, J. Gong, and Z. Yang. Cell Zooming for Cost-Efficient Green Cellular Networks. *IEEE Communications Magazine*, 48(11):74–79, 2010.

[NYZR06] T. Ng, W. Yu, J. Zhang, and A. Reid. Joint Optimization of Relay Strategies and Resource Allocations in Cooperative Cellular Networks. *Proceedings of the Annual Conference on Information Sciences and Systems*, pages 1553–1559, March 2006.

[OK10] E. Oh and B. Krishnamachari. Energy Savings through Dynamic Base Station Switching in Cellular Wireless Access Networks. *Proceedings of the IEEE Global Telecommunications Conference*, December 2010.

[OKLN11] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu. Toward Dynamic Energy-Efficient Operation of Cellular Network Infrastructure. *IEEE Communications Magazine*, 49(6):56–61, 2011.

[Ols01] X. Olsthoorn. Carbon Dioxide Emissions from International Aviation: 1950–2050. *Journal of Air Transport Management*, 7(2):87–93, March 2001.

[OPTW07] S Orlowski, M Pióro, A Tomaszewski, and R Wessäly. SNDlib 1.0–Survivable Network Design Library. In *Proceedings of the INOC*, April 2007.

[Oso05] T. Osogami. Analysis of Multi-Server Systems via Dimensionality Reduction of Markov Chains. *PhD Thesis*, 2005.

[Pau11] U. Paul. Understanding Traffic Dynamics in Cellular Data Networks. *Proceedings of the INFOCOM*, pages 882–890, 2011.

[PDF+08] S. Parkvall, E. Dahlman, A. Furuskar, Y. Jading, M. Olsson, S. Wanstedt, and K. Zangi. LTE-Advanced - Evolving LTE towards IMT-Advanced. *Proceedings of the IEEE Vehicular Technology Conference Fall*, pages 1–5, September 2008.

[Ped99] M. Pedram. Dynamic Power Management based on Continuous-Time Markov Decision Processes. *Proceedings of the Design Automation Conference*, (c):555–561, 1999.

[PF11] S. Parkvall and A. Furuskar. Evolution of LTE toward IMT-Advanced. *IEEE Communications Magazine*, (February):84–91, 2011.

[Ple84] J. Plesník. On the Sum of All Distances in a Graph or Digraph. *Journal of Graph Theory*, 8(1):1–21, 1984.

[Pro01] J. Proakis. *Digital Communications*. McGraw-Hill, 2001.

[Pru07] K. Pruhs. Competitive Online Scheduling for Server Systems. *ACM SIGMETRICS Performance Evaluation Review*, 34(4):52–58, 2007.

[PS89] D. Peleg and A. Schäffer. Graph Spanners. *Journal of Graph Theory*, 13(1):99–116, March 1989.

[PSS09] J. Park, E. Song, and W. Sung. Capacity Analysis for Distributed Antenna Systems using Cooperative Transmission Schemes in Fading Channels. *IEEE Transactions on Wireless Communication*, 8(2):586–592, February 2009.

[PVC+09] B. Puype, W. Vereecken, D. Colle, M. Pickavet, and P. Demeester. Power reduction techniques in multilayer traffic engineering. *Proceedings of the International Conference on Transparent Optical Networks*, 1, 2009.

[QW00] Q. Qiu and Q. Wu. Dynamic Power Management of Complex Systems using Generalized Stochastic Petri Nets. *Proceedings of the Annual Design Automation Conference*, pages 352–356, 2000.

[QWB+09] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs. Cutting the electric bill for internet-scale systems. *Proceedings of the ACM SIGCOMM Conference on Data Communication*, page 123, 2009.

[Rat13] S. Rathbun. Estimation of Poisson Intensity Using Partially Observed Concomitant Variables. 52(1):226–242, 2013.

[RB03] P. Reynaud-Bouret. Adaptive Estimation of the Intensity of Inhomogeneous Poisson Processes via Concentration Inequalities. *Probability Theory and Related Fields*, 153:103–153, 2003.

[RC09] S. Ramprashad and G. Caire. Cellular vs. Network MIMO: A Comparison Including the Channel State Information Overhead. In *Proceedings of the IEEE Symposium on Personal, Indoor and Mobile Radio Communication*, pages 878–884, September 2009.

[RFF09] F. Richter, A. Fehske, and G. Fettweis. Energy Efficiency Aspects of Base Station Deployment Strategies for Cellular Networks. *Proceedings of the IEEE Vehicular Technology Conference Fall*, September 2009.

[RFGB10] F. Richter, G. Fettweis, M. Gruber, and O. Blume. Micro Base Stations in Load Constrained Cellular Mobile Radio Networks. *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications Workshops*, pages 357–362, 2010.

[RKM05] Z. Ren, B.H. Krogh, and R. Marculescu. Hierarchical Adaptive Dynamic Power Management. *IEEE Transactions on Computers*, 17(3):409–420, March 2005.

[RMHL10] P. Reviriego, J. Maestro, J. Hernadez, and D. Larrabeiti. Burst Transmission in Energy Efficient Ethernet. *IEEE Internet Computing*, 2010.

[Ros09] S. Ross. *Introduction to Probability Models*. Elsevier Science, 2009.

[RRMF13] B. Rengarajan, G. Rizzo, M. Marsan, and B. Furletti. QoS-Aware Greening of Interference-Limited Cellular Networks. *IEEE International Symposium on World of Wireless, Mobile and Multimedia Networkss*, pages 1–9, June 2013.

[Rui11] R. Ruismäki. Mobile Traffic Forecasts 2010-2020. *UMTS Forum*, 44, 2011.

[Sau10] M. Sauter. *From GSM to LTE: An Introduction to Mobile Networks and Mobile Broadband*. Wiley Online Library: Books. Wiley, 2010.

[SBdM00] T. Simunic, T. Benini, and G. de Micheli. Quantitative Comparison of Power Management Algorithms. *Proceedings of the Design, Automation and Test in Europe Conference and Exhibition*, pages 20–26, 2000.

[SDA+11] B. Song, S. Das, F. Akashi, C. Chevallier, and S. Soliman. Network Scaling for Achieving Energy Efficient Cellular Networks - A Quantitative Analysis. In *Proceedings of the IEEE Vehicular Technology Conference Fall*, pages 1–5, September 2011.

[SER11] L. Saker, S. Elayoubi, and L. Rong. Capacity and Energy Efficiency of Picocell Deployment in LTE-A Networks. *Proceedings of the IEEE Vehicular Technology Conference Spring*, pages 1–5, May 2011.

[Ser12] M. Sereno. Cooperative Game Theory Framework for Energy Efficient Policies in Wireless Networks. *Future Energy Systems: Where Energy, Computing and Communication Meet*, 2012.

[Sey05] J. Seybold. *Introduction to RF Propagation*. Wiley, 2005.

[SFdSF12] V. Suryaprakash, A. Fehske, A. F. dos Santos, and G. Fettweis. On the Impact of Sleep Modes and BW Variation on the Energy Consumption of Radio Access Networks. *Proceedings of the IEEE Vehicular Technology Conference Spring*, pages 1–5, May 2012.

[SFH04] M. Schinnenburg, I. Forkel, and B. Haverkamp. Realization and Optimization ofSoft and Softer Handover in UMTS Networks. In *Proceedings of the European Personal Mobile Communications Conference*, pages 603–607. IET, 2004.

[She10] X. She. MIMO and CoMP in LTE-Advanced. 12(2):20–28, 2010.

[SHL06] A. Sadek, Z. Han, and K. Liu. An Efficient Cooperation Protocol to Extend Coverage Area in Cellular Networks. In *Proceedings of the IEEE Wireless Communications and Networking Conference*, volume 3, pages 1687–1692. IEEE, 2006.

[Sir02] V. Siris. Resource Control for Elastic Traffic in CDMA Networks. *Proceedings of the Conference on Mobile computing and Networking*, 2002.

[SK10] M. Sawahashi and Y. Kishiyama. Coordinated Multipoint Transmission/Reception Techniques for LTE-Advanced. *IEEE Wireless Communications*, (June):26–34, 2010.

[SK12] K. Son and B. Krishnamachari. SpeedBalance: Speed-Scaling-Aware Optimal Load Balancing for Green Cellular Networks. *Proceedings of the IEEE INFOCOM*, pages 2816–2820, March 2012.

[SKYK11] K. Son, H. Kim, Y. Yi, and B. Krishnamachari. Base Station Operation and User Association Mechanisms for Energy-Delay Tradeoffs in Green Cellular Networks. *IEEE Journal on Selected Areas in Communications*, 29(8):1525–1536, September 2011.

[SLH+01] D. Staehle, K. Leibnitz, K. Heck, B. Schroder, A. Weller, and P. Tran-Gia. Analytical Characterization of the Soft Handover Gain in UMTS. *Proceedings of the IEEE Vehicular Technology Conference Fall*, pages 291–295, 2001.

[Soi08] A. Soifer. *The Mathematical Coloring Book: Mathematics of Coloring and the Colorful Life of its Creators*. Springer, 2008.

[SSBNS06] O. Simeone, O. Somekh, Y. Bar-Ness, and U. Spagnolini. Low-SNR Analysis of Cellular Systems with Cooperative Base Stations and Mobiles. In *Asilomar Conference on Signals, Systems and Computers*, pages 626–630. IEEE, 2006.

[SSPS09] O. Simeone, O. Somekh, H. Poor, and S. Shamai (Shitz). Local Base Station Cooperation Via Finite-Capacity Links for the Uplink of Linear Cellular Networks. *IEEE Transactions on Information Theory*, 55(1):190–204, January 2009.

[Ste04] J. Steele. *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. MAA problem books series. Cambridge University Press, 2004.

[Sti70] S. Stidham Jr. On the Optimality of Single-Server Queuing Systems. *Operations Research*, 18(4):708–732, 1970.

[SYLY03] Y. Shu, M. Yu, J. Liu, and O. Yang. Wireless Traffic Modeling and Prediction using Seasonal ARIMA Models. *Proceedings of the IEEE International Conference on Communications*, 3:1675–1679, 2003.

[TDD12] R. Torrea-Duran and C. Desset. Adaptive Energy Efficient Scheduling Algorithm for LTE Pico Base Stations. In *Proceedings of the Future Network & Mobile Summit*, pages 1–8, 2012.

[TGAA13] D. Tsilimantos, J. Gorce, E. Altman, and S. Antipolis. Stochastic Analysis of Energy Savings with Sleep Mode in OFDMA Wireless Networks. In *IEEE International Conference on Computer Communications*, 2013.

[Tot72] L. Toth. *Lagerungen in der Ebene, auf der Kugel und im Raum.* Springer-Verlag, 1972.

[TV05] D. Tse and P. Viswanath. *Fundamentals of Wireless Communication.* Cambridge University Press, 2005.

[TZJ11] O. Tipmongkolsilp, S. Zaghloul, and A. Jukan. The Evolution of Cellular Backhaul Technologies: Current Issues and Future Trends. *IEEE Communications Surveys & Tutorials*, 13(1):97–113, 2011.

[TzJwHj04] C. Tian-zhou, H. Jiang-wei, and D. Hong-jun. The Dynamic Power Management for Embedded System with Poisson Process. *Journal of Zhejiang University SCIENCE*, 6(Suppl. I):70–74, August 2004.

[U.S12] U.S. Energy Information Administration. *Annual Energy Review 2011.* 2012.

[U.S13] U.S. Energy Information Administration. *Annual Energy Outlook 2013.* 2013.

[Vaz01] V. Vazirani. *Approximation Algorithms.* 2001.

[VDC+12] W. Vereecken, M. Deruyck, D. Colle, W. Joseph, M. Pickavet, L. Martens, and P. Demeester. Evaluation of the Potential for Energy Saving in Macrocell and Femtocell Networks using a Heuristic Introducing Sleep Modes in Base Stations. *EURASIP Journal on Wireless Communications and Networking*, (1):170, 2012.

[VGZ94] A. Viterbi, K. Gilhousen, and E. Zehavi. Soft Handoff Extends CDMA Cell c Coverage and Increases Reverse Link Capacity. *Mobile Communications Advanced Systems and Components*, pages 541–551, 1994.

[VH11] W. Vereecken and W. Heddeghem. Power Cconsumption in Telecommunication Networks: Overview and Reduction Strategies. *IEEE Communications Magazine*, 49(6):62–69, 2011.

[VK10] N. Vasić and D. Kostić. Energy-Aware Traffic Engineering. In *Proceedings of the International Conference on Energy-Efficient Computing and Networking*, pages 169–178. ACM, 2010.

[Weia] E. Weisstein. Circle-Circle Intersection. *From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/Circle-CircleIntersection.html.*

[Weib] E. Weisstein. Hyperfactorial. *From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/Hyperfactorial.html.*

[WK08] D. Woldegebreal and H. Karl. Network-Coding-Based Cooperative Transmission in Wireless Sensor Networks: Diversity-Multiplexing Tradeoff and Coverage Area Extension. *Wireless Sensor Networks*, pages 141–155, 2008.

[WMBW09] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz. Primary User Behavior in Cellular Networks and Implications for Dynamic Spectrum Access. *IEEE Communications Magazine*, (March):88–95, 2009.

[WO08] M. Webb and Others. Smart 2020: Enabling the Low Carbon Economy in the Information Age. *The Climate Group. London*, 2008.

[WXW11] P. Wan, X. Xu, and Z. Wang. Wireless Coverage with Disparate Ranges. In *Proceedings of the International Symposium on Mobile Ad Hoc Networking and Computing*, page 1, New York, USA, May 2011. ACM Press.

[ZC05] F. Zhang and S. Chanson. Power-Aware Processor Scheduling under Average Delay Constraints. *Proceedings of the Real Time and Embedded Technology and Applications Symposium*, pages 202–212, 2005.

[ZGY+09] S. Zhou, J. Gong, Z. Yang, Z. Niu, and P. Yang. Green Mobile Access Network With Dynamic Base Station Energy Saving. In *Proceedings of the ACM International Conference on Mobile Computing and Networking*, volume 9, pages 10–12, 2009.

[ZH10] R. Zakhour and S. V. Hanly. Large System Analysis of Base Station Cooperation on the Downlink. In *Proceedings of the Annual Allerton Conference on Communication, Control, and Computing*, pages 270–277. Ieee, September 2010.

[ZHS04] R. Zheng, J.C. Hou, and L. Sha. On Time-Out Driven Power Management Policies in Wireless Networks. In *Proceedings of the IEEE Global Telecommunications Conference*, volume 6, pages 4097–4103. Ieee, 2004.

C. Glossary

activation	A transient state in which a base station (BS) changes from sleep to active denoted as S_U. xii, xiii, xvi, 10–12, 20–28, 31, 33–35, 37, 38, 40, 42–44, 47, 60, 77–79, 91, 94, 115, 118, 120–122, 124, 125, 127, 129, 131, 159
active	A state in which a BS is ready to perform work (but does not necessarily perform work) denoted as S_A. iii, xii, xv, 3, 20, 23, 25, 28, 32, 34, 35, 40, 47, 48, 53, 72, 73, 78–81, 84–86, 90–94, 114, 116–118, 120–122, 124, 129, 130, 157, 158
cell	The area a BS provides coverage to (either data or signaling traffic). 62, 64–66, 68, 70, 80, 90, 107, 108, 120, 159
configuration	A state of a radio access network (RAN) which describes the activity state of each BSs denoted as C. xii, 45–49, 51, 55, 57, 78, 86, 115, 133, 134, 138
cooperation	Multiple BSs transmitting to the same user equipment (UE) to increase its signal strength. iii, xii–xiv, xvi, xix, xx, 2, 12–15, 17–19, 60–73, 75–81, 85–89, 91, 93–98, 100, 101, 104, 108–114, 118–121, 124, 125, 127–131
data traffic	The traffic that a UE wants to transmit (compare signaling traffic). xv, 2, 4, 5, 11, 12, 17, 78, 80, 129, 157–159
deactivation	A transient state in which a BS changes from active to sleep denoted as S_D. iii, xiv, xvi, xvii, 10, 11, 19, 21–24, 28, 29, 31, 35, 37, 40, 43, 44, 46, 56, 58, 91, 115, 118, 120, 121, 125, 127, 131, 159
deployment	The distribution of BSs in an area or plane. xii, xvi, xix, 14, 61, 64–67, 72, 78, 80, 86–91, 94, 116–118, 123–128, 158

hexagonal — A deployment in which each BSs has six neighboring BSs which are all at the same distance to it and each other. 64, 66, 69, 72, 73, 78–80, 84, 90, 91, 94, 117, 118, 125, 159

hotspot — A (small) area where users generate above-mean traffic (e.g., in a hotel or restaurant). 11, 12, 16, 117, 123–127, 131, 132

idle — A state in which a BS is active, but is not performing any work. 3, 4, 10, 12, 26, 28, 29, 32, 46, 48, 57, 58, 118, 120, 124, 129, 159

interference — Signals from other transmissions that alter the received signal. iii, x, 5, 17–19, 63, 95, 101, 117, 120, 121, 130, 131

latency — The time it takes the RANs to fully process a request from a UE (e.g., a file transmission) denoted as L. iii, xii–xiv, xviii, 5, 8, 10, 13, 16, 17, 20–33, 35–55, 57, 114, 124, 129, 133

load — The ratio of work a system performs to the work in could maximally perform (depending on the context this can be averaged over different time scales). 4–10, 13–16, 19, 23, 38–40, 45, 46, 55–58, 61, 62, 78, 80, 115–118, 121, 123–127, 129–131, 159

macro — A size category of BSs with long range and usually multiple sectors which I consider to serve only or mostly signaling traffic. xx, 11–14, 62, 79, 80, 114–118, 120, 122–129

noise — Random fluctuation which of a signal denoted as N. x, xiv–xvi, 4, 5, 63, 97, 101, 117, 158

outage — The state of a channel in which the signal-to-noise ratio (SNR) is too low to transmit at the outage capacity. xiv, xix, 5, 61, 62, 75, 76, 79, 95–105, 107–113, 130, 158

path loss — The reduction of signal strength over traveled distance. xiii, 3, 5, 19, 60, 63–66, 68–73, 75–77, 80, 107, 108, 117, 130

pico — A size category of BSs with short range which I consider to serve only data traffic. 11–13, 79, 80, 115–118, 120–130

policy — A procedure that describes how to determine which BS are activated and which are deactivated at which time ((de-)activation policy). xii–xv, xviii, 10, 20–44, 48, 120–126, 132, 159

power cycle — The activation and deactivation of a BS or the other way around. iii, xviii, xx, 11, 16, 20, 23, 24, 27–29, 32–35, 37, 40, 42–45, 59, 78, 114, 120, 122, 125, 128–132

power profile — The function from load to power consumption of BS or the complete RAN. iii, 6–9, 11, 12, 45, 52, 59, 101, 118, 131, 132

range — The distance over which a BS can reliably transmit (data or signaling traffic) denoted as r_S for signaling traffic and r_D for data traffic. iii, 4, 5, 11, 12, 16, 21, 60–66, 71–73, 75, 78, 80, 81, 84, 85, 94, 112, 129, 130, 158

scheme — A procedure that describes how to determine which BS serves the data traffic of which UE (association scheme). xiii–xvi, xix, 9, 12, 21, 79, 85, 86, 88, 91, 93, 94, 96, 100, 101, 106, 108–113, 120, 159

sector — The area of a cell that is covered by a single antenna. 19, 116, 117, 158

signaling traffic — Traffic that is needed to control the flow of data traffic. xv, xvi, 2, 4, 5, 11, 12, 17, 60, 63, 72, 73, 77, 78, 117, 120, 131, 157–159

sleep — A state in which a BS is *not* ready to perform work and consumes less power than being idle denoted as S_S. iii, 1, 6, 8, 11, 12, 14–16, 20, 24, 25, 29, 44, 45, 47, 51, 79, 90, 114–116, 118, 120, 121, 129, 131, 157

spacing — The distance of neighboring BSs in a hexagonal deployment, also known as inter-site distance (ISD) denoted as ξ. xvi, 60, 64–66, 68–70, 72, 81–94

traffic — The flow of data between UEs and BSs. xviii, 3, 4, 8–12, 15, 16, 20, 23, 45, 52, 80, 91, 115, 117, 129, 131, 132, 157–159